
GEB-BENCH: ABSTRACT STRUCTURES TOLD IN MANY VOICES

Tong Zhang

Fudan University
tongzhang25@m.fudan.edu.cn

Zhiyuan Shi

aron.c.shi@gmail.com

Yun Peng

Fudan University
yunpeng4@sigsoft.org

Tao Xie*

Peking University
Beijing Tongming Lake Information Technology
Application Innovation Center
taoxie@pku.edu.cn

Project page: <https://gebbench.stonezhang.com>

ABSTRACT

Humans routinely reuse an abstract pattern across situations that share little in appearance, which lets one insight transfer between tasks that look unrelated. This ability underlies analogy, and it is a candidate ingredient of general intelligence. It is therefore worth asking whether multimodal models can carry a pattern across the very different ways the same pattern can be told. We introduce GEB-BENCH, a benchmark for recognition and correspondence across four voices, namely scene images, short stories, mathematical statements and structural skeletons. Its motifs, among them recursion, self-reference and Möbius identification, are held fixed while the voice that presents them changes. Motif identification asks which named structure a presented example realizes. Cross-voice matching asks which example in another voice shares its intended motif. A separate finite-graph pilot asks models to follow an explicit structural edit whose answer cannot be recovered from the endpoint name alone. In this suite, under the declared answer-extraction rule, every evaluated vision model is less accurate when choosing among four stories than when identifying the motif from four options with the same source scene and ordered candidate motif identities. Removing the image or rotating the answer positions changes this gap differently across models, and requiring the requested response format moves several scores further. The finite pilot shows that strong endpoint recognition can coexist with much weaker composition and intermediate-state accuracy. These findings indicate that naming a motif is not sufficient evidence of matching its realizations or of tracking exact structural edits. GEB-BENCH supplies paired tasks and retained responses for measuring this distinction and for testing how future models use visual and textual structure.

1 INTRODUCTION

People connect situations that share little in appearance. A chess exchange, a biological trade-off and a business negotiation can share one underlying structure, and once a person grasps that structure, an insight about one case can guide the others (Gentner, 1983). This flexible reuse of relational patterns is a hallmark of human cognition (Gentner, 1983; Falkenhainer et al., 1989) and a proposed building block of general intelligence (Chollet, 2019). The ability matters because it lets a limited store of experience transfer to novel problems, and because it supplies a unified way to organize tasks whose surface forms differ.

Hofstadter’s *Gödel, Escher, Bach* gives the canonical illustration (Hofstadter, 1979). A single abstract form can appear as a Bach canon, an Escher print and a Gödel sentence, and the shared form,

*Corresponding author.

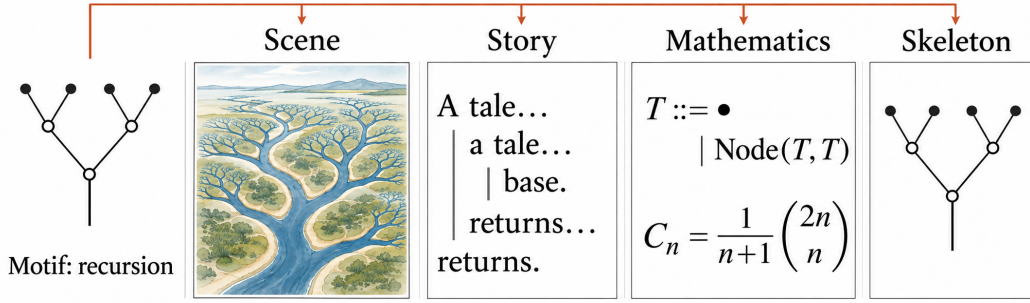


Figure 1: One motif, four voices (schematic). Recursion appears as a river-delta scene, nested telling with return, a full-binary-tree grammar and a skeleton. Counts, depths and angles are unscored nuisance variables that change the surface without changing the intended motif. GEB-BENCH tests recognition within a voice and matching across voices.

not the medium, is what the reader is invited to notice. A crab canon pairs a melody with its own reverse, and Escher’s *Drawing Hands* draws a loop in which each hand draws the other. For a multimodal model the corresponding test is whether it can recognize that an image, a narrative and a symbolic description carry one pattern, rather than merely name the pattern in one of them.

We therefore ask a focused question. When a model can name a structural motif in one representation, can it also find the same motif expressed in another, and can it act on a stated structural change? Prior multimodal benchmarks establish that models handle equivalent presentations of a problem unevenly (Fu et al., 2024a; Tang et al., 2025), and compositional-visual work shows that sensitivity to wording and ordering is widespread (Thrush et al., 2022; Yue et al., 2024). Those findings motivate a benchmark that separates recognizing a motif from matching it across representations, and that keeps the two directly comparable on a fixed item. We introduce GEB-BENCH for this purpose.

GEB-BENCH is organized around a fixed library of structural motifs. A motif is an intended recognition target, such as a cycle, a recursion or a Möbius twist. A voice is a way of presenting a motif, and the benchmark uses four, crossing modality with descriptive versus formal presentation (Figure 1). Scenes are the descriptive visual voice, natural or artistic images whose composition is meant to express the motif. Skeletons are the formal visual voice, programmatic diagrams that make selected relations explicit. Stories are the descriptive language voice, where the motif is enacted by the telling itself, for example through a form device such as a tale whose two halves mirror each other. Mathematics is the formal language voice, using definitions and formulas together with an attributed theorem. The four voices share one motif list, so the same answer set applies wherever a motif is presented. Beyond the benchmark, we use this shared list to build paired questions that hold the item fixed and change one factor at a time.

We chose tasks of a kind whose answers are closed forms. Every item has one intended answer, drawn from a finite set, and a program compares the reply with the frozen key by exact string match. No judge model and no graded rubric enter the score. This choice follows from the question, not from convenience. A comparison between recognizing a motif and matching it across voices is only as clean as the two scores it contrasts, and an item that admits a single finite answer lets identification and matching differ along one axis while the scene and the candidate motifs are held fixed. The same design makes structural edits checkable on labeled graphs, where an edited edge set and every intermediate state can be replayed exactly. The three task families then probe distinct questions. Motif identification asks which named structure a realization presents. Cross-voice matching asks which realization in another voice shares its intended motif, and it is scored against recognition on the same item. A finite-graph pilot asks for the result of a specified edit, whose answer cannot be read off the endpoint alone. Section 3 gives the protocol, and the task construction is inspired by compositional notation (Fong & Spivak, 2019).

The results separate the three questions. Under the declared compatibility scoring rule, **every evaluated vision model is less accurate when choosing among four stories than when identifying the motif from four options with the same source scene and ordered candidate motif identities.**

The image and candidate order are held fixed, so the drop tracks the change from a motif menu to full stories rather than a change of source item. Removing the image or rotating the answer positions gives a less uniform picture, and the effects differ across models, so the gap alone cannot settle whether a model ignored the image or failed to read a story. Requiring the requested response format moves several scores further, which separates a correct answer extracted from a reply from compliance with the requested output. In the finite pilot, **strong endpoint recognition coexists with much weaker composition and intermediate-state accuracy**, so naming a terminal structure does not summarize the operations that reach it. These findings are developed under the metric definitions and with the model-by-model evidence reported in Section 4.

Contributions.

- We introduce a motif benchmark that measures abstract transfer across four voices and separates recognition from cross-voice matching on paired items.
- We provide a complete pipeline that generates the four-voice realizations, checks each item against executable clauses, and freezes the validated items with their keys, so that every reported score is reproducible from retained replies.
- We report findings that a recognition score alone does not predict, namely a recognition-matching gap across the evaluated vision models, a disassociation between endpoint recognition and exact composition, and a case in which compatibility accuracy conceals a response-format failure.

2 THE MOTIF LIBRARY

The benchmark is built on a fixed library of 25 structural motifs, and every task in this paper draws its answer set from that library. Each motif carries one recognition target, a set of mathematical associations and a story form device, and the seven conceptual reading groups shown in Figure 3 organize the library for the reader. A reading group is a presentation aid rather than a scored superclass, so the graded label is always the individual motif. The scene gallery (Figure 4) shows the visual variety of the collection, and Table 6 reports the uneven coverage of the classes. This section introduces the motifs and then describes how the story checks are made executable.

2.1 TWENTY-FIVE MOTIFS

Motifs that describe how a construction returns to its own description form the first group. Self-reference points back to its own statement in a single step, a strange loop climbs a hierarchy of levels and lands back where it began, recursion repeats a construction inside smaller copies of itself, and infinite descent invokes a smaller witness to deny that a least instance exists.

A second group concerns motion that settles or repeats. Contraction applies the same rule until a state stops changing, convergence sends independent paths to one shared endpoint, a cycle returns to its starting point after one lap, and a coprime orbit visits every position of a ring under a fixed stride whose value is coprime to the ring length.

A third group concerns orientation and containment in space. The Möbius motif reverses orientation after one traversal, a spiral grows by a fixed ratio on each return, nesting adds levels without ever returning to an outer one, and figure and ground exchange which region is the subject under the same order of events.

A fourth group concerns regularities and their violations. Mirror symmetry preserves a pattern under reversal, symmetry breaking violates an otherwise shared rule exactly once, alternation switches between two roles without exception, and projective duality swaps two classes of roles while every relational statement stays true.

A fifth group concerns periodic structure. Periodic tiling repeats a unit so that a shift coincides with the original, aperiodic tiling covers the plane under local matching rules while admitting no translational period at all, braiding composes the crossings of several strands, and interference superposes repeating patterns into a larger joint period.

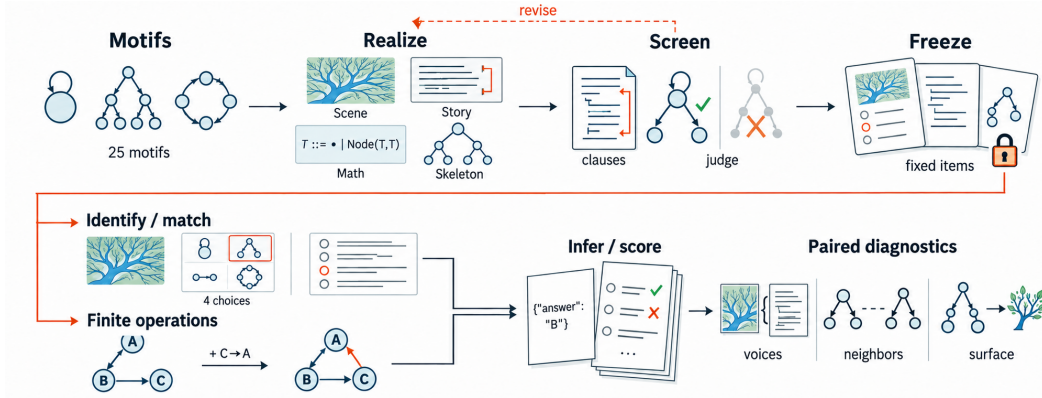


Figure 2: Construction and evaluation (schematic). Four-voice realizations undergo finite checks, and judge flags are retained separately from item admission. Frozen items feed identification and matching, and the finite operations, into common scoring and paired diagnostics. Four choices illustrate the matching and family-menu tasks, while full identification uses all 25 options, and the pictured graph edit adds $C \rightarrow A$.

A sixth group concerns relations among parts of a graph. A bottleneck is a single point that every path between two regions must cross, a hub connects otherwise separate participants, and isomorphism is a relation-preserving renaming between two structures.

The final group concerns counting. The pigeonhole principle forces a collision when the available slots are too few, and diagonalization constructs a new object that differs from each entry of a listing.

These descriptions state the intended recognition targets, and they do not certify every property from a single finite picture. A finite tiling patch, for example, cannot establish global aperiodicity. The correspondence between the gallery names, the formal catalog names and the frozen option identifiers appears in Table 7, and the signature equation, mathematical home and story device of each motif appear in the full catalog in Appendix J.

The same motif admits many realizations that differ in branch counts, nesting depths, orientations or visual style. These rendering parameters are recorded as nuisance variables and excluded from the answer, so a realization is the concrete presentation rather than a claim that two voices agree as formal definitions. Mathematical associations give a more explicit view of the intended structure, and the catalog lists them for every motif. Two examples are Menger’s theorem, which governs the bottleneck motif (Menger, 1927), and Cantor’s diagonal argument, which governs diagonalization. The associations differ in how directly they characterize a rendered example, and a shared label does not assert that all voices are equivalent mathematical definitions.

2.2 STORY FORM DEVICES

The scored stories are English. A form device concerns how the text is organized, rather than whether its plot merely mentions a motif, and a finite predicate is a specified yes-or-no test on that organization. The symmetry device, for example, checks whether lines paired around a central line have matching word-token lists after the declared normalization. The scored story *Red Mitten* repeats its first line as its last line with a different line between them, and other permitted pairs can exchange two names while punctuation is ignored. The complete text and all of its checks appear in Appendix K. Constraints of this kind connect the construction to constrained generation and instruction following (Yao et al., 2023; Zhou et al., 2023).

The construction admits stories through selected executable checks and body deduplication. Historical model-based screening is recorded separately, and the two procedures differ in what the model receives. The device judge receives the intended specification, whereas an auxiliary classifier receives the story and the motif menu without a marked answer. A classifier’s verdict is therefore not the admission rule, and a passing text check is not independent semantic certification. Contraction

is checked on a stated finite state space, and descent stories state a smaller-witness premise and a least-witness contradiction. Every model request preserves a story’s full title and body.

The ledger records retained, repaired, deduplicated and ambiguous examples, including stories that pass the stated checks despite a historical classifier rejection, and it distinguishes deterministic visual repairs from generated scenes. These flags were fixed without using the evaluated model responses. They support sensitivity analyses while preserving the limits of finite checks, since a test can verify its stated clause without establishing every possible interpretation of a story or an image.

3 TASKS AND PROTOCOL

The benchmark asks three questions about a single motif library. Can a model identify a motif from one realization, match the same motif across two voices, and carry out a specified structural edit? We chose tasks of this kind because each answer is a closed form. An item has one intended answer drawn from a finite set, and a program compares the reply with the frozen key string by string. No judge model and no graded rubric enter the score. A binary correctness indicator therefore attaches to every item, and the contrast between two tasks on the same item is an exact difference rather than a judgment about quality. This makes attribution of a failure to a task possible, since identification, matching and the operation pilot vary one factor at a time. Recognition and matching draw on the realization pools, while a separate finite pilot makes edits and their intermediate targets exactly checkable. The organization is inspired by compositional notation (Fong & Spivak, 2019); Appendix G explains the optional category-theory interpretation. The task definitions do not require that these pools form categories or that their realizations are functors.

Definition 1 (Motifs and realizations). Let $\mathcal{M} = \{m_1, \dots, m_{25}\}$ be the motif library and $v \in \{S, T, H, K\}$ a voice, namely the scene, the story, the mathematics presented as a theorem, or the skeleton. A realization $x = F_v(m; \xi)$ is one concrete image or text in voice v that realizes an intended motif m with recorded nuisance parameters ξ . An evaluation item is a scored question that comprises the presented realization, the prompt and the answer key. We omit ξ when the realization is fixed. Admissible nuisance variations preserve the intended label by construction, while model predictions need not be invariant to them. A skeleton is a programmatic two-dimensional rendering intended to reduce surface complexity. It is a reference condition for recognition rather than a theoretical upper bound, and a finite rendering need not certify the full invariant.

Definition 2 (Identification and matching). Identification presents a realization $x = F_v(m; \xi)$ and asks the model to predict its intended motif. The scored comparison is

$$\hat{m} = R_v(x), \quad \text{correctness} = \mathbf{1}[\hat{m} = m]. \quad (1)$$

Here R_v denotes the model with the prompt, menu convention and response parser fixed for the identification condition, and $\mathbf{1}[\cdot]$ is one when its condition holds and zero otherwise. Nuisance parameters can change the presented realization and hence the prediction, even when the intended label stays fixed. Cross-voice matching presents a realization in voice u and asks the model to select its intended motif’s realization in voice v from four candidates. For a fixed source pool I_u , let a_j and b_j be the binary correctness indicators for matched-arity identification and cross-voice matching on source item j , respectively. Their paired empirical contrast is

$$\Delta_{u \rightarrow v} = \frac{1}{|I_u|} \sum_{j \in I_u} (a_j - b_j). \quad (2)$$

We call this contrast the mapping tax. The reported comparison uses scenes ($u = S$) and stories ($v = T$). It holds the scene and the number of answer options fixed, while the option text, the representation and the distractor content vary. It therefore measures a difference between the administered tasks without isolating a unique internal transport process.

The primary suite combines identification and matching with a paired family-menu identification task (Family4). Table 1 lists the primary tasks and the realized item counts, while Appendix C gives every suite size. Each item has one intended answer in the frozen key. Scoring is deterministic given the stored response and that key, and semantic label validity remains a separate question. The collection reuses realizations across tasks, so it is not a disjoint train and test benchmark.

Task	Job	Given $\rightarrow$ pick	Items	Mode	Released id
Scene identification	identify	scene $\rightarrow$ motif	210	VLM	scene2motif
Skeleton identification	identify	skeleton $\rightarrow$ motif	49	VLM	skeleton2motif
Story identification	identify	story $\rightarrow$ motif	121	LLM	story2motif
Theorem identification	identify	theorem $\rightarrow$ motif	25	LLM	theorem2motif
Scene-story matching	match	scene $\rightarrow$ story	210	VLM	crossvoice
Scene-theorem matching	match	scene $\rightarrow$ theorem	210	VLM	xv_scene_thm
Story-theorem matching	match	story $\rightarrow$ theorem	121	LLM	xv_story_thm
Family-menu scene id.	paired id.	scene $\rightarrow$ motif (of 4)	210	VLM	scene2motif_family4

Table 1: Primary object-layer tasks. Identification uses the full motif menu, while matching and family-menu identification use four candidates. Family4 repeats the scene with the exact candidate motif order used in scene-story and scene-theorem matching. Image-withheld and rotation conditions are reported separately and excluded from the macro average. VLM denotes an image-input task and LLM denotes a primary text task.

GEB-BENCH is English-primary. The option menu, the theorems and the stories administered to models are English, and each story’s form device is defined on English mechanics, so the device predicates concern the administered English text.

The four 25-way identification tasks share one option menu, the 25 motif identifiers, each paired with its frozen structure description. Appendix O reproduces the full menu. The menu names the formal content of each motif rather than any surface topic, to emphasize structural content, although lexical overlap may still help. The family-menu condition shows the same structure descriptions restricted to four options. Chance performance is determined by menu arity on identification, matching and the family-menu condition (Table 1), and the ordered motif candidates of Family4 match the scene-story question of Section 4.2. Family4 predicts a motif identifier rather than a coarse taxonomy-group label. In theorem identification and in both theorem-matching tasks, selected motif-revealing terms are deterministically masked with $\square$, while named attributions, citations, formulas and other lexical cues can remain.

3.1 CANDIDATE MATCHING AND CUE CONDITIONS

Designed motif families supply related distractors. The item builder uses the same ordered candidate motif list for the scene-story, scene-theorem and family-menu questions associated with each scene. It balances correct answer positions and chooses the closest available story lengths within that constrained candidate set. This does not make all options length-balanced, and the released length strata retain this limitation. Every story is delivered with its full body and title, including the title needed to evaluate an acrostic form device.

Two further conditions vary a single cue while the item set is held fixed. The image-withheld condition removes the image fields entirely, with no dummy image, while preserving the candidates and the correct position. The request contains text only, which tests candidate-only shortcuts (Hsieh et al., 2023; Yue et al., 2025). The rotation condition preserves the image and the options but cycles their labels, which tests sensitivity to answer order (Lewis & Mitchell, 2024). The administered item, not only its task name, determines whether it enters a primary score, so rotated items retain the cross-voice task name but are excluded from the primary macro. These repeated realizations support paired comparisons and must not be counted as independent newly sampled problems.

3.2 FINITE ENDPOINT AND OPERATION PILOT

Recognizing a motif or matching two realizations does not show whether a model can track a change to the structure. A direct test on scenes or stories would require a unique edited target, which their intended motif labels do not specify. The finite pilot therefore separates endpoint association from exact operation tasks. Its finite constructions are scored as declared, so a substitution strip is a finite word and an incidence-duality example is a polygonal cycle rather than a certification of an aperiodic plane tiling or a projective plane.

The operation tasks use labeled directed graphs because both the edited edge set and every intermediate state can be replayed exactly. Models see rendered graph panels and text instructions. Relabeling changes the names of the nodes while preserving adjacency, so the same structural edit must refer to the corresponding renamed nodes. This provides a concrete test of edit transport without assuming a unique edit between a scene and a story.

Let $G = (V, E)$ be a finite labeled directed graph with $E \subseteq V \times V$, and let $\pi : V \rightarrow V$ be a permutation of its node labels. For an edge $e = (u, v)$, write $\pi e = (\pi(u), \pi(v))$ and $\pi E = \{\pi e : e \in E\}$. Adding e and then relabeling has the same endpoint as relabeling first and adding πe .

$$\pi(E \cup \{e\}) = (\pi E) \cup \{\pi e\}. \quad (3)$$

A concrete algebraic example fixes the intended distinction. Take $V = \{a, b, c\}$ and $E = \{(a, b)\}$. Let π exchange a and b while fixing c , and add $e = (a, c)$. The two operation orders give

$$\begin{aligned} \text{add then relabel} \quad & \{(a, b)\} \rightarrow \{(a, b), (a, c)\} \rightarrow \{(b, a), (b, c)\}, \\ \text{relabel then add} \quad & \{(a, b)\} \rightarrow \{(b, a)\} \rightarrow \{(b, a), (b, c)\}. \end{aligned}$$

The second chain adds the transported edge $\pi e = (b, c)$. Its first intermediate edge set differs from that of the first chain, although their final edge sets coincide. This illustrates the distinction tested by the paired questions below, and it is not an additional evaluated item. Deletion has the analogous identity. Every composition question specifies a source graph and an operation chain. The pictured answer choices have distinct labeled edge sets but are isomorphic to the same endpoint family, so naming that family cannot identify the correct graph.

Paired questions use the two operation orders of Equation (3). They share the source graph, the final endpoint and the displayed choices, but they ask for the state after the first operation, so their correct choice labels c_1 and c_2 differ. A deterministic rule restricted to the common source, endpoint and choices must return the same label $\hat{c}$ for both questions. Because $c_1 \neq c_2$,

$$\mathbf{1}[\hat{c} = c_1] + \mathbf{1}[\hat{c} = c_2] \leq 1. \quad (4)$$

Thus solving both questions requires information beyond their shared endpoints, such as the specified operation order. This statement does not identify a model’s internal algorithm and does not establish category-theory laws. Independent validation replays the serialized operations and checks unique answers, graph isomorphism and distinct intermediate targets.

3.3 MODELS, EVIDENCE AND THE TWO SCORING RULES

The evaluated roster is enumerated in the tables. The evaluation unit is a checkpoint or a hosted API identity rather than an independent architecture family. Hosted models use pinned API identifiers, and models evaluated from local weights use original checkpoint files verified against complete source manifests. This includes hosted access to the DeepSeek baseline, and the runtime provenance distinguishes these routes. The Qwen model families follow Bai et al. (2025); Wang et al. (2024); Qwen Team (2024). No model score is inherited from a different model or API identifier. Hosted API identifiers fix the evaluated endpoints, and their backend architecture or vendor mapping is not independently verified.

Each reply is scored under two declared rules. Compatibility accuracy credits an answer recovered by the fixed extraction procedure, which first attempts the requested JSON form and then falls back to a fixed extraction. Strict accuracy credits only a single JSON object that carries the requested field and a valid answer. Both are computed from the same retained reply, so the two values differ in the response rule rather than in the underlying attempt. A visible unparseable reply counts wrong, and a missing network response is never counted as an observed wrong answer. An explicit refusal or a fully evidenced exhausted generation budget is a completed nonanswer, whereas a generic empty response or a transport failure remains incomplete. Prompts request one JSON field under temperature-zero decoding, and every received reply is retained with its request hash, image hashes, provider metadata and parsing outcome.

All models share the declared image serialization and the default output cap. The native-context exception for OLMo is fixed from complete input-length checks before inference. The protocol and parameter table in Appendix C lists these values. The attempt journal is written before a retry, and the first terminal response is used without selecting among later answers. Temperature zero makes neither hosted responses nor floating-point inference universally reproducible.

Model	Scene	Skel.	Story	Thm.	Sc→St	Sc→Th	St→Th	Family4	Avg
GPT-5.5	77.6	83.7	90.9	88.0	71.9	82.4	82.6	88.1	83.2
GPT-5.2	72.9	77.6	76.9	80.0	43.8	63.8	66.1	81.9	70.4
GPT-6 Astra	85.7	91.8	92.6	92.0	74.3	90.0	81.0	89.0	87.1
Qwen2.5-VL-72B	51.4	61.2	45.5	80.0	38.6	47.1	55.4	76.2	56.9
DeepSeek-V4-Flash-0731	–	–	65.3	84.0	–	–	43.8	–	64.4
Qwen2.5-VL-32B	50.5	73.5	50.4	76.0	33.3	47.6	57.0	72.9	57.6
DeepSeek-Coder-V2-Lite	–	–	16.5	40.0	–	–	27.3	–	27.9
Qwen2.5-14B	–	–	45.5	72.0	–	–	45.5	–	54.3
Qwen2.5-7B	–	–	26.4	64.0	–	–	41.3	–	43.9
Qwen2-VL-7B	31.4	44.9	19.8	60.0	28.6	32.9	35.5	55.7	38.6
Gemma-4-31B	71.9	83.7	81.0	88.0	44.8	67.1	71.1	82.9	73.8
Qwen3-VL-32B	60.5	81.6	63.6	80.0	39.5	54.8	53.7	81.4	64.4
GLM-4.6V	59.0	77.6	50.4	68.0	33.8	50.0	52.9	76.7	58.5
Llama-4-Scout	49.5	65.3	42.1	76.0	31.9	55.7	57.0	73.8	56.4
MiniCPM-V-4.5	44.8	59.2	36.4	72.0	34.8	47.6	38.8	67.1	50.1
InternVL3.5-8B	33.8	49.0	40.5	68.0	36.2	50.0	57.9	65.7	50.1
Kimi-VL-A3B	39.5	51.0	24.0	56.0	30.5	37.1	31.4	59.0	41.1
ERNIE-4.5-VL-28B-A3B	31.0	51.0	16.5	36.0	24.8	41.4	37.2	62.9	37.6
Pixtral-12B	29.0	55.1	36.4	40.0	29.5	39.0	34.7	51.4	39.4
Nemotron-Nano-VL-8B	17.1	26.5	27.3	44.0	32.4	31.4	28.1	58.6	33.2
Random chance	4	4	4	4	25	25	25	25	–

Table 2: Task accuracy (%) for the core display group. Identification uses the full motif menu; matching and Family4 use four candidates. Family4 is identification with the same ordered candidate motifs as scene–story matching. Avg is an equal-task macro over the primary tasks, excluding cue conditions; text and vision macros have different task sets. All reported models are complete; – means not applicable. Strict-JSON sensitivity appears in Table 9.

4 RESULTS AND ANALYSIS

Holding the source scene and the ordered candidate motif identities fixed, and changing only the options from motif descriptions to four complete stories, lowers accuracy for every evaluated vision model under compatibility scoring. Table 10 reports the two task accuracies and their paired difference, whose motif-cluster intervals exclude zero across the completed vision roster. This recognition-matching gap is the first result of the benchmark.

Endpoint recognition can approach its ceiling while exact composition fails. In the finite pilot, a model that identifies the endpoint graph almost perfectly is far less accurate on the labeled composition and intermediate-state questions, while a strong model solves nearly all of them. Table 16 reports the endpoint, composition and intermediate arms side by side, and Table 17 reports the paired intermediate comparison. Naming the endpoint family therefore does not summarize performance on the operations that transform it.

A compatibility score can conceal a response-format failure. Several models earn substantial compatibility credit on the primary tasks and no strict-JSON credit at all, because the correct answer survives only inside a longer reply. Table 9 reports both scores from the same stored responses, so the divergence is a property of the reply format rather than of the answer it carries.

Tables 2 and 8 report per-task accuracy before aggregation, and Table 3 gives the roster and response counts. Every reported model completed its applicable suites and passed the response-integrity checks, and a dash marks a task that does not apply to a text-only model. Those checks establish that a reply arrived and was parsed, not that an item’s intended motif is correct.

4.1 WHAT THE REPORTED METRIC COUNTS

Accuracy is the fraction of administered items whose credited response matches the frozen key. Compatibility accuracy credits an answer recovered by the fixed extraction procedure, while strict accuracy credits only a single JSON object carrying the requested field and a valid answer. Both are computed from the same retained replies, so they differ in the response rule and not in the underlying

attempt. A missing network response is never counted as an observed wrong answer. The tasks differ in menu arity, so we also report the uniform-chance normalization

$$\kappa_{m,t} = \frac{a_{m,t} - 1/K_t}{1 - 1/K_t}, \quad \bar{a}_m = \frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} a_{m,t}, \quad (5)$$

where $a_{m,t}$ is model m 's accuracy on task t and K_t is the menu size of t . This is uniform-chance normalization rather than empirical-marginal Cohen's kappa, and it adjusts the reference floor without equating task difficulty or information content. Appendix Tables 12 and 13 report the normalization separately by model scope.

The columns ask related but distinct questions, because a motif label with a short definition offers different evidence from a full story or a masked mathematical statement. We therefore report task accuracies before forming a macro average, and the per-task results rather than one pooled leaderboard show where a model succeeds. Appendix F reports a descriptive logit fit that adjusts for model identity across the vision roster.

4.2 MATCHING IS HARDER THAN RECOGNITION

The paired comparison holds the picture and the ordered candidate motif identities fixed and varies only what the four options present. Identification offers motif names with structure descriptions, whereas matching offers the corresponding complete stories. Equation (2) averages the within-scene correctness differences, so a positive value means lower accuracy with story options. The gap is positive for every completed vision model, and its motif-cluster intervals lie above zero, which is the pattern reported in Table 10 next to the two task accuracies. The pairing holds the image pool and the uniform guessing probability fixed, while candidate length, semantic explicitness and required reading still differ from the motif-description menu.

Two further conditions test how much of this pattern depends on the administered cues. The image-withheld condition drops the scene and keeps the candidates, and the rotation condition keeps all content while cyclically moving the candidate positions. Their signed differences are the original minus the manipulated form, computed over the same source pairs. **The image matters for some models and not for others.** GPT-5.5 is more accurate with the image present than with it withheld, and its motif-cluster interval excludes zero, whereas the corresponding interval for Qwen2.5-VL-72B includes zero. The comparisons therefore do not support a uniform statement that models either use or ignore the picture. **One model reverses under a pure answer-order change.** The original ordering is less accurate than its rotation for GLM-4.1V under the motif-cluster analysis, and because the candidate content is identical, that order sensitivity limits how far a raw choice can be read as a stable judgment of correspondence. Table 11 retains the individual signed differences.

A single score also hides the source of an error, since selecting another valid motif, emitting a nonanswer and returning a correct motif inside an invalid format are different observed outcomes. The stored response and its parsing category separate them without assigning an unobserved reasoning failure. The distinction matters when a model with strong compatibility scores fails the requested format often, as the compatibility and strict columns of Table 9 show for the same stored replies. Both rules apply to every model alike, and no model receives its own parser or a further attempt after a terminal response.

4.3 ERROR GEOMETRY IS AN ASSOCIATION ANALYSIS

The declared confusable families and the separately constructed typed-graph edit distances (Gao et al., 2010) are frozen motif-level covariates. Let $\mathcal{F}$ contain the possibly overlapping confusable-family sets. For distinct motifs, the family dissimilarity is

$$D_{ij}^F = 1 - \mathbf{1}\{\exists F \in \mathcal{F} : i, j \in F\}, \quad i \neq j, \quad D_{ii}^F = 0. \quad (6)$$

This binary score tests direct co-membership rather than overlap through a third motif, and it is a neighborhood proxy rather than a metric. Fresh DINOv2 (Oquab et al., 2023), CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) features are extracted from the exact JPEGs administered in the evaluation, and for each completed vision model the directed error counts are related to standardized formal and perceptual similarities with a Poisson working mean and a source-class exposure offset.

Motifs that share a declared family attract errors, and for several models so do motifs that look alike. In the completed DINOv2 fits the family coefficient is positive, and the image-similarity term also has positive intervals for several models. Joint motif-node permutations preserve the dyadic organization of the matrix, and Appendix I reports the coefficients and identifies the cases in which an interval or a fit is unavailable. These are fitted associations among the specified predictors rather than a unique internal source of error.

4.4 ENDPOINT RECOGNITION IS NOT EXACT OPERATION

The finite pilot separates recognition of a finite endpoint from execution of an exact labeled operation. Endpoint identification, statement matching and edit description use explicit finite constructions, whereas composition questions require the final labeled edge set even when every option is isomorphic to the same motif. Their accuracies have separate denominators in Table 16. **A model with near-perfect endpoint recognition can fail most exact compositions while another model solves nearly all of them.** GPT-5.5 is more accurate than GPT-5.2 on both exact composition and intermediate-state selection, and the observed successes rule out a blanket claim that these models cannot perform the specified compositions. Table 17 reports the paired comparison whose two chains share a starting graph, a final graph, displayed options and image but have different correct intermediate states. A decision rule restricted to the shared endpoint must return the same label for both members of a pair, so it cannot answer both correctly, and the joint success rate measures how often a model clears that bound. This diagnostic concerns the specified labeled operations, and Appendix G states the optional categorical reading and its limits.

4.5 PRESENTATION CONDITIONS

The auxiliary suite contains nuisance renders, a paired comparison manipulation and a motif-level naturalization ladder. Table 18 and Table 19 report the completed models under the frozen English prompts and declared assets, with ladder levels averaged within motif so that later levels do not gain weight from additional renders. **Naturalization does not produce a monotone profile, and its photorealistic end sits below the bare-skeleton baseline for every reported model.** GPT-5.5 improves at some intermediate levels relative to that baseline before declining at the last, while other models fall away earlier. A naturalized render changes cue visibility, rendering fidelity and semantic ambiguity together, so the level differences are associations within this ladder rather than an isolated effect of surface complexity. Appendix H records the level definitions and the composite-image conditions.

5 CONCLUSION

GEB-BENCH separates motif recognition, cross-voice matching and exact structural edits. The paired results show that identifying a motif need not imply matching its differently expressed realization. Source, answer-order and response-format conditions qualify this pattern, while the finite pilot separates endpoint association from following an operation chain. The benchmark supports reproducible measurement of these distinct outcomes. Construction ambiguity, shared content, generator affinity and the limits of mechanistic interpretation remain explicit in Appendix Q.

STATEMENT OF AI USE

AI systems assisted with scene and story generation, agent-based screening, experimental implementation, data cleaning and reformatting, figure preparation, manuscript drafting and editing, literature organization, methodological feedback, code review, and interpretation of results. Experimental scores were computed from retained model responses using fixed evaluation code. Executable validators and response-integrity audits support the checks described in the paper; these do not constitute independent human validation of unrestricted image or narrative semantics. The authors are responsible for the content, claims and cited sources.

REFERENCES

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL <https://arxiv.org/abs/2502.13923>.
- David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *ICML*, 2018.
- Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. VASR: Visual analogies of situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 241–249, 2023. doi: 10.1609/aaai.v37i1.25096. URL <https://doi.org/10.1609/aaai.v37i1.25096>.
- Mikhail M. Bongard. *Pattern Recognition*. Spartan Books, 1970.
- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.
- B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1979. doi: 10.1214/aos/1176344552. URL <https://doi.org/10.1214/aos/1176344552>.
- Brian Falkenhainer, Kenneth D. Forbus, and Dedre Gentner. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63, 1989. doi: 10.1016/0004-3702(89)90077-5. URL [https://doi.org/10.1016/0004-3702\(89\)90077-5](https://doi.org/10.1016/0004-3702(89)90077-5).
- Brendan Fong and David I. Spivak. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge University Press, 2019. doi: 10.1017/9781108668804. URL <https://doi.org/10.1017/9781108668804>.
- Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. IsoBench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling*, 2024a. URL <https://openreview.net/forum?id=KZd1EErRJ1>.
- Xingyu Fu et al. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024b.
- Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129, 2010. doi: 10.1007/s10044-008-0141-y. URL <https://doi.org/10.1007/s10044-008-0141-y>.
- Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170, 1983.

-
- Dedre Gentner and Laura L. Namy. Comparison in the development of categories. *Cognitive Development*, 14(4):487–513, 1999.
- Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, 2023.
- Felix Hill, Adam Santoro, David Barrett, Ari Morcos, and Timothy Lillicrap. Learning to make analogies by contrasting abstract relational structure. In *ICLR*, 2019.
- Douglas R. Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books, 1979.
- Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31096–31116. Curran Associates, Inc., 2023. doi: 10.52202/075280-1355. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.
- Lawrence Hubert and James Schultz. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29(2):190–241, 1976. doi: 10.1111/j.2044-8317.1976.tb00714.x. URL <https://doi.org/10.1111/j.2044-8317.1976.tb00714.x>.
- Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-HOI: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19056–19065, 2022. doi: 10.1109/CVPR52688.2022.01847. URL https://openaccess.thecvf.com/content/CVPR2022/html/Jiang_Bongard-HOI_Benchmarking_Few-Shot_Visual_Reasoning_for_Human-Object_Interactions_CVPR_2022_paper.html.
- Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.731. URL <https://aclanthology.org/2020.emnlp-main.731/>.
- Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018. URL <https://proceedings.mlr.press/v80/lake18a.html>.
- Martha Lewis and Melanie Mitchell. Evaluating the robustness of analogical reasoning in large language models. *arXiv preprint arXiv:2411.14215*, 2024. URL <https://arxiv.org/abs/2411.14215>.
- Yuan Liu et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024.
- Pan Lu et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.
- Karl Menger. Zur allgemeinen kurventheorie. *Fundamenta Mathematicae*, 10:96–115, 1927.
- Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The ConceptARC benchmark: Evaluating understanding and generalization in the ARC domain. *Transactions on Machine Learning Research*, 2023.
- Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-LOGO: A new benchmark for human-level concept learning and reasoning. In *NeurIPS*, 2020.

-
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL <https://arxiv.org/abs/2304.07193>.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, December 2024. URL https://openaccess.thecvf.com/content/ACCV2024/html/Rahmanzadehgervi_Vision_language_models_are_blind_ACCV_2024_paper.html.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.
- Zhenwei Tang, Difan Jiao, Blair Yang, and Ashton Anderson. SEAM: Semantically equivalent across modalities benchmark for vision-language models. In *Conference on Language Modeling (COLM)*, 2025.
- Damien Teney, Peng Wang, Jiewei Cao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. V-PROM: A benchmark for visual reasoning using visual progressive matrices. In *AAAI*, 2020.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024.
- Peng Wang et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.
- Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023. doi: 10.1038/s41562-023-01659-w. URL <https://www.nature.com/articles/s41562-023-01659-w>.
- Rujie Wu, Xiaojian Ma, Zhenliang Zhang, Wei Wang, Qing Li, Song-Chun Zhu, and Yizhou Wang. Bongard-OpenWorld: Few-shot reasoning for free-form visual concepts in the real world. In *International Conference on Learning Representations*, volume 2024, pp. 20688–20718, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html.

-
- Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik Narasimhan. COLLIE: Systematic construction of constrained text generation tasks. *arXiv preprint arXiv:2307.08689*, 2023. doi: 10.48550/arXiv.2307.08689. URL <https://arxiv.org/abs/2307.08689>.
- Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. In *Advances in Neural Information Processing Systems*, 2024.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.736. URL <https://aclanthology.org/2025.acl-long.736/>.
- Xiang Yue et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.
- Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. RAVEN: A dataset for relational and analogical visual reasoning. In *CVPR*, 2019.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, pp. 169–186. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.
- Jeffrey Zhou et al. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A EXTENDED RELATED WORK

Abstract reasoning and concept learning. ARC (Chollet, 2019), Raven-style matrices (Zhang et al., 2019; Barrett et al., 2018) and Bongard concept problems (Bongard, 1970), including Bongard-LOGO (Nie et al., 2020), probe rule induction and concept learning in controlled visual domains. ConceptARC (Moskvichev et al., 2023) organizes ARC-style tasks around systematically varied concept groups. Natural-image settings include Bongard-HOI, which tests human-object interaction concepts with action-based hard negatives (Jiang et al., 2022), and Bongard-OpenWorld, which tests free-form concept induction from positive and negative real-world examples (Wu et al., 2024). These precedents distinguish concept induction from recognizing and matching the fixed intended motifs in GEB-BENCH.

IsoBench (Fu et al., 2024a) pairs visual problem presentations with isomorphic textual representations, including mathematical encodings in relevant domains, to measure representation-dependent performance. SEAM (Tang et al., 2025) likewise pairs semantically equivalent notations in formal domains. GEB-BENCH instead compares recognition and matching of intended motifs across heterogeneous scenes, narrative form devices, and mathematical associations. These realizations need not be isomorphic encodings of one fully specified problem; the fixed motif label is the evaluation target, and its semantic adequacy remains a construction question.

Analogy and structure mapping. Structure-mapping theory emphasizes correspondences between relations rather than shared attributes (Gentner, 1983). The Structure-Mapping Engine constructs structurally consistent mappings between relational descriptions (Falkenhainer et al., 1989). Visual analogy studies (Hill et al., 2019; Teney et al., 2020) and VASR’s natural-image situation analogies (Bitton et al., 2023) extend this question to visual inputs. Developmental work also studies comparison as a route to category abstraction (Gentner & Namy, 1999). Language-only evaluations report strong zero-shot performance on several analogy tasks in GPT-3 (Webb et al., 2023), while robustness studies identify sensitivity to problem variants, including answer-order effects in story analogies (Lewis & Mitchell, 2024). These results motivate both relational matching and checks on its sensitivity to presentation; they do not establish the mechanism behind GEB-BENCH errors.

Compositional generalization. SCAN (Lake & Baroni, 2018) and COGS (Kim & Linzen, 2020) use controlled training and test splits to evaluate new combinations of familiar commands, words and structures. This is a distinct evaluation axis from consistency across realizations. The current GEB-BENCH collection reuses motifs and realizations across tasks and does not establish generalization to held-out motifs.

Multimodal and cross-representation evaluation. MMMU (Yue et al., 2024), MMBench (Liu et al., 2024) and BLINK (Fu et al., 2024b) assess broad knowledge and perception; MathVista (Lu et al., 2024) targets mathematical reasoning in visual contexts. Targeted probes also expose failures on elementary overlap and contact judgments in the models tested (Rahmanzadehgervi et al., 2024). Winoground pairs images with captions containing the same words in different orders to test visio-linguistic composition (Thrush et al., 2022). SEAM (Tang et al., 2025) instead pairs semantically equivalent textual and visual notations in formal domains such as chess, chemistry, music and graphs. GEB-BENCH relates intended motifs across scenes, story form devices and mathematical associations; surface quantities are unscored nuisance variables, and semantic validity is assessed separately from deterministic scoring.

Candidate cues and image dependence. SugarCrepe identifies language-only biases in compositional distractors and constructs fluent hard negatives with adversarial refinement (Hsieh et al., 2023). MMMU-Pro filters questions answerable by text-only models, expands candidate options and introduces a vision-only input setting (Yue et al., 2025). These studies motivate testing whether candidate content suffices without the image. GEB-BENCH uses paired image-withheld and answer-rotation conditions on its administered items, while retaining the limitations from candidate length and semantic ambiguity.

MathVerse (Zhang et al., 2025) varies the distribution of textual and diagram information across versions of visual mathematical problems to test dependence on each modality. This is a close precedent for varying textual redundancy. GEB-BENCH applies image-withheld and answer-order

conditions to its own fixed candidate sets; these diagnose cue use within the administered protocol rather than certify cue-free structural reasoning.

Interpretability and localization. A parallel line of interpretability work localizes capabilities to sparse units, such as factual knowledge (Dai et al., 2022), task skills (Wang et al., 2022), languages (Tang et al., 2024), or to transportable objects such as function vectors (Todd et al., 2024) and circuits (Conmy et al., 2023; Yao et al., 2024), though causal localization need not indicate where interventions succeed (Hase et al., 2023). GEB-BENCH contributes a test bed where cross-representation correspondence can be studied behaviorally and, in future work, through mechanistic interventions; the present evaluation is behavioral.

Verifiable and constrained generation. Instruction-following evaluation uses programmatically checkable requirements (Zhou et al., 2023); COLLIE specifies compositional generation constraints over words, sentences, paragraphs and passages (Yao et al., 2023). GEB-BENCH applies finite predicates to literary form devices, while retaining semantic clauses that those predicates do not fully certify. Model-based screening supplies another source of evidence, subject to the biases documented in LLM-based evaluation (Zheng et al., 2023). Passing either type of check does not establish unique semantic interpretation.

B MOTIF GALLERY AND READING GROUPS

The appendix specifies the evaluated protocol, complete model results, uncertainty conventions and mathematical scope. Illustrative catalog examples remain distinct from the frozen material used for scoring.

A motif is an abstract structure in the library, and a voice is one of the four ways an item presents it, the descriptive or formal form in either the visual or the language channel. We write x for the single displayed item a model receives, so an item realizes motif m in voice v as $x = F_v(m; \xi)$, where ξ collects the recorded realization choices for that item. The identification convention R_v fixes the question template, the ordered answer menu and the response parser for voice v , and it is not re-estimated from the responses. A model returns the motif $\hat{m} = R_v(x)$, so an item is scored correct when the predicted motif matches the intended one, $\mathbf{1}[\hat{m} = m]$, and every quantity in this appendix sums these per-item indicators over its stated denominator.

C REPRODUCIBILITY AND RELEASE FILES

All reported models completed the applicable suites. Response-integrity validation reproduces item coverage, parsing and summaries from retained replies. Table 3 gives the final scope.

Table 3: Scope of the full evaluated roster. Every applicable suite is complete.

Scope	Evaluated	Validated
Models	35	35
Vision models	25	25
Text models	10	10
Applicable model–suite evaluations	85	85
Response records in complete suites	85,920	85,920
Snapshot (UTC)	2026-09-19 06:48	
Score analysis (UTC)	2026-09-19 06:48:05	

Construction records and model aliases. The retained generators request `gpt-image-2` for initial images and `gpt-image-2:stable` for expansion. The final scene inventory includes skeleton-conditioned image edits, a separate text-to-image branch and flagged deterministic repairs (Table 5). English story records retain `gpt-5.5:stable` as writer and `gpt-5.2:stable` as historical screening model. The client configuration was selected from the local environment, and the records do not establish an immutable per-generation model identity. These are recorded aliases, not independent attestations of the returned model version.

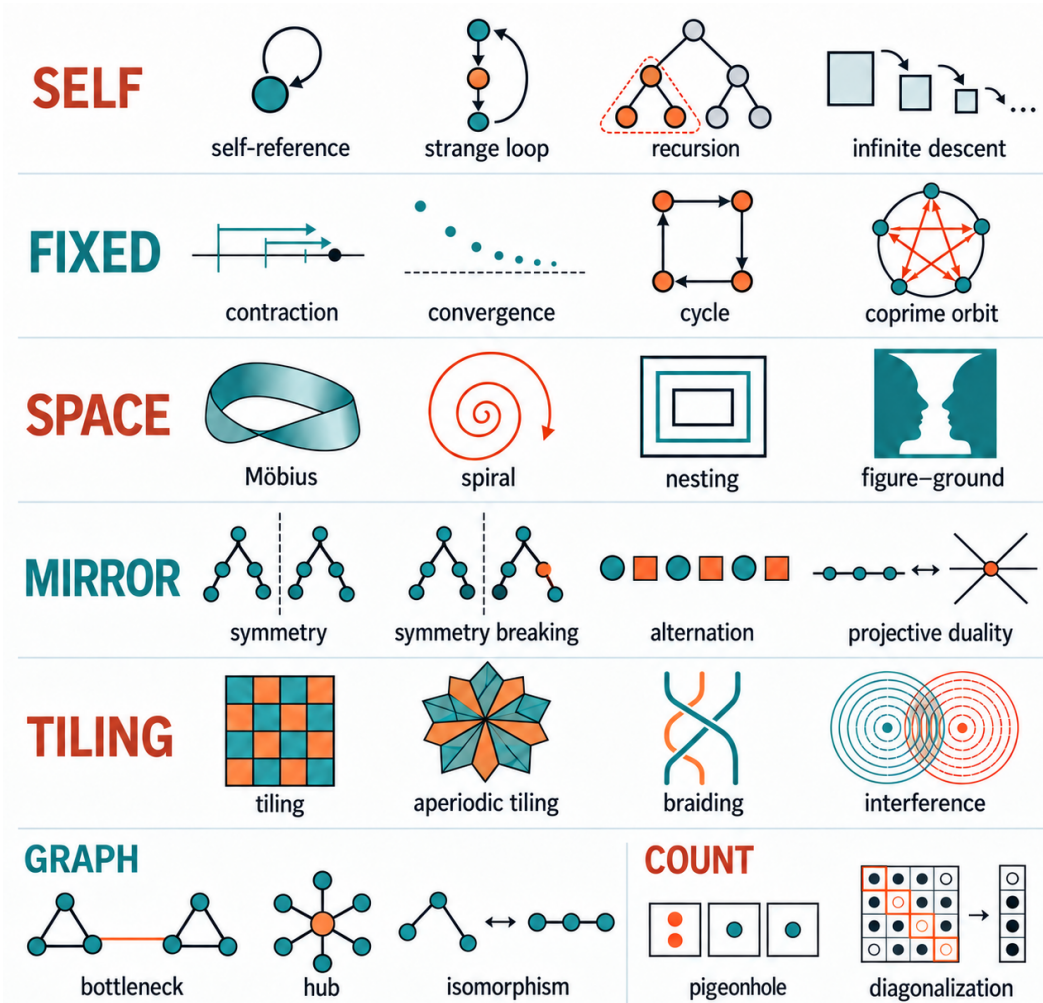


Figure 3: The 25 motifs in seven conceptual groups. Each glyph illustrates a structural feature; a finite picture need not certify a full mathematical invariant. The groups are a reading aid, while the scored label is the individual motif. Table 7 maps names and display order to the frozen scored IDs. Detailed signatures and mathematical associations appear in Table 29.

Device judges receive the intended device specification. Story classifiers receive the story, full motif menu and disambiguation notes without a separately marked answer; scene invariant checks are label-informed. These model screens are not independent human adjudication. Shared naming families between generator, judge and evaluated models can introduce affinity that this evaluation does not isolate.

Counts, names and sampling. The item inventory uses the retained available assets, not balanced random quotas. Table 6 gives the class counts. Each motif also has one theorem-association identification item. The rare classes therefore receive less weight in item-averaged accuracy; equal-task macro scores do not rebalance motifs within a task. Matching reuses these source images, and there is no disjoint train/test or unseen-motif split.

The menu and catalog share an order; the gallery and skeleton gallery share a different conceptual-arc order. The seven taxonomy groups are reading aids, not scored superclasses or learned clusters. They are distinct from the possibly overlapping confusable-family sets in Equation (6). Family4 names a four-candidate motif-ID task, not a prediction of either kind of group. Table 7 fixes the correspondence. In particular, `symmetry` is Mirror Palindrome in the catalog,

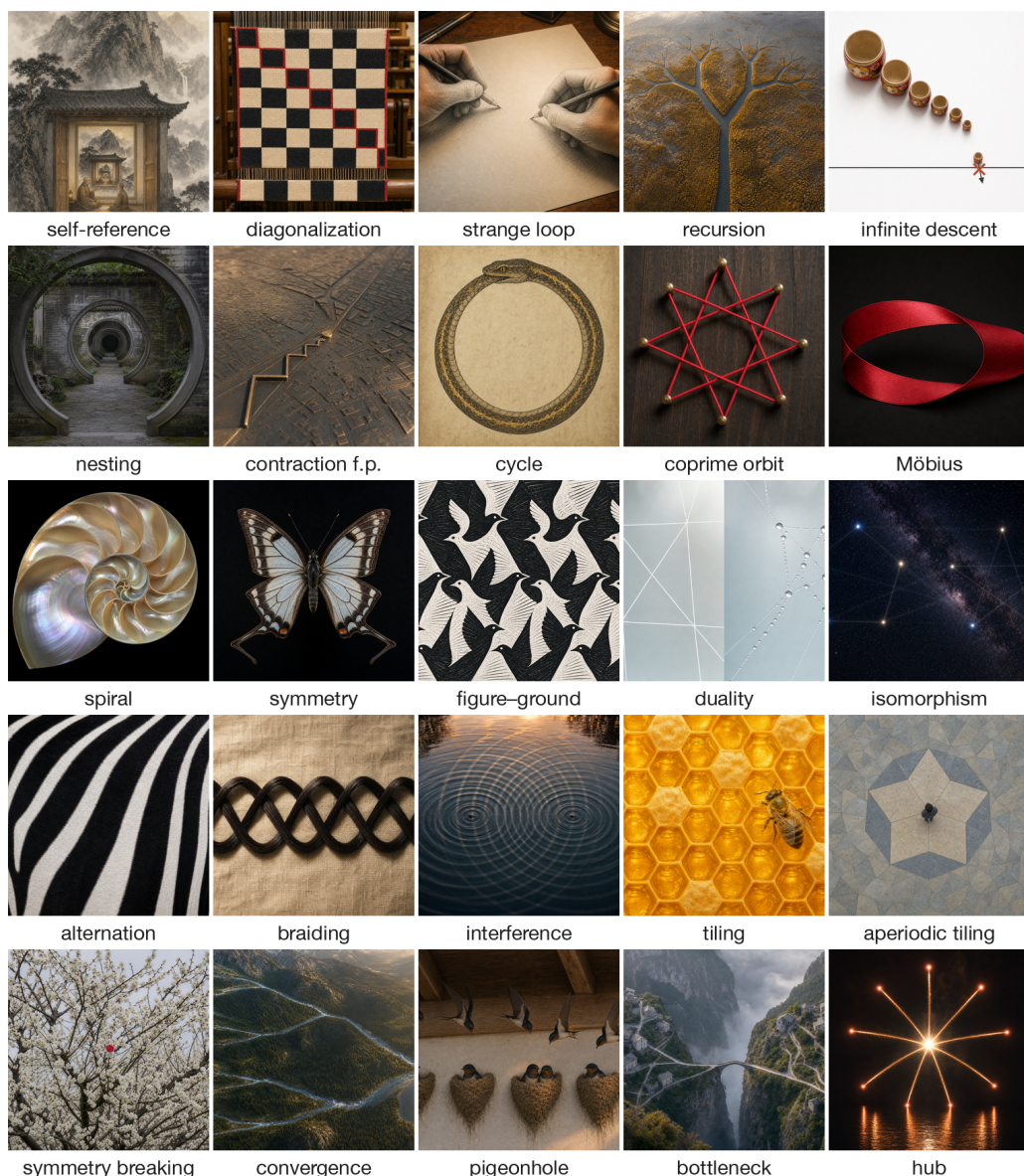


Figure 4: The 25 motifs of GEB-BENCH, each shown through one generated scene. Row by row, the scenes show self-reference, diagonalization, strange loop, recursion, infinite descent; nesting, contraction, cycle, coprime orbit, Möbius; spiral, symmetry, figure-ground, duality, isomorphism; alternation, braiding, interference, tiling, aperiodic tiling; symmetry breaking, convergence, pigeon-hole, bottleneck, hub. Each image is intended to depict its motif; medium, palette, and domain are unscored nuisance variables.

figure_ground is Figure and Ground, projective_duality is also called duality, and contraction_fixed_point appears as contraction or contraction f.p. These labels denote the same frozen IDs.

Frozen data and admission. All story bodies and titles are supplied without truncation wherever a story is used. The construction ledger records selected executable clauses, repairs and whitespace deduplication. It excludes duplicate story bodies and retains historical judge outcomes as metadata, without using that judge’s answer as the corrected inclusion rule. Visual repairs are recorded with before/after hashes and style flags. The intended-motif collection retains ambiguous and formally

Table 4: Frozen task and inference constants. Cue conditions are outside the primary macro. The total text core includes the applicable option-only conditions.

Quantity	Value
Core identification and matching	946
Paired Family4 identification	210
Image-withheld and rotation conditions	630
Full vision core / primary-score items	1,786 / 1,156
Full text core / primary-score items	687 / 267
Corrected finite pilot / auxiliary suite	696 / 680
Vision-model total	3,162
Default output-token cap / OLMo exception	4,096 / 1,353
Temperature / client workers	0 / 4
Maximum image side / JPEG quality	896 px / 88
Motif bootstrap draws / geometry image draws	10,000 / 2,000
Joint motif-node QAP draws	10,000

Retained construction record	Count
Expansion skeleton-conditioned scene edits	161
Initial skeleton-conditioned scene edits	30
Initial text-to-image scenes	14
Deterministic scene repairs	5
Final scene inventory	210
Additional repaired self-reference skeletons	3
Source English story records	125
Admitted unique English stories	121
Admitted stories with deterministic local edits	14
Admitted stories with historical classifier pass / fail	117 / 4

Table 5: Construction records joined to the frozen inventory. Recorded generation aliases do not reconstruct missing per-call provenance. Historical classifier outcomes do not define the admission rule.

uncertifiable images as labeled sensitivity views rather than claiming universal semantic certification.

Matching distractor stories minimize a character-length gap within the fixed motif candidate set, with no truncation or padding. This is a best-effort heuristic rather than uniform length balance.

Candidate-set length diagnostic	Value
Scene matching candidate sets	210
Sets within maximum/minimum length ratio two	39
Menu-description whitespace-word count, min / median / max	18 / 33 / 194

Different description lengths and remaining lexical cues can affect recognition and matching.

The historical generation and screening funnel is not a new experimental result. Prompt counts, rejected renders and retained scenes are different units; without the complete generation manifest they do not define a rejection rate. No correlation between old screening rejections and old model scores is carried into this revision. Selected executable predicates do not certify all unrestricted language semantics, and historical model-based screening is not an independent human validation set.

Frozen chronology and implementation corrections. The response policy distinguishes received model outcomes from transport failures. Its newer runner saves complete wire responses and parsed objects before retry; explicit refusal or fully evidenced budget exhaustion is a completed incorrect nonanswer. A visible but unparseable answer is also incorrect. Generic empty content or missing

Motif ID	Scenes	Skeletons	Stories
self_reference	10	3	5
recursion	10	3	5
nesting	10	3	5
cycle	9	3	5
symmetry	10	1	5
alternation	10	3	5
spiral	10	3	5
convergence	10	1	5
bottleneck	10	1	5
symmetry_breaking	10	3	5
strange_loop	3	3	1
figure_ground	6	3	5
isomorphism	10	1	5
mobius	10	1	5
braiding	8	3	5
interference	10	1	5
tiling	10	3	5
hub	10	3	5
diagonalization	5	1	5
infinite_descent	6	1	5
coprime_orbit	10	1	5
pigeonhole	5	1	5
projective_duality	8	1	5
contraction_fixed_point	9	1	5
aperiodic_tiling	1	1	5

Table 6: Items per voice in scene, skeleton and story identification. The totals are 210, 49 and 121. Multiple skeleton variants exist only for some motifs, and story deduplication leaves one `strange_loop` story. The unequal counts are retained in the declared evaluation.

responses remain incomplete. The publication tables use the first terminal outcome and do not select a later favorable answer.

The entire preliminary constant-answer pilot batch and the two affected old API core batches are quarantined. A demonstrated rotary-position implementation mismatch in the earlier GLM-4.1V build led to a complete fresh model rerun with a verified native implementation. None of those quarantined observations contributes to the tables. The final roster includes author-directed and availability-driven amendments made after other models had started; it was not preregistered before all inference. Inputs, parsing rules and output budgets remain fixed, while each replacement model’s identity and configuration is bound before its first formal request. Amendments retain their chronology and prior plans; they are not justified by choosing favorable observed scores.

The roster contains 31 locally evaluated checkpoint identities and four hosted API identities. The GPT-5.5 and GPT-5.2 rows use the dated identifiers `gpt-5.5-2026-04-23` and `gpt-5.2-2025-12-11`. The GPT-6 Astra row retains the requested and returned identifier `gpt-6-astra`; this undated API label does not independently establish backend vendor, architecture or parameter count. The cancelled Claude and Kimi K2.5 benchmarks are absent from the current roster; Kimi-VL-A3B is a separate completed model. Kimi K2.5 partial responses remain archived outside the scores. The evaluated DeepSeek row uses the distinct hosted identifier `deepseek-v4-flash-0731` in place of DeepSeek-V3.2. The unavailable Gemini identity is replaced by the distinct hosted identifier `gpt-6-astra`. These pre-inference replacements inherit no responses or scores from the removed models. Moving an unstarted native model to a different deployment requires fresh source, implementation and process evidence. It does not authorize silently replacing the model with a different API version.

Persistence and exact reproduction. Responses, request hashes, image hashes, attempt journals, model provenance and analysis receipts are stored on persistent storage. A storage migration preserved their exact bytes and resumed existing evaluations with unchanged fingerprints. Model-weight caches are separate from the authoritative response archive. Reproducing a score requires

Scored ID	Catalog name	Menu	Gallery	Group	Taxonomy
self.reference	Self-Reference	1	1	SELF	1
recursion	Recursion	2	4	SELF	3
nesting	Nesting	3	6	SPACE	11
cycle	Cycle	4	8	FIXED	7
symmetry	Mirror Palindrome	5	12	MIRROR	13
alternation	Alternation	6	16	MIRROR	15
spiral	Spiral	7	11	SPACE	10
convergence	Convergence	8	22	FIXED	6
bottleneck	Bottleneck	9	24	GRAPH	21
symmetry_breaking	Symmetry Breaking	10	21	MIRROR	14
strange_loop	Strange Loop	11	3	SELF	2
figure_ground	Figure and Ground	12	13	SPACE	12
isomorphism	Isomorphism	13	15	GRAPH	23
mobius	Möbius	14	10	SPACE	9
braiding	Braiding	15	17	TILING	19
interference	Interference: superposed pe- riodic schedules or signals	16	18	TILING	20
tiling	Tiling	17	19	TILING	17
hub	Hub	18	25	GRAPH	22
diagonalization	The Diagonal Escape	19	2	COUNT	25
infinite_descent	Infinite Descent	20	5	SELF	4
coprime_orbit	Coprime Orbit	21	9	FIXED	8
pigeonhole	Pigeonhole	22	23	COUNT	24
projective_duality	The Involutive Role-Swap	23	14	MIRROR	16
contraction_fixed_	The Shrinking Return	24	7	FIXED	5
point					
aperiodic_tiling	Aperiodic Tiling	25	20	TILING	18

Table 7: Frozen IDs, display names and one-based positions in the menu, gallery and taxonomy. Gallery also gives the skeleton-gallery order; Group is the conceptual reading group shown in the taxonomy figure.

Model	Scene	Skel.	Story	Thm.	Sc→St	Sc→Th	St→Th	Family4	Avg
GLM-4.1V-9B	46.2	67.3	47.9	72.0	29.5	51.4	49.6	80.0	55.5
GLM-4.6V-Flash	45.2	73.5	43.0	64.0	33.3	43.3	43.0	72.4	52.2
Qwen3-VL-8B	50.0	69.4	38.8	68.0	35.2	50.5	47.9	73.8	54.2
Qwen3-VL-4B	42.4	75.5	33.1	68.0	39.5	47.6	43.0	71.0	52.5
Qwen3-VL-2B	31.0	61.2	24.0	40.0	28.6	32.4	28.9	62.9	38.6
Gemma-3-27B	46.2	57.1	36.4	64.0	35.7	45.7	43.8	67.6	49.6
Gemma-3-12B	33.3	44.9	36.4	44.0	34.3	42.4	41.3	63.8	42.5
InternVL3-8B	30.5	36.7	33.1	76.0	31.0	42.9	40.5	66.7	44.7
MiniCPM-V-4	22.4	34.7	14.9	40.0	29.0	37.6	35.5	52.9	33.4
GLM-4-9B	–	–	36.4	60.0	–	–	38.8	–	45.1
Qwen3-4B	–	–	36.4	68.0	–	–	43.0	–	49.1
Gemma-2-9B	–	–	29.8	60.0	–	–	36.4	–	42.0
Llama-3.1-8B	–	–	19.0	36.0	–	–	35.5	–	30.2
Mistral-7B-v0.3	–	–	12.4	52.0	–	–	30.6	–	31.7
OLMo-2-7B	–	–	8.3	4.0	–	–	30.6	–	14.3
Random chance	4	4	4	4	25	25	25	25	–

Table 8: Remaining roster rows, with the same primary task definitions as Table 2. Text and vision macros average different applicable task sets.

the frozen input and recorded reply, not a new image generation or another hosted API call. Retained audit records map each displayed number to its source analysis row and underlying result-file hash; the appendix specifies the protocol and the limits of reproducing API-served replies.

Uncertainty and dependence. Task and paired intervals resample whole intended-motif clusters. Shared source objects, correct-story clusters and connected components through all shared candidate content supply separate sensitivity groupings. For the full scene–story candidate graph, the latter grouping has one connected component, so its interval is not estimable. A motif-cluster interval does not eliminate every dependency caused by repeated content. We leave unavailable bounds empty rather than producing a zero-width interval. Exact McNemar and Holm-adjusted values are descriptive under these dependencies. Tables therefore emphasize effect estimates and declared cluster intervals rather than treating reused content as independent experimental replication.

D COMPLETED-MODEL RESULTS

Tables 2 and 8 enumerate the full evaluated roster. Their grouping preserves display organization rather than ranking models by a shared text-and-vision aggregate.

Task accuracies use integer correct counts and denominators. Macros, chance-normalized scores and paired contrasts are computed from unrounded values; only the final displayed values are rounded. Averaging displayed task columns or subtracting displayed percentages can therefore differ from a displayed aggregate or contrast in its last digit.

Model	Scope	Compatibility [motif CI]	Strict JSON
GPT-5.5	Vision	83.2 [75.0, 90.2]	83.2
GPT-5.2	Vision	70.4 [61.8, 78.1]	70.4
GPT-6 Astra	Vision	87.1 [80.2, 93.2]	87.1
Qwen2.5-VL-72B	Vision	56.9 [48.6, 64.6]	56.9
DeepSeek-V4-Flash-0731	Text	64.4 [51.6, 76.0]	64.4
Qwen2.5-VL-32B	Vision	57.6 [49.6, 65.0]	0.0
DeepSeek-Coder-V2-Lite	Text	27.9 [19.8, 37.1]	25.7
Qwen2.5-14B	Text	54.3 [43.4, 64.8]	54.3
Qwen2.5-7B	Text	43.9 [33.4, 54.7]	43.9
Qwen2-VL-7B	Vision	38.6 [32.2, 44.9]	37.9
Gemma-4-31B	Vision	73.8 [66.2, 80.7]	73.6
Qwen3-VL-32B	Vision	64.4 [56.2, 72.1]	64.4
GLM-4.6V	Vision	58.5 [50.7, 65.9]	7.1
Llama-4-Scout	Vision	56.4 [48.1, 63.7]	20.2
MiniCPM-V-4.5	Vision	50.1 [42.5, 57.0]	50.0
InternVL3.5-8B	Vision	50.1 [43.1, 56.5]	10.7
Kimi-VL-A3B	Vision	41.1 [33.7, 48.5]	40.6
ERNIE-4.5-VL-28B-A3B	Vision	37.6 [30.1, 44.7]	0.0
Pixtral-12B	Vision	39.4 [32.1, 45.9]	5.6
Nemotron-Nano-VL-8B	Vision	33.2 [27.6, 39.0]	31.7
GLM-4.1V-9B	Vision	55.5 [48.1, 62.3]	0.0
GLM-4.6V-Flash	Vision	52.2 [44.3, 59.3]	51.9
Qwen3-VL-8B	Vision	54.2 [46.3, 61.4]	54.2
Qwen3-VL-4B	Vision	52.5 [45.4, 59.1]	52.5
Qwen3-VL-2B	Vision	38.6 [31.2, 46.3]	24.8
Gemma-3-27B	Vision	49.6 [41.3, 57.1]	49.2
Gemma-3-12B	Vision	42.5 [33.9, 51.1]	23.4
InternVL3-8B	Vision	44.7 [36.1, 52.7]	38.3
MiniCPM-V-4	Vision	33.4 [25.3, 41.6]	30.2
GLM-4-9B	Text	45.1 [32.3, 57.6]	45.1
Qwen3-4B	Text	49.1 [38.9, 59.7]	49.1
Gemma-2-9B	Text	42.0 [31.1, 53.4]	39.8
Llama-3.1-8B	Text	30.2 [19.3, 41.6]	30.2
Mistral-7B-v0.3	Text	31.7 [23.5, 40.2]	29.5
OLMo-2-7B	Text	14.3 [7.5, 23.1]	14.3

Table 9: Compatibility-parser and strict-JSON primary macro accuracy (%). Brackets are 95% whole-motif bootstrap intervals for compatibility scoring. A strict zero is an observed formatting result under the strict parser; it does not mean every compatibility answer is wrong.

Model	Family4	Sc→St	Δ [motif CI]
GPT-5.5	88.1	71.9	16.2 [4.6, 28.8]
GPT-5.2	81.9	43.8	38.1 [25.7, 50.0]
GPT-6 Astra	89.0	74.3	14.8 [5.4, 25.2]
Qwen2.5-VL-72B	76.2	38.6	37.6 [21.5, 52.9]
Qwen2.5-VL-32B	72.9	33.3	39.5 [23.2, 55.6]
Qwen2-VL-7B	55.7	28.6	27.1 [15.3, 38.8]
Gemma-4-31B	82.9	44.8	38.1 [23.1, 52.5]
Qwen3-VL-32B	81.4	39.5	41.9 [27.4, 56.0]
GLM-4.6V	76.7	33.8	42.9 [23.0, 61.9]
Llama-4-Scout	73.8	31.9	41.9 [25.7, 56.9]
MiniCPM-V-4.5	67.1	34.8	32.4 [18.4, 45.9]
InternVL3.5-8B	65.7	36.2	29.5 [13.7, 44.3]
Kimi-VL-A3B	59.0	30.5	28.6 [15.8, 41.6]
ERNIE-4.5-VL-28B-A3B	62.9	24.8	38.1 [23.4, 51.6]
Pixtral-12B	51.4	29.5	21.9 [5.3, 38.5]
Nemotron-Nano-VL-8B	58.6	32.4	26.2 [12.0, 39.5]
GLM-4.1V-9B	80.0	29.5	50.5 [36.8, 64.0]
GLM-4.6V-Flash	72.4	33.3	39.0 [22.2, 55.1]
Qwen3-VL-8B	73.8	35.2	38.6 [22.4, 54.2]
Qwen3-VL-4B	71.0	39.5	31.4 [14.0, 47.3]
Qwen3-VL-2B	62.9	28.6	34.3 [16.0, 50.2]
Gemma-3-27B	67.6	35.7	31.9 [16.3, 46.9]
Gemma-3-12B	63.8	34.3	29.5 [12.7, 45.3]
InternVL3-8B	66.7	31.0	35.7 [22.1, 49.3]
MiniCPM-V-4	52.9	29.0	23.8 [6.5, 39.4]

Table 10: Paired Family4 identification minus scene–story matching. Differences are computed before display rounding. Accuracies and differences are in percent and percentage points, respectively; brackets are 95% motif-cluster intervals over the same 210 sources. All completed vision models have a positive point estimate. The fully shared-content interval is unavailable because its graph has one connected component.

Model	Δ_{story} [motif CI]	Δ_{theorem} [motif CI]	Δ_{order} [motif CI]
GPT-5.5	47.1 [26.0, 65.7]	65.7 [52.0, 78.0]	1.9 [-2.3, 6.2]
GPT-5.2	18.6 [1.3, 34.7]	37.6 [18.3, 56.5]	2.4 [-4.2, 9.2]
GPT-6 Astra	50.0 [31.9, 65.8]	68.6 [51.2, 83.8]	0.5 [-2.4, 3.2]
Qwen2.5-VL-72B	8.1 [-6.3, 22.1]	26.7 [10.7, 42.0]	2.9 [-1.4, 7.5]
Qwen2.5-VL-32B	7.1 [-6.0, 21.0]	21.9 [4.1, 39.0]	-1.9 [-7.9, 4.8]
Qwen2-VL-7B	3.3 [-1.0, 7.6]	10.0 [1.0, 19.1]	1.9 [-3.0, 6.6]
Gemma-4-31B	20.0 [5.1, 34.7]	42.4 [24.2, 58.9]	1.4 [-4.3, 7.4]
Qwen3-VL-32B	12.9 [0.4, 25.4]	27.1 [10.6, 42.8]	-0.5 [-6.8, 5.6]
GLM-4.6V	4.3 [-9.4, 18.5]	27.6 [9.5, 45.0]	-1.4 [-6.8, 4.3]
Llama-4-Scout	4.8 [-4.2, 13.2]	29.0 [13.8, 43.6]	0.5 [-3.8, 4.8]
MiniCPM-V-4.5	7.1 [-5.3, 19.6]	28.6 [15.0, 42.4]	-2.9 [-9.3, 3.7]
InternVL3.5-8B	9.0 [-1.5, 18.7]	22.9 [7.9, 36.9]	3.3 [-2.9, 8.8]
Kimi-VL-A3B	6.7 [-4.6, 19.3]	13.8 [1.9, 26.8]	-0.5 [-7.6, 7.1]
ERNIE-4.5-VL-28B-A3B	2.9 [-5.4, 11.0]	17.6 [1.5, 32.3]	-5.2 [-13.1, 2.7]
Pixtral-12B	4.8 [-5.1, 14.9]	15.2 [-1.5, 31.3]	1.0 [-5.6, 7.7]
Nemotron-Nano-VL-8B	7.1 [0.4, 14.3]	11.0 [-1.0, 22.5]	1.9 [-3.8, 7.6]
GLM-4.1V-9B	4.3 [-5.8, 15.3]	23.8 [7.3, 40.7]	-8.1 [-13.3, -2.7]
GLM-4.6V-Flash	5.7 [-3.6, 15.8]	17.1 [3.8, 30.4]	1.0 [-5.3, 6.7]
Qwen3-VL-8B	8.6 [-0.5, 18.5]	31.0 [14.1, 46.6]	2.9 [-1.9, 8.0]
Qwen3-VL-4B	9.5 [-4.7, 24.1]	21.4 [4.9, 37.2]	1.4 [-5.5, 8.8]
Qwen3-VL-2B	-0.5 [-9.1, 8.3]	5.7 [-5.8, 18.1]	2.9 [-2.5, 8.3]
Gemma-3-27B	10.0 [-1.7, 22.0]	22.4 [5.6, 38.7]	-2.9 [-7.8, 2.4]
Gemma-3-12B	7.6 [-4.0, 20.2]	21.9 [7.7, 35.9]	-2.4 [-9.8, 4.5]
InternVL3-8B	8.1 [0.5, 16.0]	24.3 [5.9, 42.9]	1.0 [-5.2, 6.8]
MiniCPM-V-4	1.4 [-6.6, 9.9]	12.4 [-1.6, 27.0]	-1.4 [-8.7, 5.8]

Table 11: Source and order conditions (percentage-point differences, 95% motif-cluster intervals). Story and theorem are image-present minus image-withheld performance. Order is original minus a one-position left rotation. An interval spanning zero is not an equivalence test.

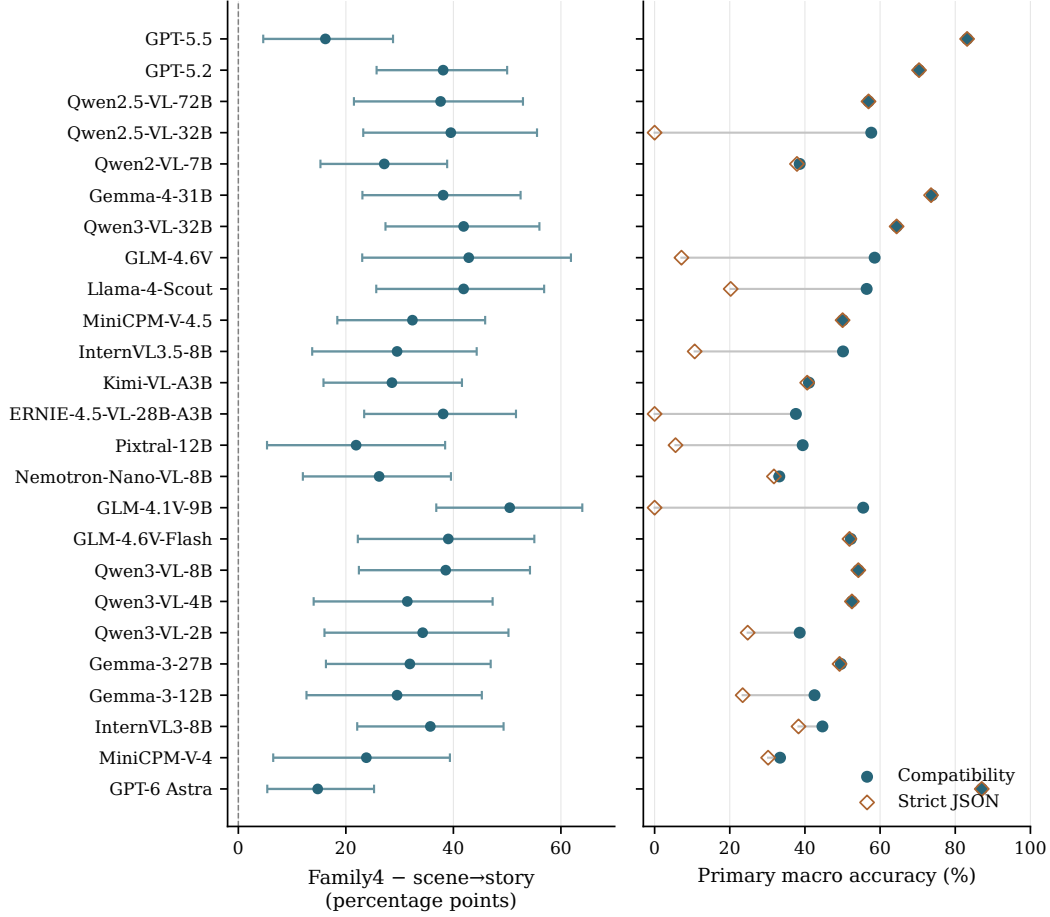


Figure 5: Completed-model paired mapping contrasts (left) and primary compatibility versus strict-JSON accuracy (right). The left panel uses the same motif-cluster intervals as Table 10. The right panel includes completed vision models only, so task sets are comparable. All plotted values come from the frozen evaluated collection.

E UNIFORM-CHANCE-NORMALIZED SCORES

Tables 12 and 13 report $\kappa = (a - 1/K)/(1 - 1/K)$ for each primary task and its equal-task average. This average is not obtained by chance-correcting pooled accuracy and is not empirical-marginal Cohen’s kappa. Text and vision models retain their different task sets. Negative scores below the uniform-guessing reference are not clipped. Every ladder level uses the same motif menu, so its chance normalization is $\kappa = (a - 0.04)/0.96$; the ladder figure displays accuracy rather than κ .

Model	Scene	Skel.	Story	Thm.	S–St	S–T	St–T	Family	Avg.
GPT-5.5	0.767	0.830	0.905	0.875	0.625	0.765	0.769	0.841	0.797
GPT-5.2	0.717	0.766	0.759	0.792	0.251	0.517	0.548	0.759	0.639
GPT-6 Astra	0.851	0.915	0.923	0.917	0.657	0.867	0.747	0.854	0.841
Qwen2.5-VL-72B	0.494	0.596	0.432	0.792	0.181	0.295	0.405	0.683	0.485
Qwen2.5-VL-32B	0.484	0.724	0.483	0.750	0.111	0.302	0.427	0.638	0.490
Qwen2-VL-7B	0.286	0.426	0.165	0.583	0.048	0.105	0.140	0.410	0.270
Gemma-4-31B	0.707	0.830	0.802	0.875	0.263	0.562	0.614	0.771	0.678
Qwen3-VL-32B	0.588	0.809	0.621	0.792	0.194	0.397	0.383	0.752	0.567
GLM-4.6V	0.573	0.766	0.483	0.667	0.117	0.333	0.372	0.689	0.500
Llama-4-Scout	0.474	0.639	0.397	0.750	0.092	0.410	0.427	0.651	0.480
MiniCPM-V-4.5	0.425	0.575	0.337	0.708	0.130	0.302	0.185	0.562	0.403
InternVL3.5-8B	0.311	0.469	0.380	0.667	0.149	0.333	0.438	0.543	0.411
Kimi-VL-A3B	0.370	0.490	0.208	0.542	0.073	0.162	0.085	0.454	0.298
ERNIE-4.5-VL-28B-A3B	0.281	0.490	0.131	0.333	-0.003	0.219	0.163	0.505	0.265
Pixtral-12B	0.261	0.532	0.337	0.375	0.060	0.187	0.129	0.352	0.279
Nemotron-Nano-VL-8B	0.137	0.235	0.242	0.417	0.098	0.086	0.041	0.448	0.213
GLM-4.1V-9B	0.439	0.660	0.458	0.708	0.060	0.352	0.328	0.733	0.467
GLM-4.6V-Flash	0.430	0.724	0.406	0.625	0.111	0.244	0.240	0.632	0.426
Qwen3-VL-8B	0.479	0.681	0.363	0.667	0.137	0.340	0.306	0.651	0.453
Qwen3-VL-4B	0.400	0.745	0.303	0.667	0.194	0.302	0.240	0.613	0.433
Qwen3-VL-2B	0.281	0.596	0.208	0.375	0.048	0.098	0.052	0.505	0.270
Gemma-3-27B	0.439	0.554	0.337	0.625	0.143	0.276	0.251	0.568	0.399
Gemma-3-12B	0.306	0.426	0.337	0.417	0.124	0.232	0.218	0.517	0.322
InternVL3-8B	0.276	0.341	0.303	0.750	0.079	0.238	0.207	0.556	0.344
MiniCPM-V-4	0.191	0.320	0.113	0.375	0.054	0.168	0.140	0.371	0.217

Table 12: Uniform-chance-normalized primary task scores for vision models. The first four identification tasks use the full motif menu; the four matching/Family4 tasks use four options. Avg is the equal-task mean.

Model	Story	Theorem	Story–theorem	Avg.
DeepSeek V4 Flash (0731)	0.638	0.833	0.251	0.574
DeepSeek-Coder-V2-Lite	0.131	0.375	0.030	0.179
Qwen2.5-14B	0.432	0.708	0.273	0.471
Qwen2.5-7B	0.234	0.625	0.218	0.359
GLM-4-9B	0.337	0.583	0.185	0.368
Qwen3-4B	0.337	0.667	0.240	0.414
Gemma-2-9B	0.268	0.583	0.152	0.334
Llama-3.1-8B	0.156	0.333	0.140	0.210
Mistral-7B-v0.3	0.087	0.500	0.074	0.221
OLMo-2-7B	0.044	0.000	0.074	0.040

Table 13: Uniform-chance-normalized primary scores for text-only models. These averages concern a different task set from the vision table.

F DESCRIPTIVE TASK REGRESSION

For correctness Y_{mit} on item i , task t and model m , the fixed-roster association model is

$$\text{logit } \Pr(Y_{mit} = 1) = \alpha_m + \beta_t + o_t.$$

The reference task is theorem identification. We report $o_t = 0$ and $o_t = \text{logit}(1/K_t)$, where K_t is task arity. Table 14 reports compatibility-score task coefficients with motif-cluster and shared-source-cluster intervals. The fit uses all primary responses from the vision roster; repeated models and shared source items are not independent replications. Few motif clusters limit asymptotic inference. The offset changes the reference floor rather than equating task difficulty.

Strict-JSON scores remain in Table 9. Model groups with no strict successes cause complete or quasi-separation in the corresponding logit fits. We therefore do not interpret their finite numerical coefficients, Wald intervals or p-values as reliable inference.

Task	Raw / motif	Raw / source	Chance / motif	Chance / source
Scene	-0.87 [-1.64, -0.09]	-0.87 [-1.45, -0.28]	-0.87 [-1.64, -0.09]	-0.87 [-1.45, -0.28]
Skeleton	-0.16 [-0.90, 0.57]	-0.16 [-0.85, 0.53]	-0.16 [-0.90, 0.57]	-0.16 [-0.85, 0.53]
Story	-0.96 [-1.63, -0.28]	-0.96 [-1.57, -0.35]	-0.96 [-1.63, -0.28]	-0.96 [-1.57, -0.35]
Scene→story	-1.28 [-1.92, -0.63]	-1.28 [-1.86, -0.70]	-3.36 [-4.00, -2.71]	-3.36 [-3.94, -2.78]
Scene→theorem	-0.73 [-1.46, 0.00]	-0.73 [-1.32, -0.14]	-2.81 [-3.53, -2.08]	-2.81 [-3.40, -2.22]
Story→theorem	-0.76 [-1.34, -0.17]	-0.76 [-1.35, -0.17]	-2.84 [-3.42, -2.25]	-2.84 [-3.43, -2.25]
Family4	0.22 [-0.46, 0.91]	0.22 [-0.37, 0.81]	-1.86 [-2.54, -1.17]	-1.86 [-2.45, -1.27]

Table 14: Task coefficients and 95% cluster-robust intervals in the descriptive compatibility logit model. Raw and Chance indicate the absent or chance-logit offset; motif and source identify clustering units. Theorem identification is the reference.

G FINITE OPERATIONS AND INTERMEDIATE STATES

Reading the mathematical organization. An object names a structure being related; a morphism names an allowed operation that changes one structure into another. Composition means performing compatible operations in sequence, and an identity operation changes nothing. A category requires identity laws and associative composition. A functor maps both objects and operations between categories while preserving these laws. These terms motivate the organization of the benchmark, rather than impose extra assumptions on its scoring.

For an idealized interpretation, suppose a motif category $\mathcal{C}$ and a realization category $\mathcal{A}$ have been specified, with functors $F_u, F_v : \mathcal{C} \rightarrow \mathcal{A}$ for two voices. An edit $f : m \rightarrow m'$ has realized maps $F_u(f)$ and $F_v(f)$. If $\eta_m : F_u(m) \rightarrow F_v(m)$ changes the voice, naturality requires

$$\eta_{m'} \circ F_u(f) = F_v(f) \circ \eta_m. \quad (7)$$

In words, editing and then changing voice must agree with changing voice and then applying the corresponding edit. The released scene and story pools are not proved to supply these categories, functors, or maps. The finite pilot instead checks explicitly serialized operations. Node relabeling supplies a specified representation change, and the edit’s node labels must change with it. It checks selected finite identities and intermediate targets, not the universal naturality condition.

The corrected pilot has the arm sizes in Table 15. These arms replace the old endpoint/theorem/story-square experiment. No old pilot score is pooled with the corrected pilot. The dated result snapshot includes a read-only replay of the current pilot validator. The older validation receipt predates neutral answer-choice prompts and is retained as history, not as certification of the current inputs.

For a polygon with vertices p_i and edges e_j , incidence is $I(p_i, e_j)$ if and only if $i = j$ or $i = j + 1$ modulo its size. The role swap

$$\sigma(p_i) = e_{c-i}, \quad \sigma(e_j) = p_{c-j}$$

is involutive and preserves incidence: $I(\sigma(e_j), \sigma(p_i))$ is equivalent to $j = i$ or $j = i - 1$, the original condition. The validator checks incidence and non-incidence for every serialized instance. This finite cycle relation does not instantiate a projective plane.

Table 15: Separate denominators for the corrected finite pilot.

Arm	Items
Finite endpoint identification	144
Endpoint statement matching	72
Edit description and nuisance comparison	144
Exact labeled graph composition	144
Paired intermediate-state conditions	192
Total	696

The contraction endpoint uses $f(x) = x^* + qR_\theta(x - x^*)$ on Euclidean space, with orthogonal R_θ and $0 < q < 1$. Hence $\|f(x) - f(y)\| = q\|x - y\|$, and the marked fixed point is x^* rather than the last plotted finite sample. The finite substitution-strip arm is also labeled by its finite construction, not a claim of global aperiodicity.

Composition options have distinct labeled edge sets but the same isomorphism type. Addition/deletion and relabeling supply chains with order-dependent intermediate states, while reversal and relabeling supply the commuting case. Equation (4) explains why a decision rule based only on shared endpoints cannot solve both intermediate questions in a pair. This is a property of the explicit targets, not a claim about model internals.

Model	Endpoint ID	Statement	Edit	Compose	Intermediate
GPT-5.5	98.6 / 98.6	100.0 / 100.0	98.6 / 98.6	97.9 / 97.9	94.8 / 94.8
GPT-5.2	99.3 / 99.3	84.7 / 84.7	88.9 / 88.9	38.9 / 38.9	59.4 / 59.4
GPT-6 Astra	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0
Qwen2.5-VL-72B	83.3 / 0.0	84.7 / 0.0	86.8 / 0.0	21.5 / 0.0	43.8 / 0.0
Qwen2.5-VL-32B	84.0 / 0.0	91.7 / 0.0	76.4 / 0.0	37.5 / 0.0	29.2 / 0.0
Qwen2-VL-7B	25.0 / 25.0	83.3 / 83.3	51.4 / 51.4	19.4 / 0.0	26.6 / 26.6
Gemma-4-31B	100.0 / 0.0	95.8 / 0.0	83.3 / 0.0	68.8 / 0.0	63.5 / 0.0
Qwen3-VL-32B	100.0 / 100.0	94.4 / 2.8	88.9 / 79.9	40.3 / 13.9	53.1 / 0.0
GLM-4.6V	93.8 / 70.1	88.9 / 0.0	86.1 / 0.7	61.8 / 0.0	27.6 / 0.0
Llama-4-Scout	97.2 / 0.0	97.2 / 0.0	79.9 / 0.0	38.9 / 0.0	29.7 / 0.0
MiniCPM-V-4.5	75.7 / 75.7	84.7 / 84.7	55.6 / 55.6	23.6 / 22.9	33.3 / 33.3
InternVL3.5-8B	63.2 / 63.2	45.8 / 0.0	52.1 / 0.0	25.0 / 0.0	26.6 / 0.0
Kimi-VL-A3B	38.2 / 38.2	63.9 / 6.9	43.8 / 6.9	20.8 / 0.0	20.3 / 0.0
ERNIE-4.5-VL-28B-A3B	51.4 / 0.0	65.3 / 0.0	53.5 / 0.0	28.5 / 0.0	13.5 / 0.0
Pixtral-12B	62.5 / 0.0	54.2 / 0.0	66.0 / 0.0	27.1 / 0.0	21.4 / 0.0
Nemotron-Nano-VL-8B	8.3 / 8.3	77.8 / 27.8	38.2 / 0.7	26.4 / 26.4	25.0 / 0.0
GLM-4.1V-9B	81.9 / 0.0	66.7 / 0.0	62.5 / 0.0	35.4 / 0.0	30.2 / 0.0
GLM-4.6V-Flash	70.8 / 61.1	76.4 / 11.1	48.6 / 35.4	38.2 / 0.0	28.6 / 4.7
Qwen3-VL-8B	78.5 / 78.5	97.2 / 97.2	72.2 / 72.2	26.4 / 1.4	19.8 / 15.6
Qwen3-VL-4B	66.0 / 66.0	68.1 / 68.1	71.5 / 71.5	26.4 / 26.4	21.4 / 21.4
Qwen3-VL-2B	17.4 / 0.0	79.2 / 0.0	55.6 / 0.0	16.0 / 1.4	20.3 / 0.0
Gemma-3-27B	80.6 / 0.0	75.0 / 0.0	52.1 / 0.0	27.1 / 0.0	25.0 / 0.0
Gemma-3-12B	64.6 / 0.0	68.1 / 0.0	41.0 / 0.0	28.5 / 0.0	26.0 / 0.0
InternVL3-8B	71.5 / 0.0	66.7 / 0.0	68.1 / 0.0	24.3 / 0.0	31.2 / 0.0
MiniCPM-V-4	9.7 / 0.7	66.7 / 66.7	50.7 / 33.3	24.3 / 5.6	25.5 / 0.0

Table 16: Finite-pilot accuracy (%). Each cell is compatibility / strict-JSON accuracy on the indicated arm. Endpoint association, composition and intermediate-state selection have separate denominators and should not be interpreted as interchangeable measures.

Model	Both correct [motif CI]	Same parsed choice
GPT-5.5	91.7 [87.5, 96.9]	1.0
GPT-5.2	32.3 [14.6, 51.0]	18.8
GPT-6 Astra	100.0 [100.0, 100.0]	0.0
Qwen2.5-VL-72B	18.8 [16.7, 20.8]	26.0
Qwen2.5-VL-32B	9.4 [4.2, 14.6]	18.8
Qwen2-VL-7B	3.1 [0.0, 6.2]	77.1
Gemma-4-31B	43.8 [25.0, 62.5]	13.5
Qwen3-VL-32B	29.2 [17.7, 36.5]	10.4
GLM-4.6V	6.2 [0.0, 12.5]	35.4
Llama-4-Scout	8.3 [2.1, 14.6]	43.8
MiniCPM-V-4.5	9.4 [6.2, 12.5]	62.5
InternVL3.5-8B	6.2 [2.1, 10.4]	31.2
Kimi-VL-A3B	2.1 [0.0, 4.2]	55.2
ERNIE-4.5-VL-28B-A3B	1.0 [0.0, 3.1]	29.2
Pixtral-12B	1.0 [0.0, 3.1]	56.2
Nemotron-Nano-VL-8B	2.1 [0.0, 4.2]	69.8
GLM-4.1V-9B	9.4 [2.1, 16.7]	30.2
GLM-4.6V-Flash	4.2 [1.0, 7.3]	52.1
Qwen3-VL-8B	1.0 [0.0, 3.1]	76.0
Qwen3-VL-4B	1.0 [0.0, 3.1]	68.8
Qwen3-VL-2B	1.0 [0.0, 3.1]	40.6
Gemma-3-27B	6.2 [2.1, 10.4]	36.5
Gemma-3-12B	5.2 [0.0, 10.4]	36.5
InternVL3-8B	5.2 [4.2, 7.3]	42.7
MiniCPM-V-4	4.2 [4.2, 4.2]	64.6

Table 17: Paired intermediate-state conditions, using 96 pairs with the same source, final endpoint and displayed options but different correct intermediate states. The second rate is equality of parsed outputs, including pairs where both predictions are null; it is not necessarily agreement on a valid graph choice. Joint success remains unconditional.

H AUXILIARY PRESENTATION CONDITIONS

The auxiliary suite contains nuisance renders, paired comparison conditions and a naturalization ladder. The ladder progresses from a skeleton (L0) through sketch (L1), flat-color diagram (L2), minimal vector illustration (L3), an illustration in a simple setting (L4), a photo-like moderately cluttered scene (L5), and a cluttered photorealistic scene (L6). L0 has one item per motif; later levels use three variant indices that select medium, style or scene ideas. The index is not a transmitted generative API random seed. Levels therefore mix rendering and semantic changes rather than calibrating a single entropy variable.

Near resamples a target’s nuisance realization, Far adds another scene with the same intended motif, and Foil adds a similar-surface scene with a different intended motif. Instructed uses exactly the Far composite image with an added comparison instruction; all corresponding image hashes match. The scored target is always the left panel. The common maximum image side applies to the whole composite before JPEG encoding, not to each panel. A paired target can therefore have fewer pixels than its standalone version; this is not an equal-resolution semantic intervention. Accuracy is averaged within motif before comparing ladder levels. Versioned L0 and L5 repairs retain their explicit construction and rendering-style flags.

Model	L0	L1	L2	L3	L4	L5	L6
GPT-5.5	76.0	74.7	81.3	84.0	77.3	82.7	70.7
GPT-5.2	80.0	76.0	76.0	73.3	66.7	57.3	54.7
GPT-6 Astra	96.0	92.0	92.0	90.7	90.7	92.0	80.0
Qwen2.5-VL-72B	56.0	52.0	53.3	52.0	45.3	45.3	34.7
Qwen2.5-VL-32B	64.0	53.3	57.3	46.7	44.0	42.7	28.0
Qwen2-VL-7B	44.0	40.0	40.0	32.0	30.7	33.3	21.3
Gemma-4-31B	84.0	78.7	74.7	81.3	74.7	72.0	65.3
Qwen3-VL-32B	80.0	73.3	76.0	69.3	62.7	62.7	49.3
GLM-4.6V	76.0	72.0	65.3	62.7	46.7	58.7	49.3
Llama-4-Scout	56.0	61.3	65.3	56.0	38.7	46.7	28.0
MiniCPM-V-4.5	52.0	48.0	50.7	41.3	38.7	42.7	32.0
InternVL3.5-8B	36.0	37.3	32.0	28.0	29.3	24.0	21.3
Kimi-VL-A3B	40.0	45.3	46.7	44.0	44.0	42.7	30.7
ERNIE-4.5-VL-28B-A3B	36.0	38.7	37.3	29.3	24.0	26.7	22.7
Pixtral-12B	48.0	36.0	42.7	38.7	22.7	24.0	14.7
Nemotron-Nano-VL-8B	28.0	20.0	24.0	17.3	14.7	20.0	17.3
GLM-4.1V-9B	56.0	52.0	56.0	46.7	40.0	38.7	29.3
GLM-4.6V-Flash	68.0	61.3	60.0	56.0	42.7	41.3	32.0
Qwen3-VL-8B	60.0	58.7	56.0	56.0	42.7	42.7	33.3
Qwen3-VL-4B	56.0	50.7	49.3	50.7	42.7	40.0	25.3
Qwen3-VL-2B	52.0	52.0	52.0	48.0	33.3	33.3	24.0
Gemma-3-27B	48.0	52.0	53.3	45.3	40.0	45.3	36.0
Gemma-3-12B	40.0	38.7	44.0	41.3	26.7	32.0	24.0
InternVL3-8B	24.0	29.3	32.0	30.7	30.7	29.3	22.7
MiniCPM-V-4	24.0	24.0	25.3	24.0	21.3	22.7	20.0

Table 18: Ladder accuracy (%) averaged within motif at each level. Nonmonotone profiles are retained; the table does not impose a shared slope or assert that the manipulation isolates surface entropy.

Model	Single	Near	Far	Foil	Instructed
GPT-5.5	76.0	68.0	68.0	68.0	76.0
GPT-5.2	60.0	60.0	56.0	56.0	72.0
GPT-6 Astra	80.0	76.0	84.0	76.0	88.0
Qwen2.5-VL-72B	48.0	44.0	48.0	44.0	48.0
Qwen2.5-VL-32B	36.0	40.0	36.0	32.0	40.0
Qwen2-VL-7B	36.0	32.0	28.0	20.0	28.0
Gemma-4-31B	64.0	60.0	68.0	68.0	72.0
Qwen3-VL-32B	48.0	56.0	56.0	52.0	56.0
GLM-4.6V	48.0	56.0	52.0	48.0	60.0
Llama-4-Scout	44.0	36.0	36.0	44.0	48.0
MiniCPM-V-4.5	36.0	28.0	28.0	28.0	28.0
InternVL3.5-8B	24.0	28.0	28.0	28.0	28.0
Kimi-VL-A3B	36.0	44.0	44.0	32.0	40.0
ERNIE-4.5-VL-28B-A3B	28.0	16.0	20.0	8.0	28.0
Pixtral-12B	20.0	20.0	24.0	28.0	28.0
Nemotron-Nano-VL-8B	20.0	8.0	12.0	4.0	12.0
GLM-4.1V-9B	40.0	36.0	36.0	36.0	44.0
GLM-4.6V-Flash	28.0	32.0	24.0	28.0	28.0
Qwen3-VL-8B	36.0	44.0	44.0	44.0	48.0
Qwen3-VL-4B	32.0	32.0	32.0	24.0	36.0
Qwen3-VL-2B	24.0	20.0	24.0	24.0	20.0
Gemma-3-27B	44.0	48.0	48.0	40.0	48.0
Gemma-3-12B	32.0	32.0	36.0	24.0	32.0
InternVL3-8B	36.0	28.0	36.0	32.0	36.0
MiniCPM-V-4	20.0	24.0	20.0	20.0	24.0

Table 19: Comparison-condition accuracy (%). Each condition contains 25 motif-paired items. Single is the baseline; Near, Far, Foil and Instructed use the declared composite-image conditions. The target’s effective resolution can differ after the common resize operation.

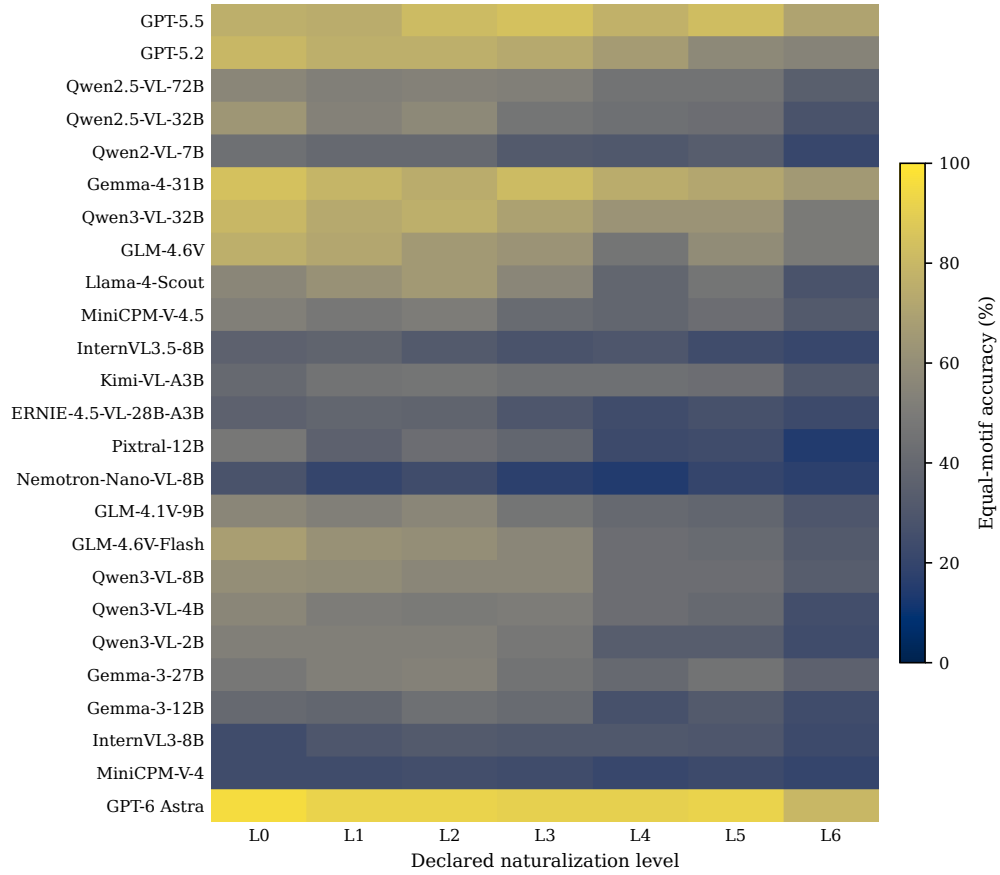


Figure 6: Ladder profiles for all completed vision models. Models and rows follow the fixed table order. Values are equal-motif accuracy at each level, with no smoothing or monotonicity constraint.

I FORMAL AND PERCEPTUAL GEOMETRY

For each model m , geometry uses the same 290 images, the 210 core scenes and 80 auxiliary nuisance renders. Let $c(x)$ denote the intended motif and $\hat{c}_m(x)$ the compatibility-parsed prediction. Define

$$N_i = \sum_x \mathbf{1}\{c(x) = i\}, \quad Y_{ij}^{(m)} = \sum_x \mathbf{1}\{c(x) = i, \hat{c}_m(x) = j\}, \quad j \neq i.$$

N_i includes correct and unparseable responses. Invalid predictions are not reassigned to a motif. If $U_i^{(m)}$ counts invalid predictions and $C_{ii}^{(m)}$ correct predictions, then $\sum_{j \neq i} Y_{ij}^{(m)} = N_i - C_{ii}^{(m)} - U_i^{(m)}$. Thus the dyadic cells in a source row are constrained counts; we do not interpret them as independent observations.

For each model and image encoder e we fit the Poisson working mean

$$\log \mu_{ij}^{(m,e)} = \log N_i + \alpha_{m,e} + \beta_{m,e}^F z_{ij}^F + \beta_{m,e}^G z_{ij}^G + \beta_{m,e}^P z_{ij}^{P,e} + \beta_{m,e}^L z_{ij}^L. \quad (8)$$

There is one global intercept per fit and no source-motif fixed effect. The exposure offset adjusts the number of presentations, not the source-specific probability of an error. We maximize $\sum_{(i,j) \in \mathcal{D}} [Y_{ij}^{(m)} \log \mu_{ij}^{(m,e)} - \mu_{ij}^{(m,e)}]$, where $\mathcal{D} = \{(i,j) : i \neq j, N_i > 0\}$. This is a working likelihood for descriptive rate associations, not a claim of an independent-Poisson data-generating process. Unmodeled source difficulty can affect the fitted associations.

All fitted covariates point toward greater similarity.

$$S^F = 1 - D^F, \quad S^G = -D^G, \quad S^{P,e} = -D^{P,e}, \quad S^L = -D^L,$$

and each is standardized over retained directed off-diagonal cells:

$$z_{ij}^k = \frac{S_{ij}^k - \bar{S}_{\mathcal{D}}^k}{\left[|\mathcal{D}|^{-1} \sum_{(a,b) \in \mathcal{D}} (S_{ab}^k - \bar{S}_{\mathcal{D}}^k)^2 \right]^{1/2}}.$$

The full fit contains all 600 directed off-diagonal cells. Image distance is one minus cosine similarity between unit-normalized motif centroids, formed by averaging unit-normalized image features and normalizing the centroid. The legacy text-description distance D^L is one minus cosine similarity between normalized `text-embedding-3-small` embeddings of `name_en + " . Invariant: " + invariant_zh`. It is not an empirical class-frequency or predicted-label prior. This matrix is retained from the legacy analysis; it was not recomputed from the current English prompts. The historical key `label_prior` is retained only for reproducibility.

Joint-node randomization follows the QAP construction (Hubert & Schultz, 1976). For each draw b , a uniform motif permutation π_b is applied to both axes of every predictor matrix, using the same permutation for all four covariates. Response counts and exposures remain fixed, and the working model is refit. With $B = 10,000$,

$$p_k = \frac{1 + \sum_{b=1}^B \mathbf{1}\{|\hat{\beta}^{k,(b)}| \geq |\hat{\beta}^k|\}}{B + 1}.$$

The family-image contrast uses the same construction with $\hat{\beta}^F - \hat{\beta}^P$ as the statistic. These are two-sided, unadjusted, joint-node-permutation reference p-values for a global predictor-response association null. They are not conditional tests of the unique contribution of a single covariate. If any permutation fit fails, all coefficient and contrast p-values for that model-encoder fit are unavailable. We retain valid observed coefficients and image-bootstrap intervals.

Image-bootstrap intervals (Efron, 1979) resample whole source-image identities 2,000 times, retaining repeated observations of a sampled identity. Each draw rebuilds counts and exposures and refits the model; the predictor matrices remain fixed. This addresses the specified image resampling unit, but does not add source fixed effects, validate the mean specification, or capture uncertainty in the legacy proxies. The leave-one-image-out operation in the separate within-motif distance diagnostic defines an image-to-centroid distance; it is not a jackknife standard-error estimator.

The named external encoders are DINOv2-base, CLIP-B/32 and SigLIP-base, with CLIP-L/14 as the declared alternate. DINOv2 and SigLIP appear in both stored analysis scopes but are the same fits,

not independent replications. Tables 24–27 print all four coefficient-tail statistics and the family–image contrast. N/A means that at least one required permutation fit failed, so a valid complete permutation reference distribution was unavailable; it is not a zero or a nonsignificance claim. These unadjusted p-values do not test a unique conditional causal effect. The image-bootstrap intervals address a different resampling question and need not agree with the QAP decisions.

Model	β^F	β^G	β^P	β^L
GPT-5.5	0.495 [0.289, 0.704]	-0.195 [-0.405, 0.062]	0.320 [0.055, 0.521]	0.352 [0.111, 0.591]
GPT-5.2	0.356 [0.136, 0.562]	0.031 [-0.181, 0.288]	0.357 [0.180, 0.509]	0.105 [-0.138, 0.357]
GPT-6 Astra	0.754 [0.451, 1.069]	-0.319 [-0.628, 0.087]	0.305 [0.029, 0.506]	0.328 [0.003, 0.626]
Qwen2.5-VL-72B	0.392 [0.227, 0.540]	0.122 [-0.011, 0.281]	0.351 [0.233, 0.461]	-0.102 [-0.311, 0.101]
Qwen2.5-VL-32B	0.356 [0.198, 0.512]	0.110 [-0.018, 0.256]	0.217 [0.076, 0.344]	0.131 [-0.059, 0.315]
Qwen2-VL-7B	0.299 [0.157, 0.437]	0.146 [-0.018, 0.300]	0.378 [0.294, 0.464]	-0.152 [-0.334, 0.025]
Gemma-4-31B	0.431 [0.248, 0.619]	-0.184 [-0.364, 0.036]	0.293 [0.085, 0.471]	0.507 [0.300, 0.692]
Qwen3-VL-32B	0.416 [0.218, 0.596]	-0.024 [-0.186, 0.168]	0.279 [0.115, 0.426]	0.297 [0.093, 0.505]
GLM-4.6V	0.374 [0.216, 0.526]	0.002 [-0.147, 0.182]	0.347 [0.196, 0.476]	0.100 [-0.116, 0.300]
Llama-4-Scout	0.383 [0.215, 0.535]	0.164 [0.034, 0.299]	0.200 [0.072, 0.322]	-0.190 [-0.419, 0.038]
MiniCPM-V-4.5	0.429 [0.280, 0.562]	0.103 [-0.037, 0.258]	0.283 [0.158, 0.384]	-0.007 [-0.206, 0.184]
InternVL3.5-8B	0.450 [0.324, 0.574]	0.204 [0.093, 0.332]	0.281 [0.185, 0.373]	-0.297 [-0.476, -0.121]
Kimi-VL-A3B	0.341 [0.200, 0.480]	0.281 [0.129, 0.450]	0.210 [0.068, 0.334]	-0.152 [-0.362, 0.050]
ERNIE-4.5-VL-28B-A3B	0.350 [0.218, 0.478]	-0.031 [-0.170, 0.111]	0.200 [0.083, 0.305]	-0.110 [-0.304, 0.078]
Pixtral-12B	0.462 [0.332, 0.581]	0.234 [0.131, 0.345]	0.178 [0.068, 0.277]	-0.211 [-0.388, -0.041]
Nemotron-Nano-VL-8B	0.335 [0.195, 0.471]	-0.371 [-0.506, -0.245]	-0.017 [-0.141, 0.093]	0.237 [0.087, 0.392]
GLM-4.1V-9B	0.361 [0.222, 0.488]	-0.055 [-0.146, 0.053]	0.110 [-0.033, 0.229]	0.044 [-0.174, 0.245]
GLM-4.6V-Flash	0.300 [0.155, 0.438]	0.172 [0.052, 0.329]	0.106 [-0.013, 0.221]	-0.354 [-0.600, -0.122]
Qwen3-VL-8B	0.434 [0.294, 0.568]	-0.077 [-0.181, 0.044]	0.205 [0.068, 0.323]	0.047 [-0.152, 0.235]
Qwen3-VL-4B	0.256 [0.107, 0.392]	-0.233 [-0.311, -0.154]	0.040 [-0.077, 0.146]	0.479 [0.360, 0.597]
Qwen3-VL-2B	0.345 [0.205, 0.475]	0.199 [0.061, 0.354]	0.235 [0.111, 0.346]	-0.144 [-0.312, 0.033]
Gemma-3-27B	0.401 [0.276, 0.522]	0.031 [-0.073, 0.142]	0.085 [-0.072, 0.215]	-0.001 [-0.223, 0.203]
Gemma-3-12B	0.291 [0.137, 0.419]	0.203 [0.082, 0.342]	0.083 [-0.049, 0.201]	0.048 [-0.144, 0.229]
InternVL3-8B	0.356 [0.213, 0.483]	0.320 [0.219, 0.450]	0.260 [0.170, 0.345]	-0.743 [-0.966, -0.540]
MiniCPM-V-4	0.351 [0.249, 0.449]	0.028 [-0.056, 0.125]	-0.072 [-0.235, 0.063]	0.053 [-0.090, 0.188]

Table 20: DINOv2-base geometry coefficients and 95% image-bootstrap intervals for completed vision models. Family and GED remain frozen formal proxies. The text-description coefficient is the retained legacy similarity covariate, not a frequency prior.

Model	β^F	β^G	β^P	β^L
GPT-5.5	0.454 [0.231, 0.664]	-0.209 [-0.392, 0.051]	0.309 [0.080, 0.531]	0.392 [0.181, 0.601]
GPT-5.2	0.297 [0.075, 0.513]	-0.010 [-0.197, 0.219]	0.414 [0.234, 0.591]	0.163 [-0.083, 0.403]
GPT-6 Astra	0.718 [0.412, 1.057]	-0.350 [-0.642, 0.026]	0.351 [0.076, 0.615]	0.360 [0.051, 0.659]
Qwen2.5-VL-72B	0.325 [0.153, 0.486]	0.050 [-0.081, 0.210]	0.503 [0.355, 0.648]	-0.041 [-0.255, 0.150]
Qwen2.5-VL-32B	0.333 [0.163, 0.485]	0.122 [-0.008, 0.273]	0.101 [-0.044, 0.246]	0.171 [-0.014, 0.341]
Qwen2-VL-7B	0.220 [0.068, 0.354]	0.038 [-0.126, 0.193]	0.670 [0.543, 0.800]	-0.089 [-0.259, 0.072]
Gemma-4-31B	0.391 [0.199, 0.593]	-0.217 [-0.401, -0.001]	0.394 [0.167, 0.621]	0.535 [0.347, 0.709]
Qwen3-VL-32B	0.374 [0.181, 0.555]	-0.046 [-0.208, 0.134]	0.304 [0.117, 0.484]	0.337 [0.159, 0.514]
GLM-4.6V	0.323 [0.169, 0.475]	-0.019 [-0.167, 0.152]	0.325 [0.122, 0.518]	0.158 [-0.047, 0.352]
Llama-4-Scout	0.338 [0.164, 0.503]	0.111 [-0.018, 0.255]	0.331 [0.186, 0.476]	-0.155 [-0.377, 0.049]
MiniCPM-V-4.5	0.389 [0.248, 0.524]	0.092 [-0.045, 0.235]	0.229 [0.070, 0.387]	0.045 [-0.140, 0.231]
InternVL3.5-8B	0.398 [0.270, 0.527]	0.159 [0.051, 0.283]	0.332 [0.201, 0.464]	-0.243 [-0.426, -0.070]
Kimi-VL-A3B	0.289 [0.138, 0.428]	0.225 [0.082, 0.386]	0.352 [0.222, 0.492]	-0.112 [-0.311, 0.086]
ERNIE-4.5-VL-28B-A3B	0.320 [0.181, 0.442]	-0.050 [-0.189, 0.093]	0.203 [0.071, 0.335]	-0.076 [-0.264, 0.101]
Pixtral-12B	0.418 [0.286, 0.542]	0.178 [0.072, 0.292]	0.313 [0.196, 0.430]	-0.178 [-0.346, -0.019]
Nemotron-Nano-VL-8B	0.335 [0.204, 0.467]	-0.401 [-0.540, -0.270]	0.145 [0.041, 0.259]	0.221 [0.083, 0.363]
GLM-4.1V-9B	0.343 [0.215, 0.473]	-0.072 [-0.160, 0.026]	0.144 [-0.010, 0.301]	0.060 [-0.142, 0.250]
GLM-4.6V-Flash	0.262 [0.105, 0.398]	0.113 [-0.004, 0.256]	0.311 [0.175, 0.453]	-0.337 [-0.577, -0.126]
Qwen3-VL-8B	0.404 [0.264, 0.545]	-0.095 [-0.197, 0.019]	0.211 [0.058, 0.372]	0.080 [-0.118, 0.255]
Qwen3-VL-4B	0.278 [0.119, 0.423]	-0.167 [-0.250, -0.082]	-0.425 [-0.569, -0.282]	0.528 [0.410, 0.651]
Qwen3-VL-2B	0.304 [0.161, 0.433]	0.174 [0.041, 0.319]	0.249 [0.114, 0.384]	-0.100 [-0.270, 0.062]
Gemma-3-27B	0.380 [0.253, 0.504]	0.001 [-0.097, 0.112]	0.180 [0.058, 0.311]	0.010 [-0.189, 0.177]
Gemma-3-12B	0.265 [0.122, 0.391]	0.174 [0.052, 0.316]	0.185 [0.060, 0.320]	0.062 [-0.110, 0.232]
InternVL3-8B	0.293 [0.141, 0.422]	0.227 [0.110, 0.363]	0.501 [0.378, 0.621]	-0.686 [-0.892, -0.498]
MiniCPM-V-4	0.352 [0.243, 0.454]	0.012 [-0.074, 0.100]	0.034 [-0.068, 0.136]	0.037 [-0.093, 0.158]

Table 21: Geometry fits using CLIP-B/32, under the same exposure, predictors and uncertainty convention as Table 20.

Model	β^F	β^G	β^P	β^L
GPT-5.5	0.447 [0.223, 0.667]	-0.154 [-0.373, 0.087]	0.176 [-0.085, 0.431]	0.418 [0.192, 0.641]
GPT-5.2	0.279 [0.043, 0.499]	0.044 [-0.157, 0.289]	0.395 [0.205, 0.584]	0.198 [-0.045, 0.435]
GPT-6 Astra	0.686 [0.374, 1.013]	-0.302 [-0.604, 0.079]	0.395 [0.128, 0.679]	0.402 [0.088, 0.703]
Qwen2.5-VL-72B	0.308 [0.131, 0.476]	0.118 [-0.001, 0.276]	0.464 [0.299, 0.634]	-0.007 [-0.219, 0.189]
Qwen2.5-VL-32B	0.339 [0.180, 0.498]	0.140 [0.005, 0.292]	0.036 [-0.130, 0.196]	0.175 [0.004, 0.351]
Qwen2-VL-7B	0.189 [0.038, 0.342]	0.116 [-0.028, 0.270]	0.718 [0.575, 0.862]	-0.042 [-0.221, 0.132]
Gemma-4-31B	0.384 [0.190, 0.589]	-0.145 [-0.330, 0.075]	0.200 [-0.050, 0.434]	0.567 [0.380, 0.744]
Qwen3-VL-32B	0.355 [0.160, 0.540]	-0.009 [-0.167, 0.190]	0.313 [0.135, 0.525]	0.370 [0.179, 0.556]
GLM-4.6V	0.315 [0.158, 0.467]	0.028 [-0.116, 0.197]	0.258 [0.051, 0.463]	0.182 [-0.030, 0.369]
Llama-4-Scout	0.328 [0.164, 0.487]	0.151 [0.030, 0.288]	0.326 [0.159, 0.492]	-0.135 [-0.362, 0.076]
MiniCPM-V-4.5	0.379 [0.232, 0.510]	0.119 [-0.010, 0.269]	0.233 [0.072, 0.395]	0.063 [-0.125, 0.238]
InternVL3.5-8B	0.390 [0.254, 0.526]	0.200 [0.088, 0.319]	0.323 [0.175, 0.460]	-0.226 [-0.411, -0.050]
Kimi-VL-A3B	0.278 [0.127, 0.429]	0.266 [0.131, 0.429]	0.349 [0.216, 0.491]	-0.090 [-0.293, 0.103]
ERNIE-4.5-VL-28B-A3B	0.313 [0.175, 0.437]	-0.024 [-0.155, 0.109]	0.189 [0.048, 0.336]	-0.063 [-0.260, 0.116]
Pixtral-12B	0.415 [0.283, 0.541]	0.224 [0.131, 0.327]	0.258 [0.132, 0.391]	-0.162 [-0.338, -0.004]
Nemotron-Nano-VL-8B	0.323 [0.190, 0.454]	-0.383 [-0.520, -0.258]	0.174 [0.072, 0.275]	0.236 [0.086, 0.375]
GLM-4.1V-9B	0.351 [0.209, 0.488]	-0.043 [-0.140, 0.060]	0.018 [-0.143, 0.189]	0.067 [-0.132, 0.261]
GLM-4.6V-Flash	0.265 [0.104, 0.405]	0.159 [0.036, 0.306]	0.216 [0.061, 0.379]	-0.325 [-0.576, -0.104]
Qwen3-VL-8B	0.401 [0.258, 0.545]	-0.062 [-0.161, 0.054]	0.143 [-0.028, 0.315]	0.094 [-0.094, 0.277]
Qwen3-VL-4B	0.292 [0.137, 0.439]	-0.210 [-0.290, -0.119]	-0.356 [-0.513, -0.202]	0.478 [0.369, 0.588]
Qwen3-VL-2B	0.290 [0.144, 0.413]	0.195 [0.063, 0.341]	0.300 [0.170, 0.432]	-0.081 [-0.253, 0.078]
Gemma-3-27B	0.388 [0.246, 0.517]	0.036 [-0.064, 0.147]	0.061 [-0.074, 0.200]	0.018 [-0.171, 0.198]
Gemma-3-12B	0.267 [0.129, 0.412]	0.200 [0.081, 0.340]	0.131 [-0.008, 0.272]	0.072 [-0.098, 0.248]
InternVL3-8B	0.283 [0.137, 0.419]	0.277 [0.175, 0.399]	0.507 [0.375, 0.646]	-0.669 [-0.889, -0.471]
MiniCPM-V-4	0.376 [0.270, 0.477]	0.033 [-0.059, 0.129]	-0.143 [-0.258, -0.032]	0.032 [-0.086, 0.152]

Table 22: Geometry fits using SigLIP-base. The measured encoder space is an external covariate, not the evaluated model’s own activation space.

Model	β^F	β^G	β^P	β^L
GPT-5.5	0.446 [0.223, 0.662]	-0.179 [-0.368, 0.056]	0.188 [-0.068, 0.420]	0.406 [0.189, 0.610]
GPT-5.2	0.268 [0.035, 0.484]	-0.001 [-0.192, 0.236]	0.392 [0.205, 0.578]	0.185 [-0.056, 0.415]
GPT-6 Astra	0.697 [0.382, 1.045]	-0.339 [-0.616, 0.032]	0.322 [0.048, 0.589]	0.372 [0.076, 0.656]
Qwen2.5-VL-72B	0.291 [0.115, 0.462]	0.066 [-0.053, 0.229]	0.465 [0.311, 0.612]	-0.009 [-0.218, 0.185]
Qwen2.5-VL-32B	0.341 [0.174, 0.501]	0.140 [0.005, 0.298]	0.020 [-0.144, 0.176]	0.173 [-0.007, 0.341]
Qwen2-VL-7B	0.170 [0.007, 0.307]	0.066 [-0.081, 0.219]	0.606 [0.488, 0.722]	-0.043 [-0.206, 0.122]
Gemma-4-31B	0.387 [0.195, 0.587]	-0.167 [-0.344, 0.048]	0.184 [-0.068, 0.422]	0.551 [0.368, 0.722]
Qwen3-VL-32B	0.349 [0.138, 0.543]	-0.048 [-0.212, 0.138]	0.320 [0.137, 0.497]	0.352 [0.164, 0.523]
GLM-4.6V	0.299 [0.123, 0.453]	-0.016 [-0.157, 0.153]	0.323 [0.129, 0.517]	0.176 [-0.016, 0.368]
Llama-4-Scout	0.304 [0.118, 0.460]	0.100 [-0.019, 0.239]	0.390 [0.238, 0.536]	-0.127 [-0.347, 0.087]
MiniCPM-V-4.5	0.359 [0.209, 0.504]	0.071 [-0.065, 0.214]	0.318 [0.149, 0.468]	0.062 [-0.125, 0.230]
InternVL3.5-8B	0.368 [0.233, 0.503]	0.153 [0.050, 0.266]	0.366 [0.237, 0.492]	-0.215 [-0.401, -0.048]
Kimi-VL-A3B	0.254 [0.104, 0.401]	0.216 [0.073, 0.373]	0.399 [0.264, 0.531]	-0.084 [-0.279, 0.098]
ERNIE-4.5-VL-28B-A3B	0.309 [0.170, 0.441]	-0.043 [-0.177, 0.092]	0.178 [0.038, 0.320]	-0.062 [-0.243, 0.117]
Pixtral-12B	0.394 [0.260, 0.524]	0.178 [0.088, 0.289]	0.320 [0.188, 0.444]	-0.155 [-0.312, 0.001]
Nemotron-Nano-VL-8B	0.336 [0.200, 0.468]	-0.375 [-0.514, -0.248]	0.013 [-0.103, 0.127]	0.233 [0.095, 0.382]
GLM-4.1V-9B	0.340 [0.202, 0.479]	-0.061 [-0.142, 0.038]	0.097 [-0.076, 0.253]	0.068 [-0.132, 0.259]
GLM-4.6V-Flash	0.229 [0.074, 0.365]	0.107 [-0.002, 0.243]	0.358 [0.215, 0.497]	-0.304 [-0.530, -0.092]
Qwen3-VL-8B	0.394 [0.238, 0.533]	-0.087 [-0.188, 0.025]	0.179 [0.011, 0.350]	0.092 [-0.098, 0.266]
Qwen3-VL-4B	0.304 [0.145, 0.454]	-0.137 [-0.215, -0.053]	-0.521 [-0.701, -0.352]	0.509 [0.386, 0.634]
Qwen3-VL-2B	0.277 [0.138, 0.404]	0.162 [0.032, 0.309]	0.301 [0.176, 0.424]	-0.079 [-0.235, 0.087]
Gemma-3-27B	0.389 [0.255, 0.513]	0.033 [-0.067, 0.148]	0.044 [-0.099, 0.194]	0.017 [-0.172, 0.185]
Gemma-3-12B	0.259 [0.115, 0.395]	0.182 [0.069, 0.310]	0.147 [0.004, 0.282]	0.071 [-0.102, 0.238]
InternVL3-8B	0.235 [0.082, 0.382]	0.208 [0.102, 0.344]	0.592 [0.478, 0.711]	-0.608 [-0.801, -0.436]
MiniCPM-V-4	0.400 [0.291, 0.501]	0.077 [-0.004, 0.171]	-0.270 [-0.417, -0.134]	0.030 [-0.099, 0.163]

Table 23: Named CLIP-L/14 alternate. Results are reported separately rather than selecting the encoder with the most favorable association.

Model	p_F	p_G	p_P	p_L	p_{F-P}	Failed / B
GPT-5.5	N/A	N/A	N/A	N/A	N/A	2 / 10,000
GPT-5.2	0.1184	0.8806	0.0807	0.6415	0.9957	0 / 10,000
GPT-6 Astra	N/A	N/A	N/A	N/A	N/A	26 / 10,000
Qwen2.5-VL-72B	0.0384	0.4999	0.0608	0.6100	0.8709	0 / 10,000
Qwen2.5-VL-32B	0.0417	0.5382	0.2126	0.4797	0.5700	0 / 10,000
Qwen2-VL-7B	0.0879	0.3934	0.0412	0.4174	0.7591	0 / 10,000
Gemma-4-31B	0.1383	0.4727	0.2743	0.0849	0.7260	0 / 10,000
Qwen3-VL-32B	0.0426	0.8982	0.1206	0.1389	0.6016	0 / 10,000
GLM-4.6V	0.0770	0.9916	0.1025	0.6442	0.9274	0 / 10,000
Llama-4-Scout	0.0585	0.3965	0.3469	0.3817	0.5457	0 / 10,000
MiniCPM-V-4.5	0.0154	0.5661	0.0917	0.9693	0.5342	0 / 10,000
InternVL3.5-8B	0.0236	0.2781	0.1634	0.1583	0.5538	0 / 10,000
Kimi-VL-A3B	0.0704	0.1521	0.2528	0.4380	0.6128	0 / 10,000
ERNIE-4.5-VL-28B-A3B	0.0174	0.8428	0.2218	0.4985	0.4925	0 / 10,000
Pixtral-12B	0.0242	0.3009	0.4317	0.3445	0.3610	0 / 10,000
Nemotron-Nano-VL-8B	0.0265	0.0976	0.9294	0.1783	0.1447	0 / 10,000
GLM-4.1V-9B	0.0498	0.7895	0.5922	0.8358	0.3709	0 / 10,000
GLM-4.6V-Flash	0.1523	0.4677	0.6682	0.1318	0.5561	0 / 10,000
Qwen3-VL-8B	0.0251	0.6618	0.2727	0.8166	0.3938	0 / 10,000
Qwen3-VL-4B	0.2719	0.3692	0.8948	0.0777	0.5845	0 / 10,000
Qwen3-VL-2B	0.0547	0.2983	0.2436	0.4501	0.6953	0 / 10,000
Gemma-3-27B	0.0685	0.8885	0.7361	0.9946	0.3569	0 / 10,000
Gemma-3-12B	0.0821	0.2492	0.6320	0.7946	0.3822	0 / 10,000
InternVL3-8B	0.1375	0.3082	0.4680	0.0441	0.8613	0 / 10,000
MiniCPM-V-4	0.1346	0.9085	0.8533	0.8445	0.3812	0 / 10,000

Table 24: DINOv2-base joint-node-permutation reference p-values. p_F, p_G, p_P, p_L use the family, graph-edit, image and legacy text-description coefficients; p_{F-P} uses the family-image coefficient contrast. Values are two-sided and unadjusted under the global association null. Failed/ B records nonconvergent permutation fits; any failure makes the fit’s p-values N/A.

Model	p_F	p_G	p_P	p_L	p_{F-P}	Failed / B
GPT-5.5	N/A	N/A	N/A	N/A	N/A	2 / 10,000
GPT-5.2	0.1920	0.9623	0.0388	0.4605	0.7100	0 / 10,000
GPT-6 Astra	N/A	N/A	N/A	N/A	N/A	26 / 10,000
Qwen2.5-VL-72B	0.0803	0.7810	0.0086	0.8356	0.5181	0 / 10,000
Qwen2.5-VL-32B	0.0585	0.5114	0.5896	0.3507	0.3724	0 / 10,000
Qwen2-VL-7B	0.2098	0.8106	0.00020	0.6372	0.0883	0 / 10,000
Gemma-4-31B	N/A	N/A	N/A	N/A	N/A	2 / 10,000
Qwen3-VL-32B	0.0585	0.7979	0.0959	0.0866	0.7988	0 / 10,000
GLM-4.6V	0.1258	0.9249	0.1542	0.4752	0.9945	0 / 10,000
Llama-4-Scout	0.0938	0.5615	0.1523	0.4750	0.9819	0 / 10,000
MiniCPM-V-4.5	0.0303	0.6027	0.1921	0.7989	0.5269	0 / 10,000
InternVL3.5-8B	0.0485	0.3863	0.1319	0.2430	0.8279	0 / 10,000
Kimi-VL-A3B	0.1236	0.2334	0.0666	0.5651	0.8218	0 / 10,000
ERNIE-4.5-VL-28B-A3B	0.0274	0.7533	0.2695	0.6350	0.6213	0 / 10,000
Pixtral-12B	0.0372	0.4194	0.2179	0.4253	0.7435	0 / 10,000
Nemotron-Nano-VL-8B	0.0269	0.0683	0.4936	0.2194	0.4565	0 / 10,000
GLM-4.1V-9B	0.0636	0.7320	0.5174	0.7731	0.4980	0 / 10,000
GLM-4.6V-Flash	0.2076	0.6201	0.2575	0.1631	0.8851	0 / 10,000
Qwen3-VL-8B	0.0338	0.5978	0.2758	0.6849	0.4907	0 / 10,000
Qwen3-VL-4B	0.2411	0.4964	0.1823	0.0641	0.0760	0 / 10,000
Qwen3-VL-2B	0.0854	0.3645	0.2599	0.6009	0.8453	0 / 10,000
Gemma-3-27B	0.0795	0.9945	0.5217	0.9629	0.5735	0 / 10,000
Gemma-3-12B	0.1115	0.3202	0.3093	0.7317	0.7531	0 / 10,000
InternVL3-8B	0.2274	0.4080	0.2168	0.0630	0.6659	0 / 10,000
MiniCPM-V-4	0.1293	0.9621	0.9292	0.8821	0.4961	0 / 10,000

Table 25: CLIP-B/32 joint-node-permutation reference p-values. p_F, p_G, p_P, p_L use the family, graph-edit, image and legacy text-description coefficients; p_{F-P} uses the family-image coefficient contrast. Values are two-sided and unadjusted under the global association null. Failed/ B records nonconvergent permutation fits; any failure makes the fit’s p-values N/A.

Model	p_F	p_G	p_P	p_L	p_{F-P}	Failed / B
GPT-5.5	N/A	N/A	N/A	N/A	N/A	2 / 10,000
GPT-5.2	0.2220	0.8330	0.0445	0.3761	0.7173	0 / 10,000
GPT-6 Astra	N/A	N/A	N/A	N/A	N/A	31 / 10,000
Qwen2.5-VL-72B	0.1029	0.5059	0.0154	0.9673	0.5749	0 / 10,000
Qwen2.5-VL-32B	0.0496	0.4421	0.8507	0.3336	0.2539	0 / 10,000
Qwen2-VL-7B	0.2719	0.4882	0.00020	0.8210	0.0505	0 / 10,000
Gemma-4-31B	0.1974	0.5674	0.4722	0.0511	0.6600	0 / 10,000
Qwen3-VL-32B	0.0754	0.9618	0.0807	0.0612	0.8764	0 / 10,000
GLM-4.6V	0.1351	0.8907	0.2550	0.4028	0.8588	0 / 10,000
Llama-4-Scout	0.0954	0.4166	0.1656	0.5158	0.9950	0 / 10,000
MiniCPM-V-4.5	0.0342	0.5145	0.1734	0.7237	0.5668	0 / 10,000
InternVL3.5-8B	0.0499	0.2814	0.1503	0.2694	0.8250	0 / 10,000
Kimi-VL-A3B	0.1310	0.1662	0.0717	0.6449	0.7957	0 / 10,000
ERNIE-4.5-VL-28B-A3B	0.0300	0.8747	0.3058	0.6876	0.5863	0 / 10,000
Pixtral-12B	0.0349	0.3189	0.3240	0.4542	0.6168	0 / 10,000
Nemotron-Nano-VL-8B	0.0288	0.0815	0.4070	0.1769	0.5496	0 / 10,000
GLM-4.1V-9B	0.0545	0.8450	0.9396	0.7419	0.2526	0 / 10,000
GLM-4.6V-Flash	0.1911	0.5092	0.4430	0.1627	0.8838	0 / 10,000
Qwen3-VL-8B	0.0380	0.7274	0.4730	0.6319	0.3633	0 / 10,000
Qwen3-VL-4B	0.2098	0.3919	0.2824	0.0814	0.0915	0 / 10,000
Qwen3-VL-2B	0.0994	0.3135	0.1773	0.6712	0.9703	0 / 10,000
Gemma-3-27B	0.0703	0.8616	0.8347	0.9391	0.3531	0 / 10,000
Gemma-3-12B	0.1080	0.2474	0.4709	0.6736	0.5993	0 / 10,000
InternVL3-8B	0.2181	0.3464	0.2457	0.0644	0.6601	0 / 10,000
MiniCPM-V-4	0.0844	0.8844	0.7135	0.8897	0.2480	0 / 10,000

Table 26: SigLIP-base joint-node-permutation reference p-values. p_F, p_G, p_P, p_L use the family, graph-edit, image and legacy text-description coefficients; p_{F-P} uses the family-image coefficient contrast. Values are two-sided and unadjusted under the global association null. Failed/ B records nonconvergent permutation fits; any failure makes the fit’s p-values N/A.

Model	p_F	p_G	p_P	p_L	p_{F-P}	Failed / B
GPT-5.5	N/A	N/A	N/A	N/A	N/A	1 / 10,000
GPT-5.2	0.2419	0.9973	0.0563	0.4108	0.7101	0 / 10,000
GPT-6 Astra	N/A	N/A	N/A	N/A	N/A	30 / 10,000
Qwen2.5-VL-72B	0.1205	0.7111	0.0163	0.9626	0.5452	0 / 10,000
Qwen2.5-VL-32B	0.0529	0.4416	0.9173	0.3461	0.2240	0 / 10,000
Qwen2-VL-7B	0.3299	0.6825	0.00080	0.8108	0.1078	0 / 10,000
Gemma-4-31B	N/A	N/A	N/A	N/A	N/A	1 / 10,000
Qwen3-VL-32B	0.0826	0.7886	0.0712	0.0754	0.9175	0 / 10,000
GLM-4.6V	0.1493	0.9359	0.1523	0.4162	0.9390	0 / 10,000
Llama-4-Scout	0.1273	0.5999	0.0914	0.5426	0.7900	0 / 10,000
MiniCPM-V-4.5	0.0459	0.6955	0.0670	0.7228	0.8785	0 / 10,000
InternVL3.5-8B	0.0644	0.4022	0.0999	0.3083	0.9958	0 / 10,000
Kimi-VL-A3B	0.1732	0.2483	0.0381	0.6584	0.6062	0 / 10,000
ERNIE-4.5-VL-28B-A3B	0.0345	0.7853	0.3349	0.6981	0.5857	0 / 10,000
Pixtral-12B	0.0459	0.4217	0.2106	0.4725	0.8175	0 / 10,000
Nemotron-Nano-VL-8B	0.0208	0.0930	0.9454	0.1829	0.1973	0 / 10,000
GLM-4.1V-9B	0.0693	0.7671	0.6666	0.7350	0.4097	0 / 10,000
GLM-4.6V-Flash	0.2729	0.6474	0.1868	0.1953	0.7169	0 / 10,000
Qwen3-VL-8B	0.0409	0.6200	0.3645	0.6409	0.4541	0 / 10,000
Qwen3-VL-4B	0.1892	0.5658	0.1051	0.0642	0.0339	0 / 10,000
Qwen3-VL-2B	0.1193	0.3952	0.1755	0.6734	0.9309	0 / 10,000
Gemma-3-27B	0.0736	0.8775	0.8821	0.9386	0.3297	0 / 10,000
Gemma-3-12B	0.1239	0.2895	0.4213	0.6883	0.6647	0 / 10,000
InternVL3-8B	0.3454	0.4414	0.1722	0.0712	0.4655	0 / 10,000
MiniCPM-V-4	0.0742	0.7473	0.5295	0.8995	0.1300	0 / 10,000

Table 27: CLIP-L/14 joint-node-permutation reference p-values. p_F, p_G, p_P, p_L use the family, graph-edit, image and legacy text-description coefficients; p_{F-P} uses the family-image coefficient contrast. Values are two-sided and unadjusted under the global association null. Failed/ B records nonconvergent permutation fits; any failure makes the fit’s p-values N/A.

Model	Encoder	Failed / B
Gemma-4-31B	CLIP-B/32	2 / 10,000
GPT-5.5	DINOv2-base	2 / 10,000
GPT-5.5	CLIP-B/32	2 / 10,000
GPT-5.5	SigLIP-base	2 / 10,000
GPT-6 Astra	DINOv2-base	26 / 10,000
GPT-6 Astra	CLIP-B/32	26 / 10,000
GPT-6 Astra	SigLIP-base	31 / 10,000
Gemma-4-31B	CLIP-L/14	1 / 10,000
GPT-5.5	CLIP-L/14	1 / 10,000
GPT-6 Astra	CLIP-L/14	30 / 10,000

Table 28: Model–encoder fits with unavailable QAP distributions after nonconvergence in at least one permutation. Observed coefficients and valid image-bootstrap intervals are retained separately.

J THE FULL MOTIF CATALOG

Table 29 lists all 25 motifs with their signature equation, mathematical home, and the classical theorem attached to each. The theorems are stated (and, where relevant, corrected) as part of the propose–curate–verify pipeline; the mathematics voice presented to models is derived from these entries.

Table 29: The full motif catalog: signature equation, classical mathematical home, and a gloss of the story form device.

Motif	Signature	Home	Story device
Self-Reference	$\vdash \psi \leftrightarrow \varphi(\ulcorner \psi \urcorner)$	Mathematical logic: diagonal lemma; computability: Kleene’s second recursion theorem	Acrostic fixed point: the title asserts the acrostic; sentence initials spell the title.
Recursion	$T ::= \bullet \mid \text{Node}(T, T), \quad C_n = \frac{1}{n+1} \binom{2n}{n}$	Combinatorics: Catalan numbers and Dyck words; theory of computation: structural induction and the call stack	Nested sub-tales under fixed enter/exit formulas (“now the tale tells” / “and back we come”), two children or none, shrinking to a fixed base sentence.
Nesting	$[a_1, b_1] \supsetneq [a_2, b_2] \supsetneq \cdots \implies \bigcap_{n=1}^{\infty} [a_n, b_n] \neq \emptyset$	Real analysis: Cantor’s nested-interval theorem (compactness); formal languages: proper prefixes of the Dyck language (push-only words)	Open-only frames: openers embed ever deeper, no outer name recurs—push without pop.
Cycle	$v_1 \rightarrow v_2 \rightarrow \cdots \rightarrow v_n \rightarrow v_1, \quad \langle r \mid r^n = e \rangle \cong \mathbb{Z}/n\mathbb{Z}$	Graph theory: directed cycle C_n ; group theory: cyclic group $\mathbb{Z}/n\mathbb{Z}$; discrete dynamics: periodic orbits of functional graphs	Causal ring: “thereupon”-linked segments whose last result line equals the first verbatim.
Mirror Palindrome	$w = w_{n+1-i}, \quad w^R, \quad (w^R)_i = R \circ R = \text{id}$	Combinatorics on words: palindromic factors and rich words; group theory: dihedral group D_n and involutions	Mirrored numbered word-token lists, with optional Alice/Bob exchange for each pair; punctuation is ignored.
Alternation	$c(i) = i \bmod 2; \quad \chi(G) \leq 2 \iff G \text{ has no odd cycle}$	Graph theory: bipartite graphs and 2-coloring (König’s characterization); group theory: $\mathbb{Z}/2\mathbb{Z}$	Two-voice script: turns labelled by two fixed speaker tags in strict parity alternation.
Spiral	$r = a e^{b\theta}$	Differential geometry / classical curves: logarithmic (equiangular) spiral; discrete dynamics: geometric growth	Enlarging refrain: one refrain opens four English sections of 6/12/24/48 words.
Convergence	$\forall v \in V \quad \exists! P : v \rightsquigarrow r$	Graph theory: in-trees (anti-arborescences); rewriting: confluence (Church–Rosser); analysis: Banach fixed-point theorem	Three disjoint threads; all names first co-occur at a new place in the final line.
Bottleneck	$uv \notin E : \min_{\substack{S \subseteq V \setminus \{u,v\} \\ S \text{ separates } u,v}} S = \max \# \{ \text{internally disjoint } u\text{--}v \text{ paths} \}$	Graph theory: cut vertices and Menger’s theorem	Three threads each cross the single-log bridge exactly once, midway, then fan out and continue.
Symmetry Breaking	$\sum_p : X(p)=0 \quad \text{ind}_p(X) = \chi(M)$	Differential topology: Poincaré–Hopf and the hairy-ball theorem; equivariant bifurcation theory: pitchfork bifurcation (spontaneous symmetry breaking)	Twelve sentences obey one rule; exactly sentence 7 breaks it.
Strange Loop	$\ell(v) = \ell(u) + 1 \quad \forall (u, v) \in E \setminus \{e^*\}, \quad e^* = (v_k, v_0), \quad \ell(v_k) = k, \quad \ell(v_0) = 0$	Graph theory: topological sorting and minimum feedback arc sets (DAG criterion); the graph skeleton of Gödel’s diagonal construction	Three narration levels with a back-edge: the innermost hero is the outer narrator; last line = first.
Figure and Ground	$\exists A \ (A \in \Sigma_1^0 \wedge \overline{A} \notin \Sigma_1^0)$	Recursion theory: recursively enumerable sets and their complements (Post 1944)	The same events retold in the same order with the subject role complemented.
Isomorphism	$\varphi(a \cdot b) = \varphi(a) \circ \varphi(b), \quad \varphi : G \rightarrow H \text{ bijective}$	Group theory / universal algebra: isomorphism; model theory: isomorphism vs. elementary equivalence	Two disjoint vocabularies with a bijective dictionary; sentence k maps to sentence k .
Möbius	$M = [0, 1]^2 / \sim, \quad (0, t) \sim (1, 1 - t)$	Topology: non-orientable surfaces	Two laps; the second repeats the first with orientation words involuted; only the coda closes it.
Braiding	$\sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}$	Low-dimensional topology: Artin braid groups	Seat order announced each scene; adjacent swaps compose to the final order.

Motif	Signature	Home	Story device
Interference	$\sin \alpha + \sin \beta = 2 \sin \frac{\alpha + \beta}{2} \cos \frac{\alpha - \beta}{2}$	Fourier analysis: beats (sum-to-product); number theory: least common multiple and the Chinese Remainder Theorem	A period-3 opener rule and a period-4 ending rule coincide only at sentences 12 and 24.
Tiling	$T + \lambda = T \quad \forall \lambda \in \Lambda = \mathbb{Z}v_1 \oplus \mathbb{Z}v_2$	Discrete geometry: wallpaper groups (plane crystallographic groups)	Eight instances of one sentence template, three slots swapped, gapless and overlap-free.
Hub	$G \cong K_{1,n}$	Graph theory: star graph $K_{1,n}$ (complete bipartite) and degree centrality	Five villagers never speak directly; every quoted pair includes the miller.
The Diagonal Escape	$d(n) = 1 - f(n)(n) \Rightarrow \forall n : d \neq f(n)$	Set theory / computability: Cantor's diagonal argument	n tellers $\times n$ binary traits; the traveler's trait k flips teller k 's trait k , so it matches no teller.
Infinite Descent	$(\forall n \in \mathbb{N} : P(n) \Rightarrow \exists m < n, P(m)) \Rightarrow \forall n \neg P(n)$	Number theory: Fermat's method of infinite descent / well-ordering of the naturals	A universal positive-integer claimant law, a smaller-witness premise, and a least-witness contradiction.
Coprime Orbit	$\langle k \rangle = \mathbb{Z}_n \iff \gcd(k, n) = 1$	Elementary number theory / group theory: generators of the cyclic group $\mathbb{Z}_n$; star polygons $\{n/k\}$ (Poincaré 1810)	A lamp at stride 3 around 8: roll-call 0, 3, 6, 1, 4, 7, 2, 5, everyone once; a stride-2 control strands half the circle.
Pigeonhole	$ A > B \Rightarrow \forall f : A \rightarrow B, \exists x \neq y : f(x) = f(y)$	Combinatorics: Dirichlet's pigeonhole principle (Schubfachprinzip, 1834; used in Diophantine approximation)	13 guests, 12 month-slots; the elder asserts a collision before any roll is read; the count confirms it.
The Involutive Role-Swap	$\sigma : (a:b:c) \leftrightarrow \{ax + by + cz = 0\}, \sigma^2 = \text{id}, \vdash P \Leftrightarrow \vdash P^\sigma$	Projective geometry: the principle of duality (Gergonne–Poncelet)	Told twice under an opening dictionary of role swaps; applying it to the second telling restores the first verbatim.
The Shrinking Return	$d(f(x), f(y)) \leq q d(x, y) \ (q < 1)$ on complete $(X, d) \Rightarrow \exists! x^* : f(x^*) = x^*, \ x_{n+1} = f(x_n) \rightarrow x^*$	Analysis / dynamical systems: Banach fixed-point theorem and Picard iteration	A chant deletes its back half each round (17→9→5→3→2→1) until one word maps to itself.
Aperiodic Tiling	$\frac{N_{\text{fat}}(n)}{N_{\text{thin}}(n)} \xrightarrow{n \rightarrow \infty} \varphi = \frac{1+\sqrt{5}}{2}$	Discrete geometry / tiling theory and quasicrystal mathematics: Penrose tilings (P1 1974; P3 rhombs 1978); pentagrid algebra (de Bruijn 1981)	Twenty-one sentences typed A/B by the substitution $A \rightarrow AB, B \rightarrow A$; never a repeating block.

K A SCORED ENGLISH FORM-DEVICE EXAMPLE

The following title and body are the complete scored payload for `story:symmetry_ens2`; the story is not an English translation of the later Chinese display example.

Red Mitten

1. Alice found a red mitten on the laundry hook and took it to Bob by the rain barrel.
2. Marta poured sweet tea into three cups while the cat slept and the children hummed a farm tune.
3. Alice found a red mitten on the laundry hook and took it to Bob by the rain barrel.

Item `story:symmetry_ens2` uses this exact title and body. The read-only validator passes odd nonempty line count, numbering, absence of the declared layer markers, and reversal of the word-token lists (with optional Alice/Bob exchange). This certifies the named finite text predicates, not unrestricted narrative meaning.

For this payload, all four recorded checks pass. These are an odd nonempty-line count, consecutive numbering, absence of the declared layer markers, and mirrored word-token lists. The first and third lines are identical here; the validator also permits the specified Alice/Bob exchange in other pairs. The checks concern the stated finite predicates, not unrestricted meaning or a proof that only one motif can describe the story.



Figure 7: Self-reference: the corrected directed self-loop skeleton (left) and a historical illustrative scene (right). The explicit self-edge replaces the earlier empty nested-frame skeleton. The scene is intended to depict a painting of the hall containing it; the photograph alone is not a formal certificate that the depicted room is the same room.

L WORKED EXAMPLES. THREE MOTIFS, FOUR VOICES

We walk through three motifs end to end, *self-reference*, *recursion* and *interference*, showing for each the skeleton, a scene illustrating the same intended motif, the story with its form device, and the mathematics voice. The stories walked through here are from the released Chinese display corpus, presented in English translation; their form devices operate at the character level and are both defined and mechanically checked on the Chinese originals (the self-reference acrostic is walked through in Appendix M). The benchmark’s scored story items are English, with devices defined on English mechanics (Section 3); the Chinese corpus ships as display material only and is never scored.

L.1 SELF-REFERENCE

Scene and skeleton. Figure 7. The nuisance variables (frame counts, aspect ratios, palette) vary freely across scenes; the scored content is only the one-step self-inclusion.

Story (English translation of the Chinese original).

In a riverside town stood a tiny woodblock print shop run by Gu Huai, a craftsman whose carving knife was steadier than a needle. At night he carved primers for children and decorated the rhymes with fish, wheat, and plump ducks. He had taken a kind saying left by his late wife, divided it character by character, and tucked those tiny marks into the first cuts of each block. Even his daughter A Sui did not know; she only felt the pictures were dear to her. One rainy season the river rose and A Sui lost her way while delivering books. A chestnut-selling grandmother noticed the little marks in the new primers seemed to lead east, and remembered how A Sui loved following them. The villagers followed the signs to the old ferry and found the girl safe under a boat awning, still guarding the dry books. Afterward, Gu Huai kept carving the hidden blessing into the first strokes of his blocks.

The device is invisible in translation. In the Chinese original, the title is the assertion “the sentence-initial characters of this piece, read in order, are exactly this title,” and the thirteen sentence-initial characters spell exactly that title, a single-level self-referential fixed point, verified mechanically in Appendix M. Note that the *plot* (marks hidden in first strokes lead searchers to the child) merely echoes the device; the scored structure lives in the telling.



Figure 8: *recursion*, one structure under many surfaces: the programmatic skeleton (left) and three illustrative scenes associated with the motif: a river delta, a lightning bolt, and a winter tree. The images vary in appearance and domain while targeting the same structure; branch counts, depths, and angles are unscored nuisance variables.

Mathematics voice (translated).

Equation. $\vdash \psi \leftrightarrow \varphi(\ulcorner \psi \urcorner)$

Definition. Diagonal lemma: if a first-order theory T represents all primitive recursive functions (e.g. contains Robinson arithmetic Q), then for every formula $\varphi(x)$ with exactly one free variable one can *explicitly construct* a sentence ψ with $T \vdash \psi \leftrightarrow \varphi(\ulcorner \psi \urcorner)$: ψ is provably equivalent to “the sentence coded by $\ulcorner \psi \urcorner$ (i.e. itself) has property φ .” Self-reference is thus a constructible fixed point of a description operator, not a paradox.

Home. Mathematical logic (diagonal lemma); computability theory (Kleene’s second recursion theorem).

Theorem. Kleene (1938): for every total computable f there is an index e with $\varphi_e = \varphi_{f(e)}$; its corollary states that every Turing-complete language with an acceptable numbering contains a program that outputs its own source code (a quine).

L.2 RECURSION

Scene and skeleton. Figure 8.

Story (English translation).

At the end of Pinecone Lane stood a little umbrella shop run by A-Liu. Before the rainy-season fair, she was asked to make a great umbrella. One part of the work led young Xiaoman to an old bamboo craftsman, who sent him for green bamboo and warm hemp thread. The other part led A-Liu to the dyer Qiuniang, who asked two children to bring blue cloth and madder flowers. With bamboo, thread, blue cloth, and red patterns all fitted together, the umbrella opened at the fair, and the children listened happily to the rain tapping above them.

Form device. In the Chinese original every sub-tale opens with the fixed formula *huà shuō* (“now the tale tells”) and closes with *qiě shuō huí lái* (“and back we come”), forming matched brackets; every non-leaf tale contains exactly two sub-tales, every leaf contains verbatim the fixed base-case sentence (“this was the smallest of the errands”), and the maximum depth is three. The bracket string is a well-formed Dyck word over a binary tree, so entering a level and exiting it returns to the caller.

Mathematics voice (translated).

Equation. $T ::= \bullet \mid \text{Node}(T, T), \quad C_n = \frac{1}{n+1} \binom{2n}{n}$

Definition. A structure generated inductively from a base case by a single grammar $T ::= \text{base} \mid \text{Cons}(T, T)$; well-foundedness guarantees every unfolding reaches a base case in finitely many steps and returns level by level. Full binary trees admit a standard bijection with Dyck words. For example, the recursive encoding $E(\bullet) = \epsilon$ and $E(\text{Node}(L, R)) = (E(L))E(R)$ yields a Dyck word with one matched pair per internal node; its inverse follows the first-return decomposition. A generic enter/exit traversal is not itself this bijection.

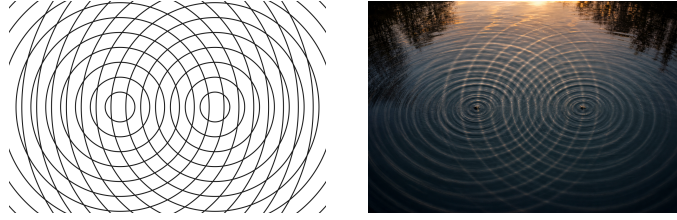


Figure 9: *interference*: skeleton (left) and the two-ripples scene (right). Two point sources with different periods; the scored content is the emergent large-scale pattern, not the wavelengths.

Home. Combinatorics (Catalan numbers and Dyck words); theory of computation (structural induction and the call stack).

Theorem. Dyck words of length $2n$, full binary trees with n internal nodes, and triangulations of a convex $(n+2)$ -gon are all counted by $C_n = \binom{2n}{n}/(n+1)$ (Euler–Segner; see Stanley, *Catalan Numbers*), with explicit bijections between any two families.

L.3 INTERFERENCE

Scene and skeleton. Figure 9.

Story (English translation).

In Evening-Glow Bay, a girl named Ali repairs blue cloth shades for fishing lamps. Behind the dye shop lives a troublesome roof-cat that keeps interrupting her work, while the tide keeps bringing odd little gifts to the water gate: a silver bell, a torn kite, the scent of herbs, a ferry, starlight, and at last two shadows. Ali turns the found things and the cat’s mischief into a fish-tailed lamp shade with a white button eye. When the fishermen light it, the whole shore laughs with delight, and the cat seems to bless the work too.

Form device. The Chinese original has exactly 24 sentences. Sentences 3, 6, 9, . . . , 24 open with the same fixed word *piānpīān* (“just then, of all things”), the cat’s mischief, period 3. Sentences 4, 8, 12, . . . , 24 close with a seven-character clause, the tide’s gift, period 4. The two patterns coincide only at sentences 12 and 24, a 3-against-4 beat run through two full periods of $\text{lcm}(3, 4) = 12$. Neither pattern alone is the motif; the motif is the emergent longer period.

Mathematics voice (translated).

Equation. $\sin \alpha + \sin \beta = 2 \sin \frac{\alpha+\beta}{2} \cos \frac{\alpha-\beta}{2}$

Definition. If f and g have periods p and q with p/q rational (take p, q positive integers), then $\text{lcm}(p, q)$ is a period of $f + g$ (not necessarily the least one; if p/q is irrational, $f + g$ is in general only quasi-periodic). For nearby frequencies the sum-to-product identity yields a slow envelope at the difference frequency, the beats that are the one-dimensional prototype of moiré fringes.

Home. Fourier analysis (beats, sum-to-product); number theory (least common multiple and the Chinese Remainder Theorem).

Theorem. If f has integer period p and g integer period q , then $\text{lcm}(p, q)$ is a period of $f + g$. When $\text{gcd}(p, q) = 1$, the phase pair $(n \bmod p, n \bmod q)$ visits all pq combinations exactly once every pq steps and returns to $(0, 0)$ (CRT), so the two patterns realign only every pq steps.

M FORM-DEVICE VERIFICATION. A COMPLETE WALKTHROUGH

This display walkthrough illustrates the self-reference story device using a retained historical judge record. The device operates on the *Chinese original*, which is shipped in the released dataset; the check there is character-exact. The corrected scored English stories instead have their selected

executable clauses and ambiguity flags recorded separately; we walk through this Chinese-corpus device because it is the most compact to display. This section presents an English-only schematic of the same verification. The title of the story is itself a complete assertion.

“The first characters of every sentence in this piece, read in order, are exactly this title.”

The title consists of exactly thirteen Chinese characters, and the story body has exactly thirteen sentences.

Step 0. The text. The thirteen sentences (delimited in the original by the ideographic full stop), in English translation:

1. This town sits by a river—peaches for sale in spring, chestnuts in autumn—and it also holds a tiny woodblock print shop.
2. Every one of the children’s primers comes from the shop’s craftsman, Gu Huai, whose knife is steadier than a needle.
3. Each night, late, Gu Huai lights a bean-oil lamp and carves the book blocks the children will collect in the morning.
4. Beside every rhyme he loves to add fish, wheat ears, and plump ducks, so that children learning their characters will not fear books.
5. From the first page on, he has split up a kind saying left by his late wife and quietly set the pieces where the knife first falls on each block.
6. The characters are tiny, hidden at the tips of leaves, on the point of a duck’s bill, in the first stroke of a wheat awn.
7. Even his little daughter A Sui does not know; she only feels those pictures are especially dear.
8. In the days when reading voices filled the street, the river rose fast and washed away several door plaques overnight.
9. Just then A Sui, delivering books, lost her way coming back, and huddled trembling under a rain awning with a bundle of new primers.
10. To find his daughter, Gu Huai called along the streets until his voice went hoarse.
11. At this moment a chestnut-selling grandmother opened a new primer, saw the little corner pictures lined up leading east, and remembered how A Sui loved to follow those little marks.
12. The marks led everyone to the old ferry, where A Sui crouched under a boat awning, still guarding the unwetted books.
13. Once the title labels had dried, the children came for their books as before, and Gu Huai simply polished his knife and went on setting that kind saying where the blade first falls.

Step 1. Extract and compare. The judge extracts the initial character of each of the thirteen sentences and concatenates them into a strip. Romanized (pinyin), with per-character glosses, the strip reads as follows.

#	Initial (pinyin)	Gloss	#	Initial (pinyin)	Gloss
1	<i>běn</i>	this	8	<i>dú</i>	read
2	<i>piān</i>	piece	9	<i>qià</i>	exactly
3	<i>měi</i>	every	10	<i>wéi</i>	is
4	<i>jù</i>	sentence	11	<i>cǐ</i>	this
5	<i>shǒu</i>	first	12	<i>biāo</i>	title (1st char.)
6	<i>zì</i>	character	13	<i>tí</i>	title (2nd char.)
7	<i>lián</i>	joined			

Read in order, “*this piece[’s] every sentence[’s] first characters, joined [and] read, [are] exactly this title*”, the strip is, character for character, the title itself (13 of 13). The comparison in the released judge record is exact string equality on the Chinese text; nothing in it depends on translation.

Step 2. Close the loop. The fact just verified in Step 1 is precisely the fact the title asserts. The text therefore contains a true statement about its own form whose truth is established by the text itself, a finite textual example of self-reference. This string-equality check illustrates the motif but is not an instance proving the formal diagonal lemma. There is no climb through narrative levels (so it is not *strange loop*), and the assertion refers to this very text rather than to a second system (so it is not *isomorphism*).

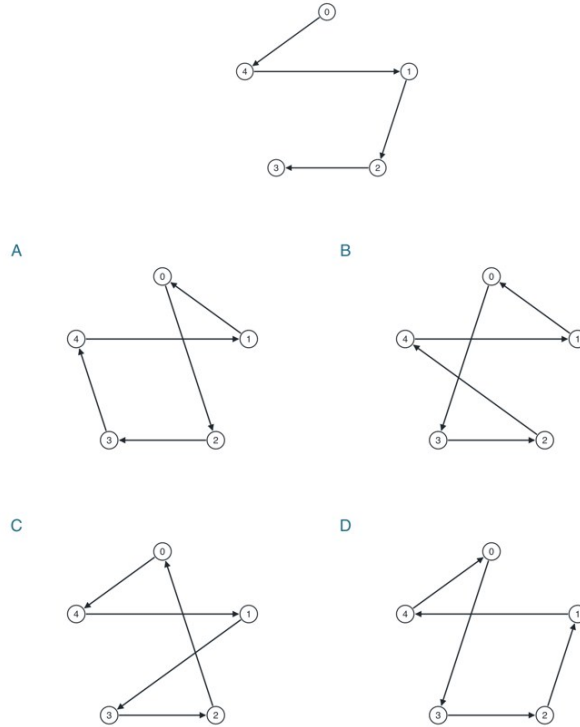
Device-judge output. The released judge record for this story reads, in translation, “1) Count: 13 sentences, each delimited by the ideographic full stop; 2) the sentence-initial characters, in order, concatenate to a string that agrees exactly with the title; 3) what the title asserts is precisely the fact verified in step 2, satisfying the self-referential fixed point.” The blind judge, shown only the story and the full 25-way menu, returned *self-reference*.

N WHAT THE MODELS RECEIVE

All pilot arms use one image and an explicit text prompt. Endpoint identification returns `motif_id`; endpoint statement, edit description, composition and intermediate-state tasks return `choice`. In the operation tasks, the image contains the source graph and candidate panels, and the text states the edit sequence. Serialized `source_edges` and `candidate_edges` are audit metadata, not additional model-visible text.

Image-withheld conditions delete both image fields from the item, producing plain-text HTTP content with the full options and JSON instruction. They do not send a blank placeholder or a dummy image. Family4 keeps the source image and ordered candidate motifs but requests `motif_id`, not an option letter. The following prompts are reproduced in full from retained requests; line wrapping and mathematical glyphs are typeset for reading. The shown JPEGs have the exact bytes used by the recorded image serialization.

N.1 EXACT GRAPH COMPOSITION



The top diagram is the starting directed graph. Numbers are vertex identities; arrowheads give edge direction. Apply the following operations in order.

1. Add the directed edge $3 \rightarrow 0$.
2. Relabel every vertex simultaneously by $0 \rightarrow 3$, $1 \rightarrow 1$, $2 \rightarrow 0$, $3 \rightarrow 2$, $4 \rightarrow 4$. Keep each edge attached to its relabeled endpoints.

Choose the final graph from panels A–D. Judge exact labeled edges, not the drawing shape. Return a JSON object with key "choice" whose value is the selected option letter (A, B, C, or D).

The graph edge lists used by the validator are not appended to this prompt.

N.2 THEOREM MATCHING WITHOUT AN IMAGE

No image is available. Select the most plausible answer using only the following options.

[A]

Kleene's Second $\square$ Theorem (1938): for every total computable function f , there exists an index e such that $\varphi_e = \varphi_{\{f(e)\}}$; corollary: in any Turing-complete programming language with an acceptable numbering, there exists a program that $\square$ (a $\square$).

[B]

Cantor's $\square$ closed interval theorem: If a sequence of bounded closed intervals satisfies $[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$ for every n , then $\bigcap_{n \geq 1} [a_n, b_n]$ is nonempty; if, in addition, $b_n - a_n \rightarrow 0$, then the intersection is exactly the singleton $\{\xi\}$, where $\xi = \sup_n a_n = \inf_n b_n$.

[C]

Hairy Ball Theorem (a corollary of Poincaré–Hopf): On every even-dimensional sphere S^{2n} ($\chi = 2 \neq 0$), every continuous tangent vector field has a zero; moreover, on S^2 one can construct a tangent vector field with exactly one zero (pull back a constant vector field on the plane by stereographic projection; the unique zero is at the north pole and has index 2). Thus a zero is unavoidable, and it can be realized as the only zero.

[D]

Dyck words of length $2n$, full binary trees with exactly n internal nodes, and triangulations of a convex $(n+2)$ -gon are three families of objects all counted by the Catalan number $C_n = \binom{2n}{n} / (n+1)$ (Euler–Segner enumeration; see Stanley, Catalan Numbers), and there are explicit bijections between each pair.

Output JSON only (no code block): `{"choice":"one of A-D"}`

No image accompanies this complete prompt. The visible attributions illustrate that selected masking does not eliminate every lexical cue.

N.3 FAMILY4 RECOGNITION



Select the intended structural motif of this image from the four options.

Options:

- self-reference: The object contains a sentence describing itself, and the description checks out verbatim as true — it points back to itself in one step, with no climb through levels and no reference to a second system.
- nesting: Each time a new level is entered, the previous level is never returned to; depth only ever increases.
- symmetry-breaking: One rule covers the whole field, with exactly one precisely locatable violation; the violation is unique and everywhere else is without exception.
- recursion: The whole is generated from smaller copies of itself by the same rule; however many levels you enter, that many you exit, terminating at a base case and then returning level by level along the original path.

Output JSON only (no code block): `{"motif_id":"..."}`

O EVALUATION PROMPTS

The object-layer templates below summarize the frozen question wording. Substitution fields contain the complete frozen menu, story title and body, or candidate statements. The entire substituted request is hashed and checked against the raw response record. Matching candidate blocks use bracketed A–D labels and full text; motif-identification menus list the candidate IDs and their descriptions. All templates append the same single-field JSON instruction appropriate to their answer type.

Table 30: Frozen object-layer question text before the JSON instruction. Braced names denote substitution fields, not material hidden from the model.

Task	Question and substituted content
Scene / skeleton	Select the intended structural motif of this image from the menu. Options: {menu}.
Family4	Select the intended structural motif of this image from the four options. Options: {menu4}.
Story	Select the intended structural motif enacted by the form of the story. Story: {full title and body}. Options: {menu}.
Theorem	Select the motif most closely associated with this mathematical statement. Statement: {text}. Options: {menu}.
Scene–story	Select the story whose form most closely matches the intended structure of the image. {full options}.
Scene–theorem	Select the mathematical statement most closely associated with the intended structure of the image. {full options}.
Story–theorem	Select the mathematical statement most closely associated with the structure enacted by this story. Story: {full title and body}. {full options}.
Image withheld	No image is available. Select the most plausible answer using only the following options. {full options}.

The appended instruction is `Output JSON only (no code block):` followed by `{"motif_id":"..."}` for identification, including Family4, or `{"choice":"one of A–D"}` for matching. The strict parser requires exactly that field with a valid value. The compatibility parser can extract an embedded field, then a task-appropriate motif identifier or the last standalone valid option letter. Its exact behavior is frozen and both strict and compatibility outcomes are stored. Pilot and auxiliary items carry their complete explicit prompts in their own immutable item files.

Table 31: The corrected frozen answer menu, reproduced from the administered item file. Wording is part of the model input; global invariant descriptions are not certificates of finite images.

Motif	Administered description
self_reference	The object contains a sentence describing itself, and the description checks out verbatim as true — it points back to itself in one step, with no climb through levels and no reference to a second system.
recursion	The whole is generated from smaller copies of itself by the same rule; however many levels you enter, that many you exit, terminating at a base case and then returning level by level along the original path.
nesting	Each time a new level is entered, the previous level is never returned to; depth only ever increases.
cycle	Following same-level causal arrows through $n \geq 3$ mutually distinct events returns you to the starting event; one lap restores everything verbatim (no escalation, no growth, no flip), all points equal in status, no hierarchy.
symmetry	The sequence of events read backwards is item-for-item identical to reading it forwards; folded at the temporal midpoint, it coincides point by point.
alternation	Exactly two states strictly take turns: adjacent items always differ, items one apart are always the same, the period is constantly two, without a single exception.
spiral	Each time it comes back around to the same bearing it grows one notch by the same fixed ratio, never returning to the original scale.
convergence	Multiple threads, mutually unaware, each advance on their own and finally all merge into the same endpoint, and there it ends.
bottleneck	Each of the two shores is internally multiply connected, but every path between the shores must pass through one and the same midway point, continuing separately beyond it; deleting that point splits the graph into exactly two still-connected halves, rather than shattering it into isolated pieces.
symmetry-breaking	One rule covers the whole field, with exactly one precisely locatable violation; the violation is unique and everywhere else is without exception.
strange_loop	Climbing step by step up the hierarchy of 'who narrates whom', the final step nevertheless lands back at the starting point; no consistent level can be assigned.
figure_ground	Figure and ground complementarily fill the whole canvas and the ground itself forms a figure: the second half takes the first half's ground as its figure and the first half's figure as its ground, while the narrated events and their order remain unchanged (not reversed order, and not a translation between two vocabularies).
isomorphism	Between two things made of entirely different material there exists a one-to-one dictionary preserving all operations and relations, translatable term by term in the same order, without a single exception.
mobius	Traveling once around the band returns you to the starting place with orientation flipped (left and right exchanged); only after two laps is everything restored point by point.
braiding	Three or more strands cross adjacently several times; the final arrangement of the ends equals exactly the composition of all the crossings in order, and the order cannot be freely exchanged.
interference	Two repeating schedules or signals are superposed, producing an interaction pattern. For discrete periods p and q , their joint phase repeats after $\text{lcm}(p,q)$; this is a common period, not generally the least period of the sum. The story device checks periods three and four explicitly.
tiling	Shift the entire pattern by one cell and it coincides exactly with the original, seamless and without overlap, without a single exception.
hub	Between any two endpoints stands only the center point; remove the center and the whole graph shatters into isolated points.
diagonalization	A table enumerated row by row; the new row built by flipping each entry along the diagonal is guaranteed to differ from row n at position n , hence differs from every row — forever outside the table. Unlike self_reference (an object pointing back to itself in one step): the diagonal element does not talk about itself; the attack targets the table's entire enumeration, not itself. Crucially unlike symmetry-breaking (exactly one isolated violation in an otherwise regular field): here it is not one anomalous cell but n systematic differences along a moving diagonal — row n differs at position n , the position shifting right by one each row, at least one specified difference per row, and this construction guarantees 'not equal to any row'; symmetry-breaking is a single-point flaw that neither escapes nor refutes any enumeration.

Motif	Administered description
<code>infinite_descent</code>	Every case forces a strictly smaller case of the same kind, and the natural numbers admit no bottomless descending chain, so the first case does not exist at all — the proof’s blade swings backwards: not advancing to an end, but cancelling the beginning. Exactly opposite to recursion: recursion exits as many levels as it enters and stands on the existence of a base case; descent stands on the absence of one and condemns every witness to nonexistence. Unlike <code>zeno_halving</code> : Zeno’s steps shrink by a fixed ratio, total length finite, the journey completes; descent’s decrease is arbitrary but each step is an integer, the chain is forbidden by well-ordering, the conclusion is ‘impossible’. Unlike <code>contraction.fixed_point</code> : contraction iterates forever and converges to a genuinely existing fixed point; the descending chain converges to nothing, must hit floor 1 in finitely many steps, so the whole chain together with its starting point is judged nonexistent. Opposite in direction to <code>ordinal_ascent</code> : ordinal ascent is unbounded upward with every level real; descent is bounded below with the whole chain illusory. Unlike <code>invariant_barrier</code> : no conserved quantity is invoked; the only killing move is that strict decrease cannot go on forever.
<code>coprime_orbit</code>	On a ring of n points, jump with a fixed stride k , connecting as you go: if the stride is coprime to the ring length, a single stroke visits every point before closing; if not, the ring splits into gcd parts. Unlike <code>cycle</code> (adjacent arrows closing head-to-tail in one lap): the large stride weaves a star shape, traversal is conditional on coprimality, and there is a failure mode — not coprime means splitting. Unlike <code>interference</code> (two independent periods superposed, an lcm-sized larger period emerging): here there is only one orbit and one number-theoretic condition. Unlike <code>hub</code> (relaying through a center): there is no central point at all; traversal is guaranteed purely by $\gcd(k,n)=1$.
<code>pigeonhole</code>	More items than cages forces two into the same cage: the collision is proved certain, yet no one can point to where it happens. Unlike <code>convergence</code> (all threads, knowingly or not, ending at the same KNOWN point): here counting only guarantees that at some UNSPECIFIED place at least two meet; no witness is constructed. Unlike <code>bottleneck</code> (every path forced through the same midway point): there is no forced path and no relay hub, only a single count.
<code>projective_duality</code>	Within a single system, two classes of roles are swapped wholesale and every relational proposition remains true; swapping twice restores the original verbatim (an involution, $\sigma^2=id$). Exclusions: unlike <code>isomorphism</code> (a relation-preserving dictionary between two heterogeneous materials) — duality is a role swap inside one and the same material and is necessarily an involution; unlike <code>figure_ground</code> (figure and ground complementarily filling one canvas) — duality produces two pictures each complete in itself, not relying on negative space; unlike <code>transformed_canon</code> (the same theme reappearing in a lagged voice under a fixed transformation) — duality is a wholesale sequential restatement that swaps role vocabulary while truth values stay unchanged, not a voice-level pitch/time transformation; unlike <code>chirality</code> (mirror image and original cannot be superposed, the point being ‘it does not fit’) — the two dual statements both hold within the same system and the swap restores everything exactly; unlike <code>mobius</code> (the flip requires traveling one full lap along a path) — duality has no path, it is an immediate wholesale role swap; unlike <code>symmetry</code> (item-by-item reversed palindrome) — duality does not reverse order, it only swaps roles.
<code>contraction_fixed_point</code>	The same rule repeatedly transforms its own output toward a state that it leaves unchanged. A strict contraction claim requires a declared complete metric space and a factor $q < 1$; the story device uses a finite closed chant orbit and an explicitly verified ultrametric. Other images are motif-level illustrations of iterative stabilization.

Motif	Administered description
<code>aperiodic_tiling</code>	Everywhere obeying the same local matching rules, seamlessly covering the whole plane, yet no translation whatsoever makes the whole picture coincide with itself; at the same time any local patch reappears within a bounded gap everywhere — ‘seen before, never repeated’. Exclusion clauses: unlike tiling (shift one cell and the whole coincides; here it is provable that no translational period exists at all); unlike symmetry_breaking (a periodic background with exactly one locatable violation; here there is no violation anywhere, yet no period exists at all); unlike spiral (phi here is the frequency ratio of the two tile kinds, not level-by-level fixed-ratio growth; the figure has no self-similar scaling center); unlike transformed_canon (the same theme reappearing verbatim under a fixed transformation; here no block ever recurs exactly at a fixed interval); unlike interference (two periods superposed still yield an lcm period; here the limiting structure’s frequency is irrational, so there is never a period).

Description length and retained contrast terms. The supplied menu contains structural descriptions rather than uniformly short invariant lines (lengths are tabulated in Appendix C). Some descriptions retain older contrast names, namely `infinite_descent` mentions `invariant_barrier`, `ordinal_ascent` and `zeno_halving`; `projective_duality` mentions `transformed_canon` and `chirality`; `aperiodic_tiling` also mentions `transformed_canon`. None is a selectable ID in the frozen menu. They remain in the quoted evaluation text; removing them would change the evaluated input, so their cueing effect is a limitation of this version. Masking deterministically replaces only the listed motif-revealing terms. Named attributions, years, journals and formula cues can remain, as the complete image-withheld example shows.

P SKELETON GALLERY

Figure 10 shows the programmatic skeleton for each of the 25 motifs: plain 2-D vis-graphs rendered from coordinate-level recipes. These images serve as the skeleton-voice recognition items and as references in the skeleton-conditioned generation branch. Separate text-to-image and deterministic-repair paths are described in Appendix C.

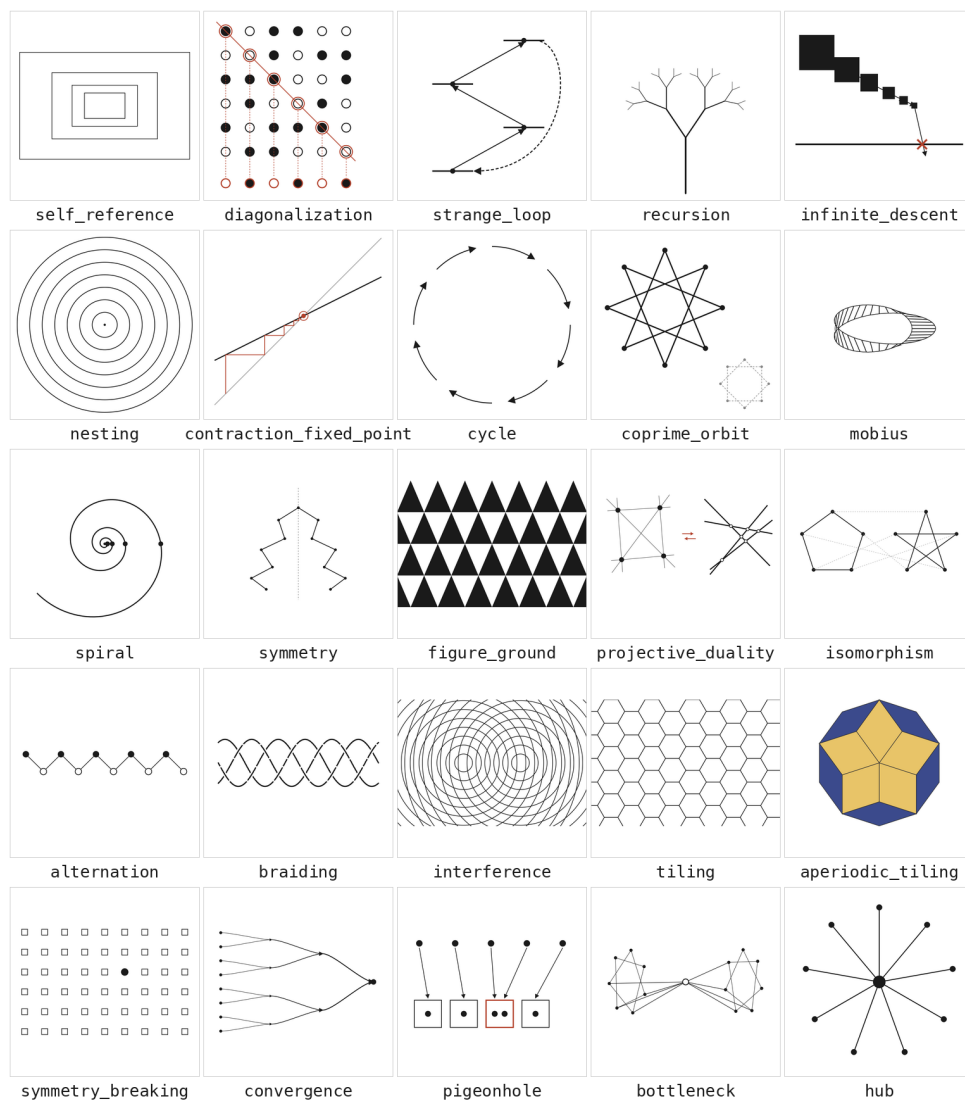


Figure 10: All 25 motif skeletons, in the conceptual-arc order of Figure 4. Every rendering parameter (node counts, depths, angles, sizes) is a nuisance variable; only the structure is scored.

Q LIMITATIONS AND REPRODUCIBILITY

The benchmark scores the intended motif, not a proof that each example uniquely realizes a mathematical invariant. Finite images cannot certify all global properties; mathematical statements can be associated with a motif without characterizing every realization. Executable story clauses cover explicit finite predicates, while semantic interpretation and unrestricted equivalences remain only partially checked. Generation and screening records are provenance, not independent semantic validation, and no independent human semantic validation was performed for this version. The evaluated collection retains ambiguity and scope flags; successful program checks do not certify unrestricted interpretation. The four-voice layout crosses the visual or language channel with descriptive or formal form as a presentation design; it is not a calibrated difficulty factorization, and no voice is claimed to be semantically certified.

The collection reuses motifs, scenes and story options across tasks. It is not an independent train/test split, and connected-component clustering can leave too little independent variation for an interval. Motif and source-cluster analyses are sensitivity estimates, not a remedy for every shared-content dependency. The writer/judge aliases `gpt-5.5:stable` and `gpt-5.2:stable` overlap with evaluated API naming families; generator-evaluator affinity and training-data overlap are not isolated. The naturalization and comparison manipulations also change rendering fidelity or available target resolution, limiting causal claims.

All reported models completed their required suites and response-integrity checks. Input and model-source manifests, retained replies and analysis receipts support the results. The model roster includes documented availability-driven amendments; it was not preregistered before all inference. Hosted API calls or different hardware can produce different replies despite fixed decoding settings.

Reproducibility has distinct targets. Score replay uses frozen items and retained terminal replies to reconstruct reported metrics. Inference replication sends the same serialized requests to a specified model and may yield different replies. Construction replication regenerates release files and requires separate semantic checks. A new image or changed API is not an exact reproduction of this evaluation. Comparisons should report input and response-policy versions with the model identifier, retaining compatibility and strict scoring as separate views.

Future work can test disjoint sources and motifs, vary generators independently of evaluators, and collect human judgments of ambiguity and device fidelity. Those studies would address generalization and generator affinity beyond what the present conditions establish. They should preserve versioned label policies and report ambiguous cases rather than silently revising the evaluated answer key.