LEVI: BENCHMARKING VISUAL IDENTITY FIDELITY WITH VERIFIABLE LOW-ENTROPY CONSTRUCTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern generative models can produce visually plausible images while violating exact identity constraints. Evaluating such failures requires distinguishing violations of the requested identity from permissible variation in appearance. We introduce Low-Entropy Visual Identity (LEVI), which characterizes visual identities determined by compact facts and explicit rules. The low-entropy requirement concerns identity uncertainty, while appearance may remain diverse. We formulate identity fidelity in terms of the probability that a conditional image distribution assigns to the identities permitted by the prompt. Visual Identity Spaces (VIS) formalize this distinction by specifying admissible identity values and which different realizations express the same identity. Building on this formulation, we develop Verifiable Visual Grammar (VVG), a declarative language whose typed primitives and compositional constraints describe identity-bearing structures. Using this grammar, we build a benchmark for image generation. It assesses each output against the task's identity requirements, drawing evidence from the final pixels. The curated benchmark contains 1,375 tasks spanning 120 base identity spaces across six domains. Among ten evaluated model configurations, the highest whole-task pass rate is 58.69%. These results reveal a persistent gap between visually plausible images and faithful realization of their specified identities.

1 INTRODUCTION

Modern generative models have advanced in visual quality and instruction following, translating complex text requests into detailed, coherent images. Their capabilities now extend from natural scene synthesis to visual creation with more demanding content and layout requirements, such as integrating text and graphics into posters and infographics (Google, 2026; ByteDance Seed, 2026). Improvements in overall presentation, however, do not always ensure that critical details satisfy the request (Zhang et al., 2025; Kamath et al., 2025). Figure 1 illustrates plausible images whose identities are altered: a sign retains a realistic appearance while its character sequence changes. Similar discrepancies affect objects whose identities depend on geometry or relations among their elements. These details may occupy a small part of the image yet change what the object expresses, requiring assessment beyond overall visual quality.

These failures involve a distinction between identity certainty and appearance diversity. For the text in Figure 1, the font or material may vary while the specified character sequence remains fixed. Global metrics such as FID and CLIPScore (Heusel et al., 2017; Hessel et al., 2021) and reference-based similarity do not separate a wrong identity from a legitimate change in appearance, and recent benchmarks with executable checks (Zhang et al., 2025; Liu et al., 2026a; Zhao et al., 2026) do not define when different-looking outputs express the same target (Table 1). We model visual identity explicitly and use the conditional entropy chain rule to distinguish these two sources of uncertainty. When a condition fully specifies the identity, the desired identity uncertainty is zero, although appearance can remain diverse. Since a model that always draws the same wrong word also has zero identity entropy, we measure identity fidelity: the probability of realizing an identity permitted by the prompt, which gives equal credit to every valid appearance. Evaluating this property requires specifying which facts determine identity and which changes preserve it.

We introduce Low-Entropy Visual Identity (LEVI) , a benchmark for image generation in which identity is tightly constrained while appearance remains diverse. To specify what must be preserved,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

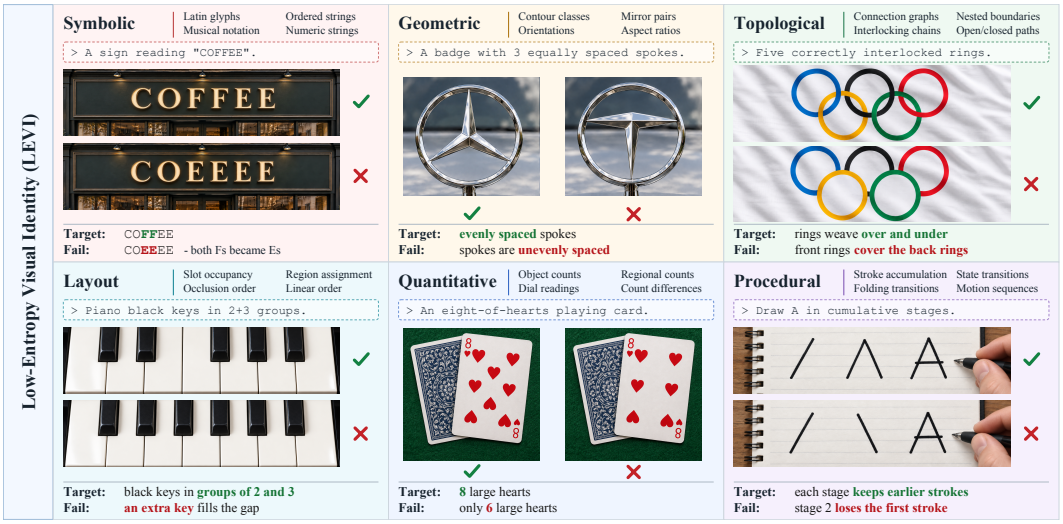


Figure 1: Visual plausibility can coexist with identity errors. Each domain pairs matching (✓) and failing (×) examples for one prompt. Target/Fail rows identify the mismatch; header entries list representative VIS. All examples are conceptual.

we formalize Visual Identity Spaces (VIS), which describe admissible identity values and define when different images express the same identity. This distinction allows us to assess whether an image preserves a specified identity while accepting permissible variation in appearance. A VIS may describe a standalone identity, while compatible spaces can be composed under shared constraints to express identities that depend on several components. Defined transformations and projections relate identities within and across spaces.

To turn these spaces into benchmark tasks, we develop Verifiable Visual Grammar (VVG), a declarative language for describing identity-bearing structures. Its typed components and rules specify how elements can be combined and which relations must hold, giving each description an identity under VIS semantics. Using VVG, we construct a benchmark spanning six identity domains: Symbolic, Geometric, Topological, Layout, Quantitative, and Procedural, illustrated in Figure 1. Tasks test both individual and combined identity requirements. Evaluation reads evidence from the final image and checks whether it satisfies the target specified at construction, linking task definition and scoring through the same identity semantics. Table 1 compares LEVI with existing image-generation benchmarks in terms of visual targets, rule requirements, verification, and identity modeling. At the identity level, we examine whether a benchmark supports constructing verifiable visual structures with explicit semantics, expressing standalone and composed identity requirements, and treating visually different realizations of the same identity as equivalent during evaluation.

Our contributions are as follows:

- We introduce LEVI and formalize visual identity fidelity. By modeling visual identity, we use conditional entropy to distinguish identity uncertainty from appearance diversity. VIS specifies admissible identities and their equivalent realizations, allowing fidelity to be defined as the probability of realizing the specified identity.
- We develop VVG to turn identity semantics into constructible and verifiable visual requirements. Typed components, relations, and compositional constraints express identity-bearing structures, supporting task construction from standalone and composed identity requirements and verification of generated images under the same semantics.
- We construct the LEVI benchmark with 120 base identity spaces and 1,375 curated tasks across six domains. Among ten evaluated model configurations, the highest whole-task pass rate is 58.69%. Identity-space analysis reveals differences obscured by overall rankings, while composition counts alone do not explain difficulty.

Table 1: Comparison with recent image-generation benchmarks. ✓ denotes explicit support in the primary protocol and × denotes no explicit support. LEVI denotes the proposed design.

Benchmark	Structured domains	Rule-governed targets	Executable checks	Verifiable construction	Identity-space expressivity	Identity equivalence
STRICT	×	×	✓	×	×	×
MMMG	✓	×	✓	×	×	×
StructBench	✓	×	×	×	×	×
SpatialGenEval	×	×	×	×	×	×
MathGen	×	✓	✓	×	×	×
GraphicDesignBench	✓	×	✓	×	×	×
RIGBench	✓	✓	×	×	×	×
GenPuzzle	✓	✓	✓	×	×	×
TRACEBench	×	✓	×	✓	×	×
LEVI (ours)	✓	✓	✓	✓	✓	✓

2 Related Work

Image-generation evaluation. Image-generation benchmarks increasingly assess whether an image satisfies specific requirements beyond overall visual quality. Global measures such as FID and CLIPScore summarize distributional similarity or image–text alignment (Heusel et al., 2017; Hessel et al., 2021), whereas TIFA, GenEval, and T2I-CompBench++ examine objects, attributes, counts, and relations (Hu et al., 2023; Ghosh et al., 2023; Huang et al., 2023). GenEval 2 and FineGRAIN further investigate fine-grained assessment and its agreement with human judgment (Kamath et al., 2025; Hayes et al., 2025). Knowledge and spatial benchmarks extend the target to factual dependencies and spatial relations (Huang et al., 2025; Luo et al., 2025; Wang et al., 2026d). Reasoning-oriented evaluations require deriving the visual content from implicit instructions or contextual rules (Sun et al., 2026; Liu et al., 2026b; Zhao et al., 2026). These developments establish the importance of precise visual requirements. LEVI focuses on their identity semantics: which facts determine the target and which changes preserve it.

Visual identity and target specification. Defining a visual target requires separating identity from its appearance. Reference-conditioned benchmarks assess whether a generated subject or product retains the properties of its reference (Peng et al., 2025; Singhania et al., 2025; Khanna et al., 2026; Zheng et al., 2026). MMIG-Bench combines identity, semantic, and aesthetic assessment (Hua et al., 2025). For text, graphics, and mathematical images, target properties can instead be specified through strings, structural descriptions, and task-specific constraints (Zhang et al., 2025; Zou et al., 2024; Nishina & Matsui, 2024; Deganutti et al., 2026; Zhuo et al., 2026; Liu et al., 2026a). TRACE-Bench makes the link between specification and assessment explicit: a compositional operator formula drives both multi-reference task construction and aligned evaluation (Wang et al., 2026a). LEVI develops this connection at the level of identity spaces. VIS specifies admissible identities and equivalent realizations, including standalone spaces and their compatible compositions; VVG expresses structures under those semantics. Table 1 distinguishes this identity formulation from task scope, target rules, and evaluation methods.

Verification of generated images. Verification must recover the relevant facts from the generated pixels. ControlNet++ checks re-extracted spatial conditions, while DiffQRCoder assesses whether a rendered code remains readable (Li et al., 2024; Liao et al., 2024). MathGen and ProVisE use external procedures to assess visual outputs (Liu et al., 2026a; Wang et al., 2026c); GenPuzzle combines programmatic checks for discrete outputs with task-specific visual judging (Zhao et al., 2026). VBVR applies rule-based scorers to generative visual reasoning (Wang et al., 2026b), and SpatialReward connects grounded observations with spatial relation checks (Zhou et al., 2026). The reliability of the observation procedure remains consequential: MJ-Bench evaluates systematic weaknesses of multimodal judges (Chen et al., 2024). LEVI separates visual readout from the fixed identity check. The reader supplies observed evidence, and the checker applies the same identity semantics used in construction, preserving the distinction between an identity violation and a measurement failure.

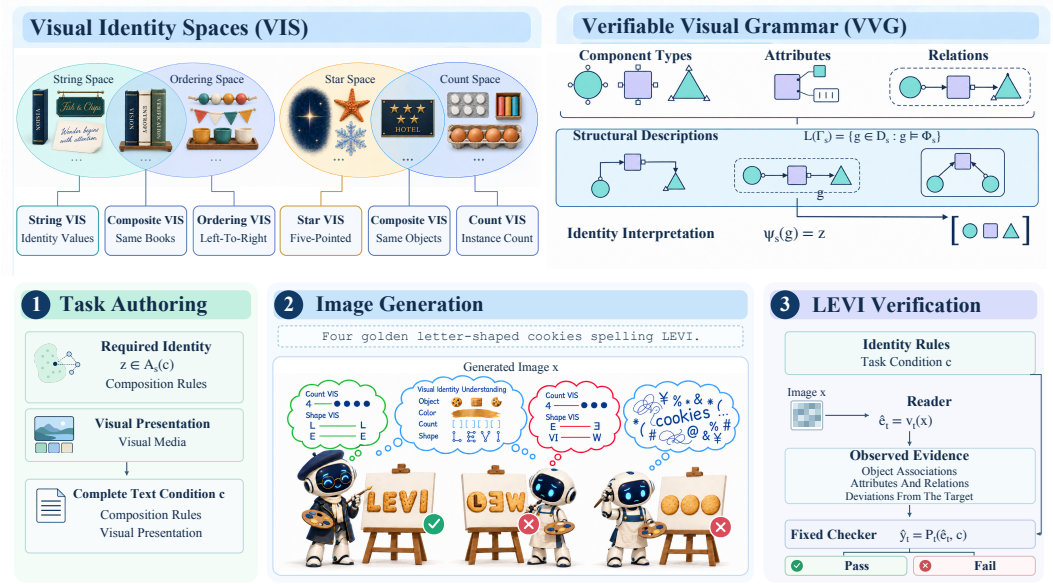


Figure 2: From visual identity to output-based verification. Top: VIS groups equivalent realizations and combines compatible identities; VVG describes a structure g whose interpretation is an identity z. Types, attributes, and relations jointly define admissible descriptions. Bottom: task authoring supplies a complete text condition, generation produces pixels, and verification checks evidence against the shared identity conditions. Robots illustrate conceptual distinctions in identity understanding.

3 METHOD

We distinguish uncertainty about identity from diversity in its visual realization. Identity fidelity measures agreement with the target; VIS organizes admissible values and their equivalence; VVG expresses these requirements as structures that support construction and evaluation.

3.1 LOW-ENTROPY VISUAL IDENTITY

Image generators model a conditional distribution $q_\theta(X \mid C)$ over images X given text C. The condition constrains visual identity while allowing variation in its realization. Let s specify the identity rules and let the deterministic map ϕ_s assign each image its expressed identity, giving $Z_s = \phi_s(X)$. Using discrete representations, with continuous quantities quantized at a specified resolution, the conditional entropy chain rule (Cover & Thomas, 2006) gives

$$H(X \mid C = c) = H(Z_s \mid C = c) + H(X \mid Z_s, C = c). \tag{1}$$

Under q_θ, the first term measures uncertainty about the expressed identity, and the second measures variation among realizations of a given identity. Low-Entropy Visual Identity (LEVI) concerns tightly specified identities with diverse visual realizations. For an exact target $z^\star$, the desired generation distribution assigns all probability to its realizations, giving zero identity entropy while retaining permissible visual variation.

Entropy measures the concentration of generated identities. Identity fidelity (IF) measures their agreement with the condition. Writing $A_s(c)$ for the permitted identity values, we define

$$\mathrm{IF}_\theta(c, s) = \Pr_{X \sim q_\theta(\cdot \mid c)} \left[\phi_s(X) \in A_s(c)\right]. \tag{2}$$

For an exact target, $A_s(c) = \{z^\star\}$. This probability aggregates all permitted realizations, giving equivalent appearances equal credit.

3.2 VISUAL IDENTITY SPACES

A Visual Identity Space (VIS) specifies which visual differences change identity. For an ordered sequence of symbols, a painted △ ∘ □ and an engraved version can express the same identity, while exchanging the circle and square changes it. The declared rules specify which changes of appearance are allowed. Under identity rules s, let $\mathcal{X}_s^{\mathrm{adm}}$ contain images expressing admissible identities. The relation $x \sim_s x'$ groups images that express the same identity into an equivalence class. These classes form the corresponding VIS:

$$\mathcal{Z}_s = \mathcal{X}_s^{\mathrm{adm}}/\!\sim_s, \qquad \phi_s(x) = [x]_{\sim_s}, \quad x \in \mathcal{X}_s^{\mathrm{adm}}. \tag{3}$$

Each point $z \in \mathcal{Z}_s$ represents one identity, and ϕ_s maps its equivalent image realizations to that point. Variation within a class preserves identity; movement between classes changes it. Images outside the admissible domain receive the value $\perp$, so Equation 2 remains defined for every output.

An identity space can stand alone or combine with other spaces. When component values $z_1, \ldots, z_m$ completely determine a composite identity, its space can be written as

$$\mathcal{Z}_s = \left\{ (z_1, \ldots, z_m) \in \prod_{j=1}^{m} \mathcal{Z}_{s_j} : \chi_s(z_1, \ldots, z_m) = 1 \right\}, \tag{4}$$

where $\chi_s = 1$ selects combinations that can be jointly realized. This representation includes component identities and every relation that determines their joint identity. Components can belong to the same or different domains; compatibility constraints allow them to depend on one another. The spaces may be discrete, continuous, or mixed.

Figure 2 gives two concrete combinations. The book spines jointly specify the strings `VISION`, `ENTROPY`, and `VERIFICATION` and their left-to-right order on the same books. The hotel sign combines a five-pointed star identity with a count of five instances of that shape. Composition binds these requirements to the same objects, retaining both component identities and their relations. The six identity domains organize these visual facts; each can contain multiple VIS with its own identity values and equivalence rules.

3.3 VERIFIABLE VISUAL GRAMMAR

VIS specifies the meaning of a visual identity. Verifiable Visual Grammar (VVG) expresses that identity through a declarative structural description. A grammar Γ_s defines component types and interfaces, assigns attributes to their components, and constrains valid relations and combinations. In the sequence example, components are symbol occurrences, each owns a symbol value, and successor relations specify their order. Constraints require values from the declared alphabet and one complete linear sequence. A description g fixes these values and relations; the language $\mathcal{L}(\Gamma_s)$ contains the descriptions satisfying the constraints. An interpretation map assigns each description its identity:

$$\psi_s : \mathcal{L}(\Gamma_s) \longrightarrow \mathcal{Z}_s, \qquad g \longmapsto z \tag{5}$$

Here, g specifies a structure and $\psi_s(g)$ gives its meaning in VIS. Descriptions that map to the same point express the same identity. A target $z^\star$ can therefore be specified by any valid description satisfying $\psi_s(g) = z^\star$, allowing construction to vary while preserving the intended identity. Grammar membership and task acceptance are distinct: both △ ∘ □ and △ □ ∘ are legal sequences, but only the first satisfies an exact request for that order.

A description is verifiable when each identity-defining attribute and relation is linked to observable image evidence and an explicit decision rule. These rules specify what must hold in the generated image, allowing the constraints used in construction to be checked on its visual realization. For the sequence, the reader recovers the visible symbols and their order, and the checker compares this evidence with the requested sequence under the same identity rules. Figure 2 connects these semantics to task authoring and output-based verification.

Instantiation assigns component values, composition establishes their declared relations, and transformations modify the description consistently with VIS operations. Each resulting structure therefore has an explicit identity interpretation.

Visual media give the resulting structures concrete form. Their appearance can vary within the equivalence class of the specified identity. The required identity and composition rules are expressed in the complete text condition c, together with the visual presentation. This condition is passed to image generation, while the structural description supplies the target for subsequent verification.

3.4 Identity-Grounded Benchmarking

The benchmark evaluates identity fidelity by connecting the target specified during construction to the content of the generated image. For an exact identity, $\psi_s(g) = z^\star$ specifies that target and $\phi_s(x) = z^\star$ defines successful realization. The target description determines the attributes and relations relevant to this comparison, while the equivalence rules specify the visual variation that evaluation accepts.

Each identity requirement is associated with evidence that can be observed in the final image and a rule for deciding whether it is satisfied. Readout records the relevant content together with its object associations, so that attributes and relations can be assigned to the components they describe. This evidence preserves the generated structure, including its deviations from the target. The checker then applies the target requirements to these observations.

Figure 2 connects this evidence-based assessment to task construction and image generation. For task t, a reader v_t extracts evidence $\widehat{e}_t$ from image x. A fixed checker P_t assesses it under the identity rules and task condition c. For resolved evidence,

$$\widehat{e}_t = v_t(x), \qquad \widehat{y}_t = P_t(\widehat{e}_t, c), \quad \widehat{y}_t \in \{0, 1\}. \tag{6}$$

The semantic map ϕ_s defines the identity expressed by an image, while the reader and checker provide an evidence-based estimate of agreement with the target. For a composite identity, the decision combines requirements on individual components with those on their joint structure. Successful realization requires these conditions to hold together. The same evidence also identifies which component or relation accounts for an identity violation, connecting the overall judgment to interpretable differences in the output.

Equivalence rules determine how variation in appearance enters this assessment. Realizations within a permitted identity class satisfy the same identity criterion even as their appearance varies. Applying these rules consistently across construction and checking gives model comparison a common semantic basis, while retaining the identity-specific requirements of each task.

4 Evaluation and Analysis

We evaluate visual identity fidelity at three levels: overall performance across models, variation across identity domains, and behavior within individual and composite VIS. Qualitative comparisons connect these scores to the visible facts that distinguish faithful realizations from errors.

4.1 Benchmark and Evaluation Setup

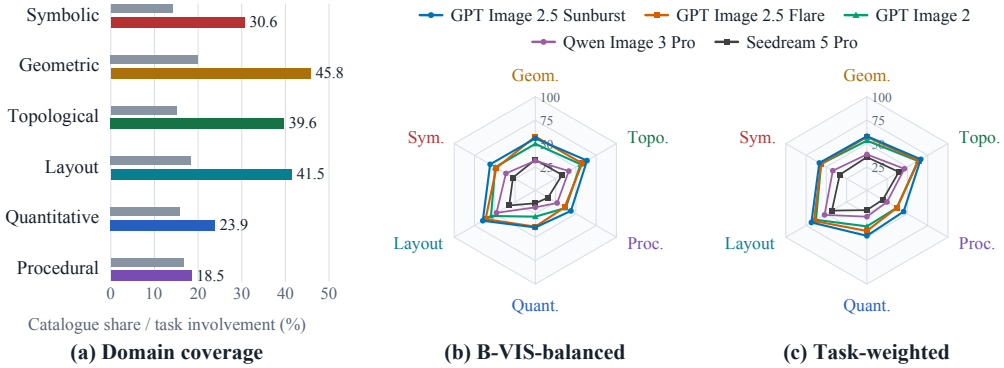
The curated benchmark contains 1,375 tasks spanning all 120 base visual identity spaces (B-VIS), each assigned to one of six domains. Tasks instantiate individual spaces or combine compatible identities and relations into composite VIS. A task can therefore contribute to several domain subsets. Figure 3 contrasts the catalogue's domain composition with task involvement in the evaluation set and compares the models' performance profiles.

The ten configurations in Table 2 each receive the complete text prompt without reference images or construction records. Evaluation covers 13,750 model–task cases, with one output per case.

The reported metric is the fraction of retained tasks whose required conditions all pass automatic assessment. GPT-6-Sol with medium reasoning effort assesses the generated image against the prompt and its identity criteria. Missing images and unresolved cases receive zero credit and remain distinct from observed identity violations.

The challenge set was curated post hoc for redundancy, ambiguity, sampling balance, and observed ease across models. Reported rates are conditional on this selection. Task-specific judging instructions include two revisions, whose equivalence has not been established; their provenance is

Figure 3: Domain coverage and top-five model profiles. (a) B-VIS catalogue share (grey) and task involvement (colour). Whole-task pass rates (%) are (b) B-VIS-balanced or (c) task-weighted. Radar axes share a 0–100% scale; domain subsets overlap.



(a) Domain coverage **(b) B-VIS-balanced** **(c) Task-weighted**

retained. Independent human agreement has not been measured. One missing output and one output with mismatched model provenance receive zero credit, leaving 13,748 assessed outputs within the full denominator.

4.2 Overall Performance

Table 2 ranks the ten configurations by overall whole-task pass rate. GPT Image 2.5 Sunburst leads at 58.69%, followed by Flare at 55.71% and GPT Image 2 at 53.24%. Qwen Image 3 Pro ranks fourth at 39.71%, 13.53 percentage points behind GPT Image 2. Even the leading configuration fails to receive credit on more than two fifths of tasks.

Table 2: Whole-task pass rates (%, ↑) on the post-hoc curated challenge set. Medals mark the top three overall; **bold**/underline mark the best/second-best per column.

Rank	Model	Overall	Sym.	Geom.	Topo.	Layout	Quant.	Proc.
1	GPT Image 2.5 Sunburst	**58.69**	**58.67**	**57.94**	**66.42**	**68.42**	**48.48**	**45.28**
2	GPT Image 2.5 Flare	55.71	56.53	57.14	63.85	65.26	43.29	37.01
3	GPT Image 2	53.24	56.53	53.17	62.94	62.81	38.41	37.01
4	Qwen Image 3 Pro	39.71	42.04	38.41	46.24	52.28	28.05	24.80
5	Seedream 5 Pro	33.67	33.49	35.56	39.45	43.16	21.04	19.69
6	Nano Banana 2	32.44	34.44	34.76	37.25	39.30	19.82	21.65
7	Nano Banana Pro	30.91	35.87	31.75	33.76	37.89	20.43	22.44
8	Grok Image 2	28.29	34.68	27.94	33.21	34.74	19.51	18.90
9	MAI Image 2.5	23.49	29.69	24.92	25.32	30.18	14.63	13.39
10	FLUX.2 Max	15.27	19.71	14.76	19.08	19.82	8.54	6.69

Sunburst also leads every task-weighted domain column. Flare places second, tying GPT Image 2 on Symbolic (56.53%) and Procedural (37.01%).

4.3 Domain-Level Analysis

Figure 3 presents two complementary views of LEVI capability. The task-weighted view measures overall task-solving capability under the evaluated task distribution: success requires all identity conditions in a task to hold. The B-VIS-balanced view highlights strengths and weaknesses across LEVI identity categories by giving each represented B-VIS equal weight within its domain, rather than allowing frequently sampled identity types to dominate. Both views retain the same whole-task success criterion.

Reweighting changes the Geometric leader: Flare reaches 57.57% under B-VIS balancing, compared with Sunburst's 55.58%. Sunburst retains the lead in the other five domains. Thus, the frequency of individual identity types affects the observed model ordering even within a domain.

The whole-task profiles in Figure 3(c) show a common contraction on Quantitative and Procedural task subsets. Sunburst reaches 68.42% on Layout and 66.42% on Topological, compared with 48.48% on Quantitative and 45.28% on Procedural. Every configuration scores lower on both of the latter subsets than on its Layout subset. These gaps describe performance on different task populations, rather than independent measurements of domain-specific reasoning abilities.

The middle of the leaderboard also contains rank reversals. Nano Banana 2 exceeds Nano Banana Pro overall by 1.53 points, yet Pro performs better on Symbolic (35.87% versus 34.44%) and Procedural (22.44% versus 21.65%). Domain membership identifies what a task involves, not which condition caused its failure. A topology-and-layout task contributes to both subsets even when only one of its requirements is violated.

4.4 Identity Spaces and Composition

The mean pass rates across ten models are 35.45%, 35.46%, and 38.31% for tasks involving one, two, and at least three distinct B-VIS types. Cross-domain tasks reach 38.24%, compared with 34.11% for single-domain tasks. These are different task populations, not matched component additions: more identity types do not imply lower success in this set. Table 3 reports the full structure breakdown. Within domains, Sunburst reaches 60.65% on tasks involving orientation but 43.48% on height order.

Table 3: Whole-task pass rates (%) by task structure. Share is the fraction of tasks in each group; Mean averages all ten models. The two groupings partition the same task set separately.

Task structure	Share (%)	Mean	Sunburst	Flare	GPT Image 2
One B-VIS type	10.40	35.45	62.24	52.45	49.65
Two B-VIS types	30.62	35.46	56.06	52.97	50.59
Three or more B-VIS types	58.98	38.31	59.43	57.71	55.24
Single-domain	26.69	34.11	55.31	50.95	50.41
Cross-domain	73.31	38.24	59.92	57.44	54.27

4.5 Visible Identity Errors

Figure 4 compares five models on four tasks, with a vector illustration of the target at the left of each row. The references make the identity requirements visible but are not supplied to the models. The cases illustrate correct realizations, local errors, and larger structural departures under the same prompt.

These illustrative cases have one or two passing outputs per row. In the two-tray task, Sunburst and GPT Image 2 preserve the requested counts and arrangements; the other three models produce twelve rings instead of ten. The blister task exposes local state errors: GPT Image 2 opens the bottom row, Qwen omits the middle-left seal, and Nano Banana Pro retains seals in the top row. FLUX.2 Max changes the row count.

On the keyboard, Qwen repeats small ellipses, Nano Banana Pro introduces nine key positions, and FLUX.2 Max extends paint across the exposed base. For the umbrella, Qwen produces ten panels; Nano Banana Pro and FLUX.2 Max change the specified color order. The reference encodes the required identity, not an appearance template; these selected cases do not estimate error frequencies.

Figure 4: Identity-level comparisons, one task per row. Vector references illustrate the text-only targets; outputs are uncropped. Pass/fail denotes whole-task assessment.

Vector reference	GPT Image 2.5 Sunburst	GPT Image 2	Qwen Image 3 Pro	Nano Banana Pro	FLUX.2 Max
(a) Two trays: 12 five-point stars in 3-by-4 rows; 10 rings in 2-by-5 rows.					
	Pass	Pass	Fail	Fail	Fail
(b) Blister pack: 3-by-7 cavities; sealed columns by row: none / 1,2,3,4 / all seven.					
	Pass	Fail	Fail	Fail	Fail
(c) Six key positions: lower-middle cap absent; one shared ellipse, no paint on the exposed base.					
	Pass	Pass	Fail	Fail	Fail
(d) Eight panels, clockwise from the top seam: blue, red, yellow, blue, yellow, red, red, yellow.					
	Pass	Pass	Fail	Fail	Fail

5 Discussion and Conclusion

LEVI separates identity requirements from permissible appearance variation. VIS defines the identities and their equivalence; VVG makes their visual requirements explicit. Across 1,375 curated tasks and ten model configurations, the highest whole-task pass rate is 58.69%. B-VIS balancing changes the Geometric leader, while component-type counts do not explain task difficulty. The examples expose concrete errors in counts, spatial assignments, and local component states despite recognizable objects. Identity-level evaluation makes these distinctions inspectable.

Limitations. Single outputs do not measure generation variability. Scores depend on post-hoc curation and automatic judgments with task-specific instruction revisions (Section 4.1). Domain and construction-batch imbalance remains after B-VIS reweighting. Whole-task rates cannot isolate component errors or establish intrinsic domain difficulty; those claims require matched comparisons and independent human validation.

AI USE STATEMENT

We used generative AI tools to generate synthetic data sets, including the images produced by all evaluated models. Two authors curated the tasks and the vector references with AI tools and manually reviewed for precision. Automatic assessment uses GPT-6-Sol, as described in Section 4.1. We did not use generative AI tools for any other task that requires disclosure. Additionally, we used LLM-based agents to suggest style polishing for readability. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

ByteDance Seed. Beyond generation, it understands design: Introducing Seedream 5.0 Pro, July 2026. URL https://seed.bytedance.com/en/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro.

Zhaorun Chen et al. MJ-Bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024. URL https://arxiv.org/abs/2407.04842.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Adrienne Deganutti, Elad Hirsch, Haonan Zhu, Jaejung Seol, and Purvanshi Mehta. Graphic-Design-Bench: A comprehensive benchmark for evaluating AI on graphic design tasks. *arXiv preprint arXiv:2604.04192*, 2026. URL https://arxiv.org/abs/2604.04192.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023. URL https://arxiv.org/abs/2310.11513.

Google. Nano Banana 2: Combining pro capabilities with lightning-fast speed, February 2026. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/.

Kevin David Hayes, Micah Goldblum, Vikash Sehwag, Gowthami Somepalli, Ashwinee Panda, and Tom Goldstein. FineGRAIN: Evaluating failure modes of text-to-image models with vision language model judges. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2512.02161.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://arxiv.org/abs/2104.08718.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.08500.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. URL https://arxiv.org/abs/2303.11897.

Hang Hua, Ziyun Zeng, Yizhi Song, Yunlong Tang, Liu He, Daniel G. Aliaga, Wei Xiong, and Jiebo Luo. MMIG-Bench: Towards comprehensive and explainable evaluation of multi-modal image generation models. *arXiv preprint arXiv:2505.19415*, 2025. URL https://arxiv.org/abs/2505.19415.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. URL https://arxiv.org/abs/2307.06350.

Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Weilong Dai, Hao Jiang, Fei Wu, and Leilei Gan. T2I-FactualBench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 27501–27524, 2025. URL https://aclanthology.org/2025.acl-long.1334/.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. URL https://arxiv.org/abs/2512.16853.

Mukund Khanna, Raj Singh Yadav, and Kunal Singh. ProductConsistency: Improving product identity preservation in instruction-based image editing via SFT and RL. *arXiv preprint arXiv:2606.19103*, 2026. URL https://arxiv.org/abs/2606.19103.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback. *arXiv preprint arXiv:2404.07987*, 2024. URL https://arxiv.org/abs/2404.07987.

Jia-Wei Liao, Winston Wang, Tzu-Sian Wang, Li-Xuan Peng, Ju-Hsuan Weng, Cheng-Fu Chou, and Jun-Cheng Chen. DiffQRCoder: Diffusion-based aesthetic QR code generation with scanning robustness guided iterative refinement. *arXiv preprint arXiv:2409.06355*, 2024. URL https://arxiv.org/abs/2409.06355.

Ruiyao Liu et al. MathGen: Revealing the illusion of mathematical competence through text-to-image generation. *arXiv preprint arXiv:2603.27959*, 2026a. URL https://arxiv.org/abs/2603.27959.

Yutong Liu, Nan Huang, Xu Cao, and James M. Rehg. Thinking in pictures: A systematic benchmark for reasoning-driven image generation. *arXiv preprint arXiv:2609.02864*, 2026b. URL https://arxiv.org/abs/2609.02864.

Yuxuan Luo, Yuhui Yuan, Jun-Wen Chen, Haonan Cai, Ziyi Yue, Yu-Wei Yang, Fatima Zohra Daha, Ji Li, and Zhouhui Lian. MMMG: A massive, multidisciplinary, multi-tier generation benchmark for text-to-image reasoning. *arXiv preprint arXiv:2506.10963*, 2025. URL https://arxiv.org/abs/2506.10963.

Kunato Nishina and Yusuke Matsui. SVGEditBench: A benchmark dataset for quantitative assessment of LLM's SVG editing capabilities. In *CVPR Workshop on Graphic Design Understanding and Generation*, 2024. URL https://arxiv.org/abs/2404.13710.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. DreamBench++: A human-aligned benchmark for personalized image generation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.16855.

Aditi Singhania et al. Beyond the pixels: VLM-based evaluation of identity preservation in reference-guided synthesis. *arXiv preprint arXiv:2511.08087*, 2025. URL https://arxiv.org/abs/2511.08087.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, Aoxue Li, and Xihui Liu. T2I-ReasonBench: Benchmarking reasoning-informed text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8919–8944, 2026. URL https://aclantholo gy.org/2026.findings-acl.433/.

Haoran Wang, Chaofan Ma, Ran Yi, and Lizhuang Ma. TRACE-Bench: Decomposing and diagnosing multi-reference image generation. *arXiv preprint arXiv:2608.16765*, 2026a. URL https://arxiv.org/abs/2608.16765.

Maijunxian Wang et al. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*, 2026b. URL https://arxiv.org/abs/2602.20159.

Xu Wang, Kaixiang Yao, Miao Pan, Xiaohe Zhou, Xuanyu Liu, Wenqi Zhang, and Xuhong Zhang. Show, don't tell: Evaluating spatial cognition in generative pixels rather than LLM text. *arXiv preprint arXiv:2607.21072*, 2026c. URL https://arxiv.org/abs/2607.21072.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Manyuan Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*, 2026d. URL https://arxiv.org/abs/2601.20354.

Tianyu Zhang et al. STRICT: Stress test of rendering images containing text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://arxiv.org/abs/2505.18985.

Changpeng Zhao, Yiren Song, and Jinpeng Wang. GenPuzzle: Benchmarking visual reasoning in image generation models. *arXiv preprint arXiv:2609.05902*, 2026. URL https://arxiv.org/abs/2609.05902.

Peng Zheng, Ruiqi Liu, Rui Ma, and Zuxuan Wu. CopyCat: Improving fine-grained subject consistency in subject-to-image models within seconds. *arXiv preprint arXiv:2608.00674*, 2026. URL https://arxiv.org/abs/2608.00674.

Sashuai Zhou, Qiang Zhou, Junpeng Ma, Yue Cao, Ruofan Hu, Ziang Zhang, Xiaoda Yang, Zhibin Wang, Jun Song, Cheng Yu, Bo Zheng, and Zhou Zhao. SpatialReward: Verifiable spatial reward modeling for fine-grained spatial consistency in text-to-image generation. *arXiv preprint arXiv:2603.22228*, 2026. URL https://arxiv.org/abs/2603.22228.

Le Zhuo et al. Factuality matters: When image generation and editing meet structured visuals. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.05091.

Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. VGBench: Evaluating large language models on vector graphics understanding and generation. *arXiv preprint arXiv:2407.10972*, 2024. URL https://arxiv.org/abs/2407.10972.

A Identity Semantics and Task Construction

A.1 Admissible Identity Values

An identity specification s states which visible facts determine identity and which variations preserve it. Let $\mathcal{X}_s^{\mathrm{adm}}$ be the set of images expressing admissible identities under these rules. The equivalence relation $x \sim_s x'$ means that the two images express the same identity. The resulting space is

$$\mathcal{Z}_s = \mathcal{X}_s^{\mathrm{adm}} / \sim_s. \tag{7}$$

Here $\mathcal{Z}_s$ is the VIS, and a point $z \in \mathcal{Z}_s$ is one identity class, not one image or physical object. For example, painted and engraved versions of the same ordered symbol sequence can belong to one class even though their pixels differ.

An identity class can be represented by deciding coordinates when those coordinates distinguish the classes without ambiguity. For a fixed-length representation,

$$\mathcal{Z}_s \cong \left\{ (z_1, \ldots, z_k) \in \prod_{j=1}^{k} \mathcal{D}_j : \chi_s(z_1, \ldots, z_k) = 1 \right\}. \tag{8}$$

The symbol $\cong$ denotes a one-to-one representation of the classes; k is the number of coordinates, z_j is coordinate j, $\mathcal{D}_j$ is its value domain, and χ_s is the legality test. For a sequence, the coordinates may record symbol values and their order. Variable-length structures require an appropriate length-dependent encoding.

Legality restricts which coordinate combinations represent identities. It does not make the coordinates independent. Multiple descriptions of the same class must be identified before counting distinct identities. Likewise, two different identities are mutually exclusive outcomes of the same semantic projection; they need not be orthogonal vectors or equally spaced points. The six domains organize identity facts but do not impose a common distance on all VIS.

A.2 Composition, Operations, and Semantic Granularity

Consider a VIS of finite, nonempty sequences over $\Sigma = \{\triangle, \circ, \square\}$, where Σ is the symbol alphabet. The identity $z = (\triangle, \circ, \square)$ specifies values and order. This example separates four operations that have different meanings.

Composition. A sequence identity can be combined with a count identity for the same symbol occurrences. If n denotes that count and $|z|$ the sequence length, a compatible pair must satisfy $n = |z|$. The pair $(z, 3)$ is admissible, whereas $(z, 4)$ is not. Composition retains component identities and their bindings to the same objects. If relative position, attachment, or another relation changes the joint identity, that relation must also be represented. A tuple of component identities alone is sufficient only when it determines all such facts.

Projection. The map $P(z) = |z|$ retains count while discarding symbol values and order. Here P maps identities in the sequence VIS to identities in the count VIS. Both $(\triangle, \circ, \square)$ and $(\triangle, \square, \circ)$ map to 3. This semantic projection between spaces differs from the image-to-identity map ϕ_s used in the main text.

Transformation. Reversal maps $(\triangle, \circ, \square)$ to $(\square, \circ, \triangle)$. A transformation specifies its source and target spaces and any preconditions. Changing paint to engraving without changing the sequence changes the realization rather than the identity.

Refinement. On a common image domain, an ordered-symbol specification distinguishes more classes than a count-only specification. Forgetting order and symbol values merges classes; adding those distinctions separates them. Thus, the same image can support different identities under different specifications. Comparisons across spaces require their semantics to be stated explicitly. Their sharing an image does not by itself define a shared coordinate system.

A.3 Image Realizations and Identity Fidelity

The deterministic semantic map ϕ_s assigns an image its identity. To cover all generated images, extend the map by a value $\perp$ for images outside the admissible domain:

$$\phi_s : \mathcal{X} \to \overline{\mathcal{Z}}_s, \qquad \overline{\mathcal{Z}}_s = \mathcal{Z}_s \cup \{\perp\}, \qquad \phi_s(x) = \begin{cases} [x]_{\sim_s}, & x \in \mathcal{X}_s^{\mathrm{adm}}, \\ \perp, & x \notin \mathcal{X}_s^{\mathrm{adm}}. \end{cases} \tag{9}$$

$\mathcal{X}$ is the image domain, $[x]_{\sim_s}$ is the identity class of x, and $\perp$ is a semantic value, not an automatic judge's confidence or an operational error code.

Let $q_\theta(x \mid c)$ be the generator's probability of image x under text condition c, with model parameters θ. For identity z, define its realization set and induced probability by

$$\mathcal{F}_s(z) = \{x \in \mathcal{X} : \phi_s(x) = z\}, \qquad q_{\theta,s}(z \mid c) = \sum_{x \in \mathcal{F}_s(z)} q_\theta(x \mid c). \tag{10}$$

The sum assumes discrete rendered images. $\mathcal{F}_s(z)$ includes all appearances expressing z under the chosen rules. Rope and wire can realize the same path identity when material does not determine that identity; a material-specific function would require a different specification.

The prompt can permit one identity or several. Let $A_s(c) \subseteq \mathcal{Z}_s$ contain all identities satisfying its requirements. Identity fidelity is

$$\mathrm{IF}_\theta(c, s) = \sum_{z \in A_s(c)} q_{\theta,s}(z \mid c). \tag{11}$$

For an exact request, $A_s(c) = \{z^\star\}$, where $z^\star$ is the target. If the prompt asks for one of each symbol in any order, all six permutations are credited. A chosen vector reference illustrates one permitted realization; it does not remove other valid answers.

This probability is a population quantity. The empirical whole-task pass rate estimates success under the task set and assessment procedure, rather than directly observing ϕ_s or the generator's full distribution. The current automatic assessment is specified in Section 4.1.

A.4 Entropy, Dependence, and Correctness

Let X be the generated image and $Z_s = \phi_s(X)$ its expressed identity. For a fixed text condition $C = c$, use the same distribution $q_\theta(\cdot \mid c)$ for all entropy terms. Since ϕ_s is deterministic, $H(Z_s \mid X, C = c) = 0$. Applying the entropy chain rule in two orders gives

$$\begin{aligned} H(X, Z_s \mid C = c) &= H(X \mid C = c) + H(Z_s \mid X, C = c) \\ &= H(X \mid C = c), \\ H(X, Z_s \mid C = c) &= H(Z_s \mid C = c) + H(X \mid Z_s, C = c). \end{aligned} \tag{12}$$

Here H is Shannon entropy in bits, using base-two logarithms. The first term in the resulting decomposition measures uncertainty about identity; the second measures appearance variation within an identity. Continuous coordinates must be quantized at a declared resolution for this discrete argument.

When an identity has coordinates $Z_1, \ldots, Z_k$, their uncertainty decomposes as

$$H(Z_1, \ldots, Z_k \mid C = c) = \sum_{j=1}^{k} H(Z_j \mid Z_{<j}, C = c). \tag{13}$$

$Z_{<j} = (Z_1, \ldots, Z_{j-1})$ contains the preceding coordinates and is empty for $j = 1$. Omitting this conditioning requires an independence assumption. A count determined by a complete object list contributes no additional conditional uncertainty. Decomposability therefore does not imply independent information or independent controllability.

Entropy measures concentration, not correctness. A generator that always produces one wrong identity has zero identity entropy and zero fidelity for an exact target. A generator supported entirely

on several permitted identities can have positive entropy and perfect fidelity. Low desired identity entropy characterizes an exact specification; success still requires agreement with that specification.

One can express fidelity as a population penalty $\mathcal{L}_{\mathrm{id}}(\theta; c, s) = -\ln \mathrm{IF}_\theta(c, s)$, where $\ln$ is the natural logarithm and zero fidelity gives infinite penalty. This notation describes the desired probability mass. It is not a training loss evaluated in the reported experiments, and no improvement from optimizing it is claimed.

A.5 VVG Descriptions and Text-to-Image Tasks

A VIS defines meaning; VVG defines structural descriptions with that meaning. For a grammar Γ_s, let $\mathcal{D}_s$ contain complete finite descriptions and let Φ_s be its constraints on types, attributes, and relations. The admissible language and its interpretation are

$$\mathcal{L}(\Gamma_s) = \{g \in \mathcal{D}_s : g \models \Phi_s\}, \qquad \psi_s : \mathcal{L}(\Gamma_s) \to \mathcal{Z}_s. \tag{14}$$

g is a structural description, $g \models \Phi_s$ means that it satisfies the constraints, and $\psi_s(g)$ is its identity. A template with unassigned values is not a complete description. Distinct descriptions can express the same identity; their equivalence is determined by their interpretation.

In a sequence grammar, each occurrence has a symbol value and successor relations define the order. Constraints require symbols from the alphabet and a single complete chain. Interpretation reads values in chain order. For prompt c, the task-compatible descriptions are

$$\mathcal{L}(\Gamma_s; c) = \{g \in \mathcal{L}(\Gamma_s) : \psi_s(g) \in A_s(c)\}. \tag{15}$$

Both $(\triangle, \circ, \square)$ and $(\triangle, \square, \circ)$ are legal sequences, but only the first satisfies an exact request for that order. Grammar membership and task success are different tests.

Instantiation assigns values; composition binds components and relations; transformations change a description while respecting the intended identity operation. If $T : \mathcal{Z}_s \to \mathcal{Z}_{s'}$ is an identity transformation and $\widetilde{T}$ its structural implementation, consistency requires

$$\psi_{s'}(\widetilde{T}(g)) = T(\psi_s(g)) \tag{16}$$

where both sides are defined. Here s' denotes the target specification, and $\widetilde{T}(g)$ must belong to $\mathcal{L}(\Gamma_{s'})$. Constructors implement these requirements over a stated range; producing valid examples does not establish exhaustive coverage of a VIS.

A visual medium m gives a description concrete form. Let η collect other permitted presentation choices, such as lighting and viewpoint, and let R denote an authoring realization process. Identity preservation requires

$$x = R(g, m, \eta), \qquad \phi_s(x) = \psi_s(g). \tag{17}$$

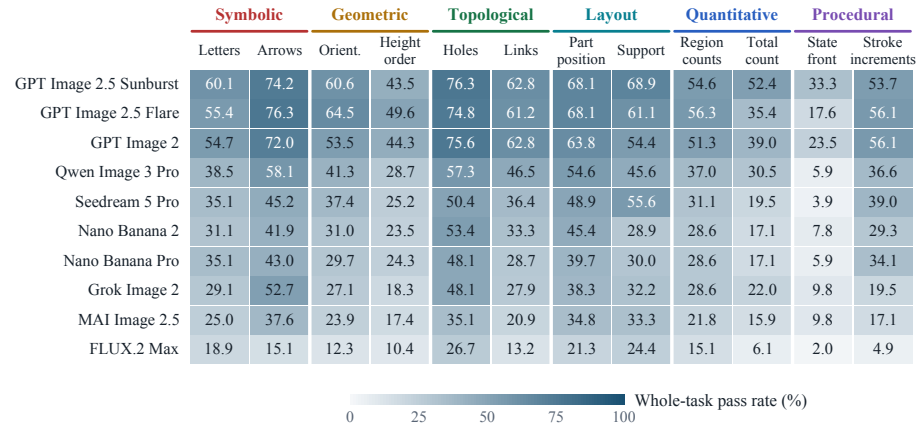
The equality is a condition on a faithful realization, not a guarantee for every renderer. A viewpoint or material choice is appearance variation only when it preserves the deciding facts and their object associations.

For the benchmark, the generator receives the complete text condition c, including the required identity facts and their concrete presentation. It does not receive g or the vector reference. Assessment concerns the generated pixels: each requirement must have visible evidence and an explicit acceptance rule. Extra objects, incorrect relations, and malformed structures remain errors to inspect, rather than observations to repair using the intended description.

B Extended Experimental Results

All results here use the same 1,375 retained tasks and full denominators as Table 2. Each heatmap column groups tasks involving a B-VIS; the two most frequent types per domain are selected independently of model performance. Rates assess complete tasks, not isolated predicates.

Figure 5: Whole-task pass rates (%) for the two most frequent B-VIS in each domain. Columns group tasks by the identities they involve; they do not measure isolated component accuracy.



For Sunburst, orientation and height-order groups reach 60.65% and 43.48%, respectively. Flare leads Sunburst on partition counts (56.30% versus 54.62%) but trails on total count (35.37% versus 52.44%). These differences show why domain averages alone do not describe every identity type.