"Draw Me the Moon that Looks Like a Boat": EVALUATING VISUAL SIMILE CANONICITY IN TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a text-to-image (T2I) model is asked for "a curved moon", it will draw a crescent with ease. However, when it is prompted with a simile such as "a moon that looks like a boat", it often draws a plain boat, a moon beside a boat, a full moon, or a moon with apparent masts and sails. Such difficulty is because *"looks like"* 's implication shifts with every pair of things being compared to in the prompt, e.g, a moon that resembles a boat takes a boat's curve, while a cherry that resembles a ruby takes a ruby's color. To bridge this gap so that T2I models can better handle such prompts with implicated resemblance, we thus formalize the **Visual Simile Generation (VSG)** task: given a simile in form of "$\langle A \rangle$ that look(s) like $\langle B \rangle$", a T2I model must: 1., install $\langle A \rangle$ as the dominant subject, 2., identify and transfer the implicated attribute from the comparison object $\langle B \rangle$ onto the subject $\langle A \rangle$, and 3., render no concrete features of $\langle B \rangle$ literally. We curate **VSG-Bench**, consisting of 549 simile prompts spanning 13 resemblance relations over four axes. At evaluation, each prompt licenses the one attribute of $\langle B \rangle$ a generation is expected to transfer, and names the concrete features of $\langle B \rangle$ it must leave out. We also establish the composite **CANONSCORE** metric, which scores a generation on each of the three requirements separately: $\text{CS}_{\text{anchor}}$ for the subject, CS_{resem} for the transferred attribute, and CS_{bound} for staying clear of $\langle B \rangle$'s concrete features. We benchmark 16 recent T2I systems spanning various architectures. We find no current T2I system can satisfy all three requirements simultaneously. More signifcantly, corroborated by human raters, our full paradigm effectively exposes **the resemblance-literalization tradeoff**, that the T2I systems conveying the intended resemblance attribute best are also the ones literalizing the most of the comparison objects.

1 INTRODUCTION

Anyone asked to draw *"a moon that looks like a boat"* should picture a crescent moon curving upwards, like the way a boat's hull does. Nobody would expect to see a mast, planking, or a boat parked beside the moon. The phrase never says which part of the boat should be drawn, yet the answer is immediate, because we know what moons and boats are in real life. People communicate with comparisons like this all the time, because likening one thing to another conveys a vivid idea in fewer words than naming the attributes outright (Colston & Gibbs, 2021; Lai & Nissim, 2024).

However, current Text-to-image (T2I) models are less likely to arrive at that answer. Systems such as GPT Image 2 (OpenAI, 2026) and Nano Banana 2 (Google, 2026) may be able to render high-resolution compositions and follow prompts that state their attributes if they are given in fine details. But when asked for a moon that is likened a boat, the T2I models would draw a boat, a moon and a boat, or an ordinary moon instead. What makes such a ***visual simile*** a difficult prompt is that *"looks like"* does not mean one thing. Its implication shifts with every pair it joins, and depends entirely on what those two things are: a moon takes a boat's contour, while a cherry takes a ruby's color or gloss. Hence, a capable T2I model must recover that implicated attribute of resemblance before it can draw anything, then carry the attribute across while leaving the rest of the compared object behind. We observe that current T2I systems fail in three recurring ways, shown in Figure 1, using *"a moon that looks like a boat"* as a running example:

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.

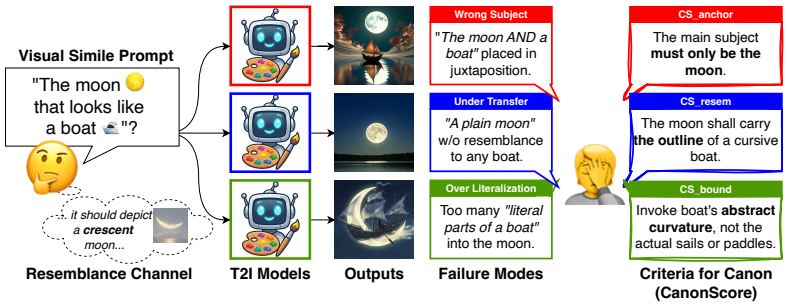


Figure 1: Left: We find many existing text-to-image models still struggle at visualizing similes such as "the moon that looks like a boat", failing in three recurring modes: *Wrong Subject*, *Under Transfer*, and *Over Literalization*. **Right:** We propose **CANONSCORE**, an evaluation metric that scores each of the three requirements separately (CS_{anchor}, CS_{resem}, and CS_{bound}), reporting which requirement a generation met rather than one combined number.

Wrong Subject. The image stops being a depiction of the moon: the moon and a boat appear as two separate objects sharing a scene, or the boat takes over as the main entity, leaving any resemblance nothing to attach to.

Under Transfer. The moon is held as the subject, but the comparison leaves no visual consequence. The image is just an ordinary full moon, where the boat resemblance leaves no trace.

Over Literalization. Too much of the boat arrives on top of the moon. Boat-exclusive components like the mast, timber planking, or the sails are visible. The image is no longer the single upward curve moon the comparison invokes.

Existing evaluation cannot register these three failures separately. Embedding-based text-image alignment scores such as CLIPScore (Hessel et al., 2021), object-presence frameworks like GenEval (Ghosh et al., 2023) and Dall-Eval (Cho et al., 2023), and preference models like PickScore (Kirstain et al., 2023) all return a single number in which a lost subject, an absent resemblance and an over-literal one are indistinguishable. This is the same gap that defeats the models themselves: so long as the implication of "*looks like*" is left unstated, there is nothing a generation can be held to and nothing a metric can score against.

In light of all of these findings, we formalize the task of **Visual Simile Generation (VSG)** and build an evaluation that reports which of its three requirements an image met. Our key contributions are three-fold:

The Visual Simile Generation task. We formalize VSG and give the three requirements a generation must satisfy, each paired with the way generation fails it: *Wrong Subject*, *Under Transfer*, and *Over Literalization*. Attaching one transferable attribute to every prompt turns "*looks like*" from an unfalsifiable instruction into a testable one.

VSG-Bench. We release a diagnostic benchmark of 549 prompts over 13 resemblance relations in four axes. Every prompt annotates the one attribute of the compared object that may transfer and the concrete features an abstraction must discard, which is what lets a single image be scored against each requirement on its own.

CANONSCORE. We design a composite evaluation metric that scores each of the three requirements separately and reports them side by side. Benchmarking 16 T2I systems finds every system missing all three requirements somewhere in its output, and a human survey finds that CANONSCORE reproduces the direction human raters report.

2 RELATED WORKS

Multimodal reasoning and figurative generation. Recent vision-language models have advanced both grounding and multi-step reasoning, and a growing line of work studies figurative language in multimodal settings (Bai et al., 2023; Kye et al., 2025; Shahmohammadi et al., 2023). Closest to

our setting is CaRVE (Kundu et al., 2026), which targets visual resemblance generation directly and scores it with a single composite. These systems improve conceptual alignment and iterative critique, but they are generally designed for general reasoning or metaphor understanding rather than the strict visual simile setting in which a target subject must remain dominant while a single attribute from a comparison object is transferred without literal instantiation. Benchmarks such as IRFL, V-Flute, and MultiMM (Yosef et al., 2023; Saakyan et al., 2025; Yang et al., 2025) capture broader figurative understanding, yet they do not isolate the structural failure modes that define visual simile generation.

Foundational text-to-image models and literalism. Modern text-to-image systems achieve strong literal fidelity and prompt adherence across architecture families: latent diffusion (Rombach et al., 2022), unified autoregressive generation (Seedream et al., 2025; Wu et al., 2025a), and reasoning-augmented generation (Kou et al., 2026). However, these models still struggle when the prompt requires attribute transfer rather than object depiction: Wang et al. (2026) find that frontier systems systematically misread the comparison object of a simile and render it as an object in its own right. This remains a central weakness of current T2I architectures: compositional binding and attribute transfer are not yet robust under figurative constraints.

Evaluation metrics for structured figurative alignment. Standard metrics such as CLIPScore, DALL-Eval, PickScore, and GenEval (Hessel et al., 2021; Cho et al., 2023; Kirstain et al., 2023; Ghosh et al., 2023) focus on global alignment or object presence, which is insufficient for visual similes because they reward literal recognizability rather than the intended resemblance relation. VQA-based approaches such as TIFA and VQAScore (Hu et al., 2023b; Lin et al., 2024) offer finer-grained assessment, but they still target whether a concept is present instead of whether a subject remains anchored while only a licensed attribute is transferred. This gap motivates **VSG-Bench** and CANONSCORE: a structure-aware benchmark and metric that decompose success into subject preservation, resemblance evocation, and abstraction from the comparison object. Table 4 places the closest prior benchmarks against those same criteria.

3 VISUAL SIMILE GENERATION TASK

3.1 TASK FORMULATION

We introduce the T2I task of **visual simile generation (VSG)**. A **simile** prompt is a nominal phrase that implicates a visual resemblance between two distinct entities $\langle A \rangle$ and $\langle B \rangle$, rendered in the fixed template: "$\langle A \rangle$ that look(s) like $\langle B \rangle$", where, grammatically, $\langle A \rangle$ is the subject and $\langle B \rangle$ is the comparison object of the phrase.

We apply two additional rules to regulate the entity pairs that VSG incorporates. First, $\langle A \rangle$ and $\langle B \rangle$ are restricted to **everyday, tangible entities**, whose visual attributes are concrete and common-sense: a moon, a boat, an umbrella, a parachute. This sets VSG apart from figurative benchmarks built on abstract or emotional concepts, such as "a heart of gold", where the comparison turns on connotation and no visual attribute is the one at issue. Second, the resemblance between $\langle A \rangle$ and $\langle B \rangle$ is one that **occurs naturally**: a crescent moon and a small boat sharing an upward-curving profile can be easily seen in a clear evening. VSG prompts therefore only include resemblances that any observer would recognize on sight, rather than inventing a likeness that holds only by stipulation. Together, these two rules keep the VSG task grounded to common sense and thus more **diagnostic**.

3.2 SUCCESS CRITERIA

A **canonical** generation for "$\langle A \rangle$ that look(s) like $\langle B \rangle$" depicts a singular entity, $\langle A \rangle$, transferring one borrowed characteristic of $\langle B \rangle$. That phrasing carries three independent requirements: which entity the image is of, if any characteristic is borrowed from $\langle B \rangle$ at all, and if that borrowed characteristic has been fully abstracted rather than being a literal $\langle B \rangle$ itself. The three criteria below state the requirements as separately checkable items against an image.[1]

[1]This follows the framing of simile as attribute transfer from a source domain to a target domain: in CMT, $\langle A \rangle$ is the target and $\langle B \rangle$ is the source. VSG operationalizes it narrowly, as a single named resemblance channel rather than an unrestricted metaphorical blend.

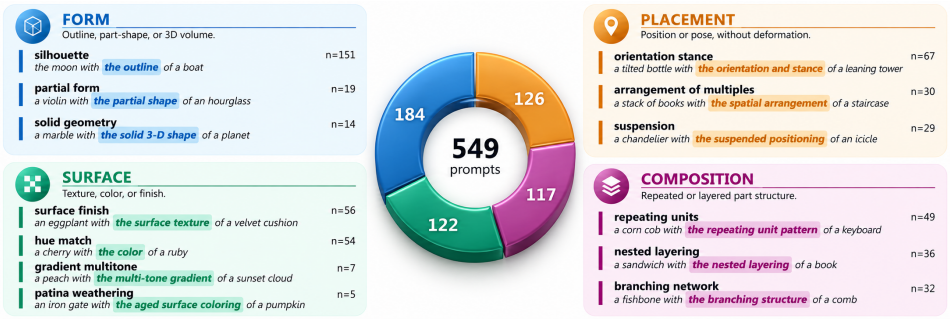


Figure 2: The VSG-Bench taxonomy and its composition. Four axes, refined into 13 relations, with one example and a count for each over the 549-prompt set. Each example paraphrases the fixed template *"an* $\langle A \rangle$ *that looks like a* $\langle B \rangle$*"*, and the **highlighted span** names the single attribute of $\langle B \rangle$ a correct image may transfer to $\langle A \rangle$. The span is a relation-level gloss of per-item text. Frequencies follow how readily each resemblance is expressed in natural similes rather than a uniform quota, so `silhouette` accounts for 27.5% of the benchmark while four relations hold fewer than n=20 items.

Subject preservation. $\langle A \rangle$ remains the anchored, dominant and recognizable visual subject, so that the image is a depiction *of* $\langle A \rangle$. This criterion is prior to the other two, which both concern what $\langle A \rangle$ is depicted as carrying, and therefore presuppose that it is $\langle A \rangle$ on the canvas.

Resemblance evocation. The image visibly conveys the **licensed channel** of $\langle B \rangle$ on $\langle A \rangle$: the single visual attribute, such as a shape, contour, placement or texture, that this prompt permits to transfer. Each benchmark entry names that channel, so this criterion is checkable against one stated attribute rather than against $\langle B \rangle$ at large, which is what makes "looks like" decidable here.

Abstraction from the comparison object. The licensed channel arrives as an abstracted characteristic rather than as an instance of $\langle B \rangle$. The **literal particulars** of $\langle B \rangle$, its concrete features that a successful abstraction discards, stay absent from the image. Each entry annotates this set, so the criterion is scored against particulars fixed in advance for that prompt.

Because each criterion is checked against an annotation the benchmark fixes in advance, the three together make canonicity decidable per image, and decidable one criterion at a time.

3.3 Failure Modes

In addition to the success criteria, we also identify three common failure modes (i.e., error types) for the task of VSG, each of which negates one of the three conditions in a distinct way: the subject can be lost, the resemblance can fail to arrive, or the object can arrive too literally.

Wrong Subject occurs when the generated image fails to establish $\langle \mathrm{A} \rangle$ as the anchored, dominant subject, whether because $\langle \mathrm{A} \rangle$ and $\langle \mathrm{B} \rangle$ are rendered as two separate entities sharing a scene, or because the image drifts towards $\langle \mathrm{B} \rangle$ as the main entity. *"An image of* $\langle A \rangle$ *and* $\langle B \rangle$ *side by side,"* or *"an image of* $\langle B \rangle$ *that looks like* $\langle A \rangle$*."*

Under Transfer occurs when $\langle \mathrm{A} \rangle$ is correctly anchored, but the resemblance to $\langle \mathrm{B} \rangle$ does not visibly arrive, leaving an unremarkable, canonical depiction of $\langle \mathrm{A} \rangle$ on which the simile has left no trace. *"An image of a plain* $\langle A \rangle$*, and nothing more."*

Over Literalization occurs when the resemblance does arrive, but as a concrete instance of $\langle \mathrm{B} \rangle$ rather than as the abstracted characteristic the simile invokes. For *"the moon that looks like a boat"*, the simile invokes a boat's upward-curving hull profile; its mast, its rigging, its timber planking and a sea beneath it are the literal particulars this prompt asks a generation to discard. *"An image of* $\langle A \rangle$ *that is overtaken by* $\langle B \rangle$*."*

We use this task formulation to guide the construction of **VSG-Bench**, a new task benchmark designed to elicit these failure modes and support consistent evaluation.

Table 1: A complete VSG-BENCH entry. The axis and relation place it in the taxonomy of Figure 2. The licensed channel is the resemblance attribute of $\langle B \rangle$ that a canonical generated image may transfer, shown here in the templated form the entry stores. $\mathcal{L}$ is the set of literal particulars, the features of $\langle B \rangle$ that a successful abstraction must discard, each listed as the probe that detects it.

Base prompt: *"the moon that looks like a boat"* $\langle A \rangle$: the moon $\langle B \rangle$: a boat		
Taxonomy	**Licensed channel**	**Literal particulars $\mathcal{L}$ of $\langle B \rangle$**
Axis: FORM **Relation:** `silhouette`	$\langle A \rangle$ *with the outline of* $\langle B \rangle$	*"a tall mast and a rigged sail"* *"wooden planks and visible grain"* *"sitting on rippling open water"*

Table 2: Fixed VQAScore inquiries used to compute the CANONSCORE sub-scores, grouped by leg, for prompts of the form "⟨A⟩ that looks like ⟨B⟩". Every sub-score falls in $[0, 1]$; a sub-score marked $\downarrow$ is a failure probe and enters its leg inverted $(1 - s)$.

Leg	**Sub-score s**	**VQAScore Probe Question**
CS_{anchor}	s_{prompt}	Does this figure show ⟨A⟩ that looks like ⟨B⟩?
	$s_{\text{subj-main}}$	Does this figure show an image of ⟨A⟩ as the main subject?
	$s_{\text{juxt}} \downarrow$	Does this figure show ⟨A⟩ and ⟨B⟩ as two separate objects?
CS_{resem}	s_{channel}	Does this figure show ⟨A⟩ with ⟨licensed channel⟩ of ⟨B⟩?
	s_{relation}	Does this figure show ⟨A⟩ with ⟨B⟩'s ⟨relation⟩?
CS_{bound}	$s_\ell, \ell \in \mathcal{L} \downarrow$	Does this figure show ⟨A⟩ with ⟨literal particular ℓ⟩ of ⟨B⟩?

3.4 VSG-BENCH, THE TASK BENCHMARK

We construct **VSG-Bench**, a diagnostic benchmark for visual simile generation consisting of 549 vetted prompt candidates following the fixed template "$\langle A \rangle$ that looks like $\langle B \rangle$." Rather than treating all resemblance prompts as interchangeable, VSG-Bench organizes them into a two-level taxonomy comprising four resemblance axes (Form, Surface, Placement, and Composition) and thirteen fine-grained relations, shown with an example prompt each in Figure 2. Each relation specifies the visual attribute that may legitimately transfer from $\langle B \rangle$ onto $\langle A \rangle$, while excluding unrelated visual properties.

Per-entry annotation. Beyond the prompt text itself, every benchmark entry includes three annotation fields used throughout the evaluation pipeline: (i) a *licensed channel* specifying the single visual attribute permitted to transfer, (ii) three *literal particulars* naming characteristics of $\langle B \rangle$ that should not appear in a successful abstraction, and (iii) a templated paraphrase of the intended resemblance. These annotations define the evaluation criterion independently of any particular model output and eliminate the need for reference images as the primary evaluation target. See Table 1 for a complete example of a benchmark entry.

Dataset construction. The 549 prompts come from three sources: 40 authored and annotated by hand, 297 generated by a language model prompted one relation at a time, and 212 drawn from MultiMM (Yang et al., 2025) and relabelled under our taxonomy. Every candidate outside the authored seed set was screened twice in parallel, by a human reviewer and by a model from a different lineage than the generator, against five stated criteria: both nouns concrete and depictable, the two from different categories, the claimed licensed channel genuinely transferable, the resemblance non-trivial, and the direction unambiguous. A candidate entered the benchmark only if both passes marked it usable and not a duplicate of an existing pair in either order. See Appendix A for the full curation procedure.

4 CANONSCORE: 3-LEGGED EVALUATION METRIC FOR VISUAL SIMILE CANONICITY

The three failure modes of VSG in Section 3.3 are not distinguishable by a single aggregate text–image alignment score, which can report that something is wrong in general in the image, but

cannot pin-point which and to what extent the three failure modes are present. Targeting the VSG task (Section 3.1) as VSG-BENCH embodies it, we hence establish the 3-legged CANONSCORE to quantify the assessment of each failure mode separately. It is computed from a battery of VQAScore probes (Lin et al., 2024): given an image I and a probe text t, VQAScore converts t into a yes–no question $q(t)$ of the form "*Does this figure show {t}? Please answer yes or no.*" and takes the alignment score to be the probability of the answer "Yes" under a frozen judge model M,

$$s(I,t) \;=\; P_M(\text{“Yes”} \mid I,\, q(t)) \;\in\; [0,1]. \tag{1}$$

All CANONSCORE legs are built from the respective combination of probes (i.e. sub-scores) listed in Table 2 and are oriented so that **higher is better**. We motivate each leg's formulation in turn below.

The three CANONSCORE legs are deliberately *not* combined into a single scalar. A T2I generation can anchor the subject and still fail to transfer the resemblance, or transfer the resemblance and fail to abstract the literal particulars. We use CANONSCORE as a *profile* of a T2I model's behavior in synthesizing visual similes, rather than as a position in an ordering. Section 5.2 shows that on real T2I systems the legs are opposed strongly enough that any weighted average would cancel most of the signal it was meant to summarize.

Leg 1, CS_{anchor}: Is $\langle A\rangle$ still the subject? The failure of Wrong Subject may arrive in two surface forms: a juxtaposed image renders $\langle A\rangle$ faithfully while also setting apparent $\langle B\rangle$ beside it, or an image that drifts to $\langle B\rangle$ as the main entity is the converse. No one probe separates both from a canonical generation, and summing or averaging would let a high score on one condition offset a failure on another, so the leg takes a product:

$$CS_{\text{anchor}} \;=\; s_{\text{prompt}} \;\cdot\; s_{\text{subj-main}} \;\cdot\; \big(1 - s_{\text{juxt}}\big), \tag{2}$$

over the three respective sub-scores in Table 2, the last of which detects juxtaposition as a failure and enters it inverted. This product is a soft conjunction: a low sub-score on any factor suppresses the leg, since an image failing any one condition is not anchored on $\langle A\rangle$.

Leg 2, CS_{resem}: Does $\langle B\rangle$'s licensed channel of resemblance transfer? Resemblance can arrive along the wrong axis: a moon given a boat's color has taken on something of the boat while missing what the relation licenses, and scoring against $\langle B\rangle$ at large would credit it. The leg is therefore conditioned on the annotated licensed channel:

$$CS_{\text{resem}} \;=\; \tfrac{1}{2}\big(s_{\text{channel}} + s_{\text{relation}}\big), \tag{3}$$

The two sub-scores in Table 2 fail in opposite ways: the templated one states the channel concretely and inherits whatever its wording implies, while the relation one names it abstractly and depends on the judge construing the term as annotated. Averaging makes this leg robust to either failing alone, and neither is redundant with the other.

Leg 3, CS_{bound}: Is $\langle B\rangle$ abstracted or instantiated? A canonical generation should be a successful abstraction that discards every literal particular of $\langle B\rangle$, so scoring high on any one is discouraged: an $\langle A\rangle$ carrying a single unmistakable part of $\langle B\rangle$ reads as $\langle B\rangle$ imported, however cleanly the rest are left out. This leg therefore takes the worst case:

$$CS_{\text{bound}} \;=\; 1 \;-\; \max_{\ell \in \mathcal{L}} s_\ell, \tag{4}$$

over the sub-scores in Table 2 applied to each particular ℓ in the annotated set $\mathcal{L}$, so that a single strongly imported literal particular determines the score and $CS_{\text{bound}} = 1$ indicates abstraction without instantiation.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

T2I systems evaluated. We evaluate sixteen recent text-to-image (T2I) systems spanning four model families: proprietary systems (GPT Image 2 (OpenAI, 2026) and Gemini Flash 3.1 (Google,

Table 3: CANONSCORE baselines on the base VSG-BENCH prompts, grouped by model category. Two undecomposed scores are given for comparison, both computed on the same images: VQAScore, scored against the same base prompt, and CaRVE (Kundu et al., 2026). The last two rows give, for each undecomposed score, the Spearman correlation between its ranking of the sixteen T2I systems and the ranking each leg induces. Both order the systems alike on subject anchoring and resemblance and **oppositely** on abstraction of $\langle B \rangle$'s literal particulars. Table 9 extends the comparison to three further scores.

Category	Model	CaRVE	VQAScore	CS_{anchor}	CS_{resem}	CS_{bound}
Proprietary	GPT Image 2	0.649	0.922	0.670 $\pm$ 0.249	0.842 $\pm$ 0.182	0.316 $\pm$ 0.321
	Gemini Flash 3.1	0.654	0.898	0.587 $\pm$ 0.283	0.813 $\pm$ 0.213	0.290 $\pm$ 0.307
CoT	Think-Then-Generate	0.643	0.879	0.548 $\pm$ 0.304	0.794 $\pm$ 0.217	0.384 $\pm$ 0.315
	Bagel-Zebra-CoT	0.638	0.796	0.467 $\pm$ 0.320	0.651 $\pm$ 0.273	0.493 $\pm$ 0.323
Unified	Qwen-Image	0.638	0.875	0.523 $\pm$ 0.308	0.780 $\pm$ 0.233	0.389 $\pm$ 0.319
	Lumina-DiMOO	0.623	0.832	0.534 $\pm$ 0.309	0.703 $\pm$ 0.256	0.477 $\pm$ 0.330
	Show-o2-7B	0.634	0.817	0.518 $\pm$ 0.303	0.700 $\pm$ 0.250	0.451 $\pm$ 0.315
	OmniGen2	0.630	0.807	0.484 $\pm$ 0.318	0.678 $\pm$ 0.265	0.527 $\pm$ 0.321
	BAGEL-7B-MoT	0.638	0.798	0.472 $\pm$ 0.313	0.651 $\pm$ 0.274	0.488 $\pm$ 0.323
	Janus-Pro-7B	0.636	0.735	0.371 $\pm$ 0.304	0.564 $\pm$ 0.292	0.544 $\pm$ 0.304
Diffusion	SD 3.5 Large	0.648	0.841	0.553 $\pm$ 0.296	0.730 $\pm$ 0.250	0.455 $\pm$ 0.325
	FLUX.2[klein]	0.630	0.829	0.551 $\pm$ 0.313	0.716 $\pm$ 0.255	0.482 $\pm$ 0.332
	SANA1.5-4.8B	0.635	0.828	0.541 $\pm$ 0.304	0.710 $\pm$ 0.253	0.428 $\pm$ 0.320
	Z-Image Turbo	0.625	0.782	0.431 $\pm$ 0.322	0.656 $\pm$ 0.281	0.532 $\pm$ 0.320
	FLUX.1[dev]	0.631	0.780	0.484 $\pm$ 0.317	0.624 $\pm$ 0.279	0.525 $\pm$ 0.325
	Kolors	0.637	0.752	0.379 $\pm$ 0.319	0.590 $\pm$ 0.293	0.521 $\pm$ 0.319
Spearman ρ vs. VQAScore		—	—	$+0.91$	$+0.98$	$\mathbf{-0.91}$
Spearman ρ vs. CaRVE		—	—	$+0.42$	$+0.47$	$\mathbf{-0.61}$

2026)), reasoning-augmented models (Think-Then-Generate (Kou et al., 2026) and Bagel-Zebra-CoT (Li et al., 2025)), unified autoregressive models (Qwen-Image (Wu et al., 2025a), Lumina-DiMOO (Xin et al., 2025), Show-o2-7B (Xie et al., 2025b), OmniGen2 (Wu et al., 2025b), BAGEL-7B-MoT (Deng et al., 2025), and Janus-Pro-7B (Chen et al., 2025)), and diffusion-based models (Stable Diffusion 3.5 Large (Esser et al., 2024), FLUX.2[klein] (Black Forest Labs, 2025), SANA1.5-4.8B (Xie et al., 2025a), Z-Image Turbo (Z-Image Team, 2025), FLUX.1[dev] (Black Forest Labs, 2024), and Kolors (Kolors Team, 2024)). All open-weight models are evaluated under their default inference configurations, while proprietary systems are queried using their public APIs.

Judges and reported scores. CanonScore is computed using VQAScore (Lin et al., 2024) with three independent vision-language judge lineages: Qwen2.5-VL-32B (Bai et al., 2025), InternVL3-14B (Chen et al., 2024), and VQAScore's in-house clip-flant5-xxl judge model (Lin et al., 2024). Unless otherwise stated, all reported CanonScore values are averaged across these three judges. We report the three CanonScore legs independently throughout : $\text{CS}_{\text{anchor}}$, CS_{resem}, and CS_{bound}.

Prompt conditions. To evaluate whether CanonScore distinguishes different aspects of visual resemblance generation, we compare three prompt conditions. The *base* condition uses the original VSG-Bench simile prompts as-is. The *Gemma-augmented* condition rewrites prompts using Gemma-4-31B-IT (Google DeepMind, 2026) before image generation. The *ViPE* condition applies the published prompt-rewriting baseline (Shahmohammadi et al., 2023).

Generation settings. All image generations are produced at 1024×1024 resolution under training-free inference. Open-weight models are executed using their default Hugging Face implementations on a single NVIDIA A100 GPU.

Table 3 individually reports each of the sixteen evaluated T2I systems as a profile over the three CANONSCORE legs: subject anchoring ($\text{CS}_{\text{anchor}}$), licensed-channel resemblance (CS_{resem}), and abstraction from the comparison object (CS_{bound}). Three claims follow, each read off a single row of the table. **First, no T2I system escapes the three failure modes.** The best $\text{CS}_{\text{anchor}}$ in the roster is 0.670 and the best CS_{bound} is 0.544, both far from the 1.0 a fully canonical generation would earn, and no model family supplies a T2I system that is strong on all three legs at once. **Second,**

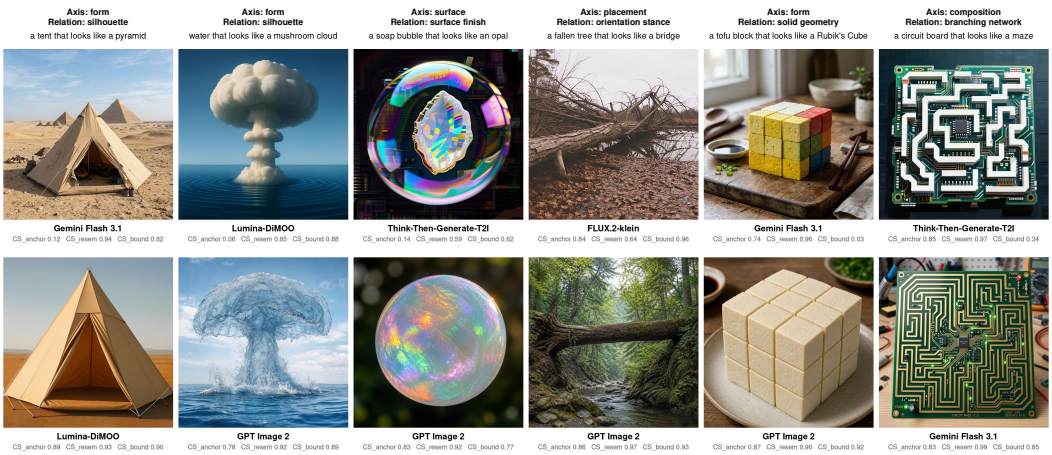


Figure 3: Qualitative showcases of six VSG-BENCH prompts, one per column, labeled with axis, relation, and base-form prompt. **Top row:** at least one CANONSCORE leg fails. Columns 1–3 fail CS_{anchor}, $\langle A \rangle$ no longer the sole subject. Column 4 trails on CS_{resem} alone, the fallen tree seen from an obscure view, not as a bridge. Columns 5–6 fail CS_{bound}, literally rendering a Rubik's Cube's colored pieces on a tofu block and a maze's walls on a circuit board. **Bottom row:** the same prompts, all three CANONSCORE legs holding. Leg scores and generating model appear beneath each image.

abstraction is the mode T2I systems fail worst: for nine of the sixteen, CS_{bound} is the lowest of that system's own three legs. **Third, resemblance and abstraction trade off against each other.** The three T2I systems with the strongest licensed-channel transfer, GPT Image 2, Gemini Flash 3.1 and Think-Then-Generate, are exactly the three weakest at abstraction, ranking 15, 16 and 14 of 16 on CS_{bound}.

Figure 3 makes these profiles concrete: for each of six prompts it pairs a generation on which a CANONSCORE leg registers a failure with one on which all three hold, so that what separates the two images is visible alongside the scores that separate them.

5.2 ESTABLISHING THE VALIDITY OF CANONSCORE

Each of the CANONSCORE legs carries its own information an undecomposed score discards. The undecomposed score is not a neutral summary of the three legs. The final row of Table 3 ranks the sixteen T2I systems by plain VQAScore and correlates that ordering with the ranking each leg induces, over the sixteen systems as observations. VQAScore reproduces the anchoring and resemblance orderings (Spearman $\rho = +0.91$ and $+0.98$) and reverses the abstraction ordering ($\rho = -0.91$); all three coefficients are significant at $p < 0.001$. It restates two legs and inverts the third, so the two proprietary T2I systems take its top two positions while ranking last and second-to-last on CS_{bound}, and Janus-Pro-7B takes its bottom position while ranking first. Over-literalization is the one failure mode an alignment score rewards, because importing $\langle B \rangle$ wholesale makes an image *more* responsive to the prompt text it is scored against. Appendix B.1 repeats the comparison for four further undecomposed scores (Table 9), including TIFA and DSG, which decompose a prompt into questions, and CaRVE, which is built for visual resemblance generation rather than for alignment in general. All five invert on CS_{bound} while agreeing on the other two legs, so the inversion is a property of scoring this task undecomposed rather than of any one score.

The CANONSCORE legs track a controlled intervention. We rewrite the VSG-BENCH prompts with Gemma-4-31B-IT and rescore the same 14 open-weight models, treating the rewrite as an experimental manipulation rather than as a proposed method. Augmentation raises CS_{bound} by 0.091 while reducing CS_{resem} by 0.098, so a gain in abstraction from $\langle B \rangle$ is bought against resemblance transfer; $\text{CS}_{\text{anchor}}$ rises by 0.078. Figure 4 shows that this is a property of the condition rather than an artifact of averaging, the per-model lines running near-parallel so that the legs separate in the same directions for essentially every model. A single scalar would register this intervention as a small net change, or as none.

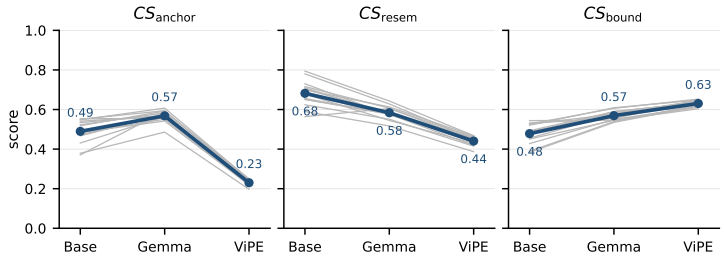


Figure 4: CANONSCORE legs across the three prompt conditions, on raw rather than normalized scores. Each thin gray line is one of the 14 open-weight models evaluated in all three conditions, and the heavy line is their mean. The model lines run near-parallel, so the shift in each leg is a property of the condition rather than of a few models driving the mean. Condition means are in Table 13.

Human raters recover the same separation. The evidence above is measured through the LLM judges. To establish the separation independently of them, we collected human ratings on the same images: ninety-three raters scored five representative T2I models on the 13 relation-matched prompts under both conditions, with the condition hidden and the two images for a prompt never shown together. Each image drew four ratings, combined into one human quantity per leg, coded so that all three rise with canonicity as the legs do (Appendix C.2). CANONSCORE reproduces the direction humans reported in thirteen of the fifteen model-leg cells, with Spearman rank correlations over the five models of $\rho = +1.00$ on $\mathrm{CS}_{\mathrm{resem}}$, $+0.90$ on $\mathrm{CS}_{\mathrm{anchor}}$, and $+0.80$ on $\mathrm{CS}_{\mathrm{bound}}$. A rank coefficient is used because with five points only the ordering is meaningful. Agreement is strongest where the dissociation lives: on $\mathrm{CS}_{\mathrm{resem}}$ all five models agree, four moving down together and FLUX.1[dev] moving up on both axes, so raters who were never told which condition they saw recovered both the common resemblance cost of augmentation and the one model that escapes it. $\mathrm{CS}_{\mathrm{anchor}}$ is the leg on which the intervention moves least: four of its five changes are small, so the panel asks raters to confirm the sign of a quantity that barely moved, and it agrees on that sign in four of the five. Its absolute values sit higher than the human ones, near 0.60 against 0.30, a constant offset that leaves the ordering intact. $\mathrm{CS}_{\mathrm{anchor}}$ is accordingly read as a comparison across conditions and models, which is what a diagnostic leg is for.

The three CANONSCORE legs therefore answer different questions, and the answers correspond to something raters can see in the images. A leg's movement therefore names a specific change in the image rather than a shift in an opaque score. Two scope conditions attach to the human result: with five models per panel a Spearman coefficient describes the ordering of a handful of points and no more, and the 13 prompts cover one item per relation, so the result pools across the taxonomy rather than speaking to any relation individually. Within those bounds the result is directional: it reports which way each model fails, not which model is better.

6 CONCLUSION

We formalize **Visual Simile Generation (VSG)** as a task whose difficulty is a measurement problem before it is a generation problem: a simile prompt does not say which attribute of the comparison object may transfer, so "*looks like*" has no criterion to be scored against. Naming that attribute per prompt is what makes the task testable, and it is what the benchmark and the metric are built on. **VSG-Bench** supplies 549 prompts over 13 resemblance relations, each annotating its licensed channel and the literal particulars a successful abstraction discards. **CANONSCORE** scores an image against those annotations along three legs, one per failure mode, reported separately rather than averaged.

Benchmarking 16 T2I systems shows that all three failure modes are present in every architecture family, and that the legs separate models a single score ranks alike. A controlled prompt intervention moves them in opposing directions, raising abstraction while lowering resemblance transfer, and ninety-three human raters scoring the same images recover that opposition without seeing the metric. What the instrument delivers is diagnostic resolution: not which model is better, but which way a model fails, which is what a T2I system aiming at figurative competence needs in order to know what to fix.

AI USE STATEMENT

Large language models (Claude Opus 4.6 and GPT-5.6-terra) were used to generate and vet candidate benchmark prompts, as detailed in Appendix A; We have utilized AI assistants, specifically Grammarly and ChatGPT, to correct grammatical errors and rephrase sentences.

ETHICS STATEMENT

VSG-Bench prompts are restricted to everyday, tangible, non-sensitive entities, and human raters in our survey were compensated fairly and provided informed consent for their participation.

REPRODUCIBILITY STATEMENT

All model configurations, inference settings, and licenses are detailed in Appendix C.1 and Appendix D, and the full benchmark construction and vetting pipeline is described in Appendix A.

REFERENCES

Anthropic. Claude opus-4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, and Sinan Tan. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. doi: 10.48550/arxiv.2309.16609.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3043–3054, 2023.

Herbert L Colston and Raymond W Gibbs. Figurative language communicates directly because it precisely demonstrates what we mean. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 75(2):228, 2021.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv*, 2024. doi: 10.48550/arxiv.2403.03206.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Google. Nano banana 2. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2026.

Google DeepMind. Gemma 4 31b (dense) model card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Model: google/gemma-4-31B; Apache 2.0 license; Accessed 2026-05-15.

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*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

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. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023a.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023b.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yugang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, and Zhijie Deng. Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. *arXiv preprint*, 2026.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

Manishit Kundu, Tejomay Kishor Padole, Sumit Shekhar, Biplab Banerjee, and Pushpak Bhattacharyya. CaRVE: Critiquing and refining visual elaborations for figurative language illustrations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42760–42777, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2122. URL https://aclanthology.org/2026.findings-acl.2122/.

Dahyeon Kye, Jeahun Sung, Minkyu Jeon, and Jihyong Oh. Chimera: Adaptive cache injection and semantic anchor prompting for zero-shot image morphing with morphing-oriented metrics. *arXiv preprint arXiv:2512.07155*, 2025. doi: 10.48550/arXiv.2512.07155.

Huiyuan Lai and Malvina Nissim. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Computing Surveys*, 56(10):1–34, 2024.

Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision language reasoning, 2025. URL https://arxiv.org/abs/2507.16746.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

OpenAI. Gpt image 2. https://openai.com/index/introducing-chatgpt-images-2-0/, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/cvpr52688.2022.01042.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. Understanding figurative meaning through explainable visual entailment, 2025. URL https://arxiv.org/abs/2405.01474.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, and Yichun Shi. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. doi: 10.48550/arxiv.2509.20427.

Hassan Shahmohammadi, Adhiraj Ghosh, and Hendrik Lensch. Vipe: Visualise pretty-much everything. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5477–5494, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.333. URL https://aclanthology.org/2023.emnlp-main.333/.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025.

Luecheng Wang, Shintaro Ozaki, Hidetaka Kamigaito, Katsuhiko Hayashi, Jingun Kwon, Manabu Okumura, and Taro Watanabe. Simile understanding in text-to-image models: An evaluation framework. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer, 2025a. URL https://arxiv.org/abs/2501.18427.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025b.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors, 2025. URL https://arxiv.org/abs/2506.06987.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. Irfl: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1044–1058, 2023.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Shaoan Zhao, Fang Zhao, Xueqiang Guo, Xinpei Su, Huanlin Gao, Qiang Hui, Ting Lu, Fuyuan Shi, Chao Tan, Bikun Yang, et al. Scalable question-centric text-to-image evaluation: Reliable ranking, fine-grained diagnosis, and cost-aware routing. *arXiv preprint arXiv:2608.24112*, 2026.

Table 4: Comparison with closely related T2I benchmarks for figurative and simile generation. **VSG-Bench** goes beyond detecting literalization by explicitly evaluating whether the intended resemblance attribute transfers while the target subject is preserved.

Benchmark	Simile T2I	Literal. Eval.	Attr. Transfer	Subject Preserv.	Failure Decomp.
T2I-ReasonBench (Sun et al., 2025)	✓	–	–	–	–
CaRVE / MetaCaRVE (Kundu et al., 2026)	–	✓	✓	–	–
Wang et al. (2026)	✓	✓	–	–	–
VSG-Bench (Ours)	✓	✓	✓	✓	✓

Literal. Eval.: detects literal rendering of the comparison object. *Attr. Transfer*: evaluates whether the intended visual resemblance attribute is transferred to the subject. *Failure Decomp.*: separately diagnoses subject, transfer, and literalization failures.

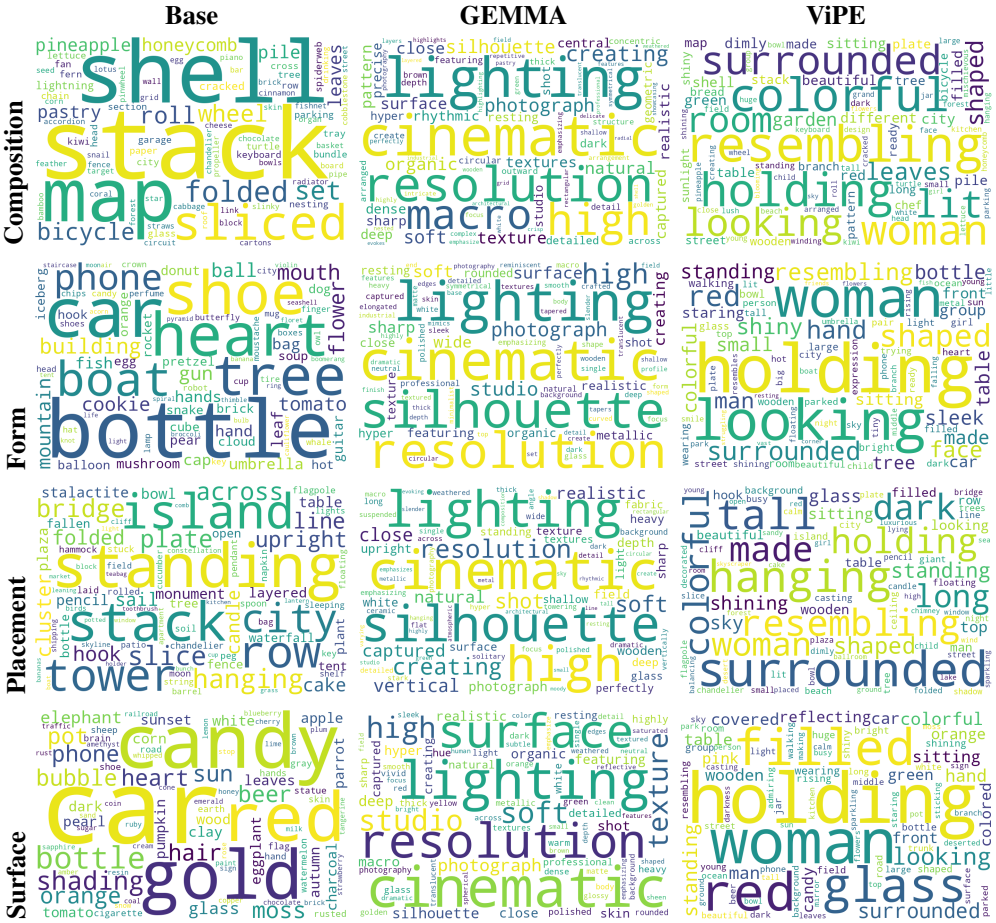


Figure 5: Word-cloud comparison across four visual dimensions for Base, GEMMA, and VIPE.

A Constructing VSG-Bench

This appendix gives the full construction procedure for the 549 prompts of **VSG-Bench**, summarized in Section 3.4. Every entry is a nominal phrase in the fixed template "$\langle A \rangle$ that looks like $\langle B \rangle$", labelled with one of four resemblance axes and one of thirteen fine-grained relations, and annotated with a licensed channel, three literal particulars, and a templated paraphrase.

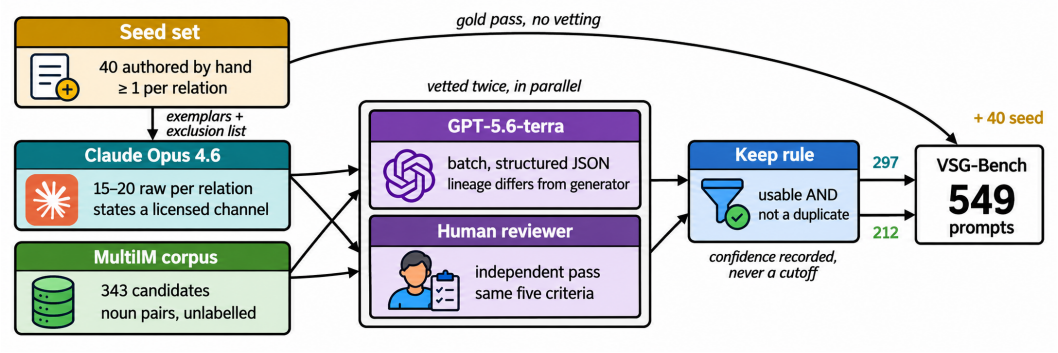


Figure 6: The VSG-BENCH curation pipeline. Three sources feed the benchmark. The seed set is authored and annotated by hand and takes a gold pass straight through; it also feeds generation, supplying the in-relation exemplars in each system prompt and the exclusion list later candidates are checked against for duplication. The generated and MultiMM streams are each vetted twice in parallel, once by a human reviewer and once by an automated vetter whose model lineage differs from the generator's, so that no model grades its own output. A candidate survives only if it is marked usable and not a duplicate; the vetter's confidence is recorded but never applied as a cutoff. Counts are the candidates entering and surviving each stream.

Table 5: Sources of the 549 VSG-BENCH prompts, with their distribution over the four resemblance axes. The seed set was authored by hand and given a gold pass; the other two streams passed the vetting procedure of Appendix A.3. Axis totals match the finalized dataset: Form 184, Placement 126, Surface 122, Composition 117.

Source	n	Form	Placement	Surface	Composition
Seed set (authored, gold pass)	40	20	9	5	6
Generated (vetted)	297	25	117	44	111
MultiMM (relabelled, vetted)	212	139	—	73	—
Total	**549**	184	126	122	117

A.1 SOURCES AND COMPOSITION

The benchmark draws on three streams, shown end to end in Figure 6 and broken down by axis in Table 5. A **seed set** of 40 handpicked prompts was authored and annotated by hand, with at least one valid instance per relation. These serve three purposes: they enter the final dataset directly, they calibrate generation as in-relation exemplars, and they are the reference list against which later candidates are checked for duplication. A **generated stream** of 297 prompts was produced by a large language model prompted per relation. A **MultiMM stream** of 212 prompts was drawn from the MultiMM corpus (Yang et al., 2025), which supplies (target, source) noun pairs without relation labels; each pair was assigned a relation and licensed channel under our scheme and then vetted like any generated candidate.

The three streams are complementary rather than redundant. MultiMM contributes overwhelmingly to Form and Surface, the two axes for which naturally occurring similes are most common, while the generated stream supplies the Placement and Composition relations that are sparse in found data. One relation, `solid_geometry`, is held as a capped control subset in which the geometry matches but no attribute transfers; it is deliberately not an expansion target.

A.2 GENERATING CANDIDATES

Candidates for the generated stream were produced with Claude Opus 4.6 (Anthropic, 2026) through a batch interface, with a separate system prompt per relation. Each system prompt carried the relation's definition, an authoring template, in-relation exemplars drawn from the seed set, and a global exclusion list of every $(\langle A \rangle, \langle B \rangle)$ pair already in the benchmark, across all relations. Generation proceeded at 15–20 raw candidates per relation.

Authoring protocol supplied to the generator, shown for `silhouette`

This document outlines the protocol for generating an evaluation dataset designed to challenge text-to-image models.

Defining a benchmark prompt A *prompt* is a nominal phrase asserting a visual resemblance between two distinct entities. Every prompt must adhere to the fixed template: `'<Subject> that looks like <Object>'`.

The relation you are authoring for `silhouette` (axis: FORM) — resemblance based on the outer 2-D outline or profile of `<Object>`. The *licensed channel* you state must name that outline specifically, not the object at large.

Mandatory constraints

- Both `<Subject>` and `<Object>` must be concrete, tangible entities observable in daily life, and must belong to different categories.
- Avoid potentially sensitive or controversial categories (e.g. human organs, insects, people).
- Avoid composite nouns that imply multiple items or background noise (e.g. "rust on a railing").
- Do not include positional modifiers (e.g. "in a cup", "on a hill").
- Do not reuse any pair on the EXISTING PAIRS list, in either order, and do not repeat a donor `<Object>` within this batch.
- The resemblance must be non-trivial, and the `<Subject>` → `<Object>` direction unambiguous.

Exemplar prompts

- *Positive:* "The moon that looks like a boat" → `{"licensed_channel": "the boat's upward-curving hull profile"}`
- *Positive:* "A sandwich that looks like a book" → `{"licensed_channel": "the book's stacked layering between two covers"}`
- *Negative:* "A knife that looks like a sword" → entities belong to the same functional subclass.
- *Negative:* "A spirit that looks like democracy" → relies on entirely abstract, non-visual concepts.

Output Return `A`, `B`, `prompt`, `relation`, `licensed_channel`, and `reasoning` for each candidate.

Figure 7: The per-relation authoring protocol given to the generator, instantiated here for `silhouette`. Each of the thirteen relations receives this protocol with its own definition and its own in-relation exemplars drawn from the seed set, alongside a global exclusion list of every pair already in the benchmark. The requirement to state a *licensed channel* at generation time is what allows the vetting step of Appendix A.3 to check the claimed attribute rather than only the noun pair.

Figure 7 reproduces the authoring protocol carried in each system prompt. The relation definition and exemplars shown there are the ones supplied for `silhouette`; the remaining twelve relations receive the same protocol with their own definition and their own seed exemplars.

Each candidate was emitted with its relation, its proposed licensed channel, and the generator's stated reasoning, so that the vetting step could check the claimed attribute rather than only the noun pair.

A.3 VETTING

Every candidate outside the seed set was vetted twice, independently and in parallel: once by a human reviewer and once by GPT-5.6-terra under the system prompt of Figure 8. The vetting model is from a different lineage than the generator, so that no model grades its own output. A candidate entered the benchmark only if it was marked usable *and* not a duplicate.

In the MultiMM stream, 233 of 343 candidates passed vetting, and 212 remained after relabelling. Rejections were dominated by relation errors rather than by unusable noun pairs: 58 wrong relation, 33 other, 16 trivial resemblance, 2 duplicate, and 1 ambiguous direction. That distribution is a

Prompt vetting system prompt

You screen candidate noun pairs for a text-to-image benchmark of visual similes of the form "an A that looks like a B".

A pair is USABLE only if ALL of these hold:
1. A and B are both concrete, physically depictable objects.
2. A and B belong to different categories (not two kinds of the same thing).
3. The stated relation and licensed_channel accurately describe what attribute of B could transfer to A -- check the claim, don't just accept it.
4. The resemblance is non-trivial: A and B do not already look alike by default outside the stated relation.
5. The direction is unambiguous: A is clearly the anchor, B clearly the donor. If it reads equally well reversed, mark ambiguous_direction.

Also check: does this pair duplicate (in either order) any pair in the EXISTING PAIRS list provided? If so, set duplicate=true regardless of usability.

For each candidate return:
{"id": "...", "usable": true/false, "duplicate": true/false,
"relation_correct": true/false, "confidence": 0.0-1.0,
"reject_reason": "one of: same_category | trivial_resemblance | ambiguous_direction | wrong_relation | duplicate | other | null"}

Confidence anchors: 0.9 = a clear, striking simile a curator would keep on sight. 0.6 = plausible but wants a human check. 0.3 = weak, probably reject. Use the full range; do not cluster around 0.7.

Return ONLY a JSON array, one object per input candidate.

Figure 8: The system prompt given to the automated vetter, reproduced as submitted. The seed-pair exclusion list and the candidate's fields (identifier, both nouns, prompt text, claimed relation and claimed licensed channel) followed in the user message. The `confidence` field was recorded for analysis but never used as an admission cutoff.

Table 6: Representative rejected candidates, with the vetter's reason and confidence. The claimed licensed channel is the one proposed at generation, not an annotation that survives into the benchmark. Rejections turn on the claimed attribute as often as on the noun pair: the two *wrong relation* rows name a channel that belongs to a different relation than the one claimed.

Candidate prompt	Claimed relation	Reject reason	Conf.
a head that looks like an egg	`silhouette`	trivial resemblance	0.80
a bunch of asparagus that looks like a bamboo grove	`repeating_units`	same category	0.88
a row of garden stakes that looks like a fence	`arrangement_of_multiples`	same category	0.94
a corn dog that looks like a frankenstein	`hue_match`	wrong relation: the channel given is a head shape, not a colour	0.68
a roll of tape that looks like a doughnut	`solid_geometry`	ambiguous direction	0.45
a city that looks like bottles	`silhouette`	ambiguous direction	0.63
a surfboard laid on sand that looks like a runway	`orientation_stance`	wrong relation, and carries a positional modifier	0.62
a fishing net that looks like a spreadsheet	`repeating_units`	other: $\langle B \rangle$ is not a physical object	0.85

consequence of treating MultiMM pairs as unlabelled. The corpus supplies noun pairs that are genuine similes, so the pair itself usually survives; what fails is our first assignment of which attribute the comparison licenses, which the vetter is explicitly asked to check.

Literal particular annotation system prompt

You annotate a benchmark of visual similes of the form "an A that looks like a B".

Each item names a RELATION and a LICENSED CHANNEL: the one attribute of B that is permitted to transfer to A. Canon means A carries B's attributes within that channel. A literal particular is an attribute of B that appears on A from OUTSIDE that channel.

Given A, B, the relation, and the licensed channel, propose exactly 3 literal attributes. Each must satisfy ALL of:
1. It lies outside the licensed channel.
2. It is NOT already true of A. Test: would a plain, correct photo of A by itself plausibly answer yes? If so, discard it.
3. It is visible in a single still image (no motion, sound, or causality).
4. It does not necessarily co-occur with either of the other two.

The RELATION determines which attribute is licensed. The licensed_channel text is supporting description only: it often mentions more than one attribute, and it often says where B is normally found. Ignore the extras and read the licensed attribute off the relation. Always return 3 literal particulars; never refuse an item. (One exception: under partial_form the channel names WHICH sub-part of B is licensed, so read that from the channel text.)

Prefer literal particulars a text-to-image model would plausibly render. Use 3 distinct categories where possible.

Categories: material_texture | color | shape_form | part_component | context_scene

Write each as a short DECLARATIVE CAPTION describing A carrying that literal particular. This string goes into VQAScore's text slot, which takes a caption, not a question -- so never write an interrogative. The scorer wraps it as "a photo of {probe}".

The caption must be headed by A's noun, lowercase, with no final period, and must name the visual evidence rather than the concept. Good: "a cherry with cut facets". Bad: "Does the cherry have cut facets?", "a cherry that looks like a gemstone". No negations, no comparatives, at most 3 coordinated items.

Return ONLY a JSON array:
[{"category": "...", "attribute": "<short noun phrase>", "probe": "<short caption>"}]

Figure 9: The instruction given to the annotating model to elicit the three *literal particulars* of every VSG-BENCH entry, the features of $\langle B \rangle$ that a successful abstraction must leave out, reproduced as submitted. The four numbered conditions are what keep a particular scorable: it must fall outside the licensed channel, so that discarding it is not in tension with the resemblance the prompt does license; it must not already hold of $\langle A \rangle$, so that an ordinary image of $\langle A \rangle$ cannot satisfy it; it must be settled by a single still image; and it must be independent of the other two, so that three probes report three separate ways of over-literalizing rather than one. Each particular is returned as a declarative caption headed by $\langle A \rangle$'s noun, since VQAScore scores a caption rather than a question, which is the form the s_ℓ probes of Table 2 take.

Table 6 gives representative rejections. Two patterns are worth naming. First, several candidates fail because the pair is too close to begin with: a head and an egg, or asparagus and bamboo, already look alike without any resemblance being asserted, so the prompt asks a generation for nothing. Second, a candidate can name a plausible resemblance under the wrong relation, as when a corn dog's likeness to a Frankenstein figure is filed under `hue_match` while the channel actually given is a head shape. Both failures are invisible to a check on the noun pair alone, which is why the claimed licensed channel is part of what the vetter screens.

Table 7: CANONSCORE by taxonomic axis, pooled over the sixteen T2I systems and the three judges. n is the number of VSG-BENCH prompts on that axis.

Axis	n	CS_{anchor}	CS_{resem}	CS_{bound}
Form	184	0.466	0.655	0.487
Placement	126	0.516	0.742	0.412
Surface	122	0.490	0.722	0.415
Composition	117	0.579	0.704	0.499

Table 8: CANONSCORE by resemblance relation, nested under its axis, pooled over the sixteen T2I systems and the three judges. n is the number of VSG-BENCH prompts carrying that relation. The two smallest relations are reported with their counts rather than suppressed, and their values should be read accordingly.

Axis	Relation	n	CS_{anchor}	CS_{resem}	CS_{bound}
Form	`silhouette`	151	0.452	0.650	0.495
	`partial_form`	19	0.538	0.734	0.414
	`solid_geometry`	14	0.520	0.600	0.496
Placement	`orientation_stance`	67	0.488	0.709	0.433
	`arrangement_of_multiples`	30	0.597	0.826	0.456
	`suspension`	29	0.499	0.732	0.316
Surface	`surface_finish`	56	0.487	0.700	0.404
	`hue_match`	54	0.481	0.715	0.459
	`gradient_multitone`	7	0.571	0.855	0.357
	`patina_weathering`	5	0.510	0.863	0.151
Composition	`repeating_units`	49	0.536	0.682	0.494
	`nested_layering`	36	0.611	0.683	0.514
	`branching_network`	32	0.608	0.760	0.491

A.4 PER-ENTRY ANNOTATION

Each surviving prompt was annotated with the three fields the evaluation pipeline consumes, described in Section 3.4 and shown in full for one entry in Table 1: its *licensed channel*, exactly three *literal particulars*, and a templated paraphrase that states the licensed channel concretely. All 549 entries carry exactly three literal particulars.

These annotations are authored per instance rather than inherited from the relation, which is what allows CANONSCORE to condition on a specific attribute instead of on $\langle B \rangle$ at large. The licensed channels are correspondingly near-unique: across the 151 `silhouette` entries there are 150 distinct licensed-channel strings, so an entry's evaluation criterion is effectively its own rather than a relation-level template applied 151 times.

Figure 9 gives the instruction used to elicit the literal particulars. It states the four conditions a particular has to meet to be scorable at all, and fixes the surface form each one is written in.

A.5 RESULTS BY AXIS AND BY RELATION

The main text reports CANONSCORE pooled over all 549 prompts. Because the thirteen relations are unevenly sized, with `silhouette` alone holding 151 entries, that pooling could in principle be reporting one relation rather than the benchmark. This section gives the same scores broken out by axis and by relation, and compares the two ways of averaging them.

The benchmark is not reporting one relation. Averaging the thirteen relations equally, so that `patina_weathering` with five entries counts as much as `silhouette` with 151, moves each leg by at most 0.035: CS_{anchor} from 0.507 to 0.531, CS_{resem} from 0.700 to 0.731, and CS_{bound} from 0.456 to 0.422. The direction of that last shift is the informative one. Weighting the small relations up lowers CS_{bound}, because the relations with the fewest entries are the ones on which abstraction is hardest, so the pooled figure the main text reports is the more conservative of the two.

The tradeoff holds on every axis. The opposition between resemblance transfer and abstraction is not an artifact of pooling axes that behave differently. Computing the same rank correlation within each axis separately, over the sixteen T2I systems, gives $\rho = -0.90$ on Form, -0.80 on Placement, -0.95 on Surface and -0.78 on Composition, against -0.89 pooled. Each axis reproduces the effect on its own prompts.

Where the axes differ. The axes are not interchangeable, and the differences are legible as properties of what each asks for. Composition scores highest on CS_{anchor} at 0.579, since a prompt about how parts are arranged leaves $\langle A \rangle$ intact as the object being arranged. Placement and Surface score lowest on CS_{bound}, at 0.412 and 0.415: an orientation or a finish is carried by the same surface that carries $\langle B \rangle$'s other particulars, so importing the licensed channel and importing what surrounds it are hard to separate. `suspension` and `patina_weathering` are the two hardest relations for abstraction, at 0.316 and 0.151, the latter over five entries.

B Validating the Efficacy of CanonScore

B.1 Efficacy in Comparison with Existing Text–Image Alignment Metrics

Section 5.1 reports that plain VQAScore (Lin et al., 2024) tracks two CanonScore legs and inverts the third. This section gives the full ranking behind that claim and states what it implies for alignment metrics as a class.

VQAScore is computed on the same 549 base prompts, the same generated images and the same clip-flant5-xxl judge as the CanonScore sub-scores, scoring each image against its unmodified base prompt. Ranking the sixteen T2I systems by that score and correlating those ranks with the ranks each leg induces gives Spearman $\rho = +0.91$ against CS_{anchor}, $+0.98$ against CS_{resem} and -0.91 against CS_{bound}. The agreement with CS_{resem} is near-total: VQAScore asks whether the image shows "$\langle A \rangle$ that looks like $\langle B \rangle$", so a T2I system that puts more of $\langle B \rangle$ into the image answers that question better, whatever it does to $\langle A \rangle$'s identity or to the literal particulars the simile asks a generation to discard.

That is what produces the inversion. The two proprietary T2I systems occupy the top two VQAScore positions of the roster and the bottom two on CS_{bound}; Janus-Pro-7B occupies the last VQAScore position and the first on CS_{bound}. The ordering is not noisy disagreement at the margin but a reversal along the length of the roster, and it locates the property an alignment score cannot represent: Over Literalization is a failure that makes the image *more* responsive to the prompt text, so any metric scoring responsiveness to that text rewards it. Raising the number of literal particulars imported raises VQAScore and lowers CS_{bound} at the same time.

This generalizes past VQAScore. CLIPScore (Hessel et al., 2021) and PickScore (Kirstain et al., 2023) score an image against the prompt or against aggregate human preference, and neither has access to the annotation that distinguishes a licensed channel from a literal particular, since that annotation is a property of the VSG-Bench entry rather than of the image or the prompt string. The decomposition is what supplies it: CS_{bound} is computable only because each entry names the set $\mathcal{L}$ its generations must abstract away from. An alignment metric is not mis-tuned for this task. It is scoring a different quantity, and on VSG the two quantities are opposed.

B.2 Ablation Study: Aggregation Choices in the Three Legs

Each leg combines its sub-scores differently: CS_{anchor} takes a product, CS_{resem} an average, and CS_{bound} a worst case. This section tests those three choices against the alternatives they were picked over, holding the sub-scores and the images fixed so that only the aggregation varies.

Each variant is read two ways.

The first is *agreement with the human survey*, the ninety-three-rater study introduced in Section 5.2 and detailed in Appendix C.2. We recompute the leg under a variant, take its change between the base and Gemma-augmented conditions on the 13 relation-matched prompts, and compare that change with the change in the corresponding human rating, across the five T2I systems the survey covers. Two quantities summarize the comparison: a Spearman correlation ρ between the two sets of five

Table 9: Five undecomposed scores against the three CANONSCORE legs, all computed on the same images and the same 549 base prompts. Each cell is the Spearman correlation between that score's ranking of the sixteen T2I systems and the ranking the leg induces. VIEScore (Ku et al., 2024), TIFA (Hu et al., 2023a) and DSG (Zhao et al., 2026) are run with their official code under a Qwen2.5-VL-32B judge, VIEScore as its overall score out of ten and DSG with dependency zero-out. Rows are ordered by agreement on CS_{bound}. *Range* is the spread of the score itself across the sixteen systems, against 0.254 for CS_{bound}.

Undecomposed score	CS_{anchor}	CS_{resem}	CS_{bound}	Range
VQAScore (Lin et al., 2024)	$+0.91$	$+0.98$	$\mathbf{-0.91}$	0.187
TIFA (Hu et al., 2023a)	$+0.90$	$+0.98$	$\mathbf{-0.88}$	0.280
DSG (Zhao et al., 2026)	$+0.90$	$+0.99$	$\mathbf{-0.86}$	0.391
CaRVE (Kundu et al., 2026)	$+0.42$	$+0.47$	$\mathbf{-0.61}$	0.031
VIEScore (Ku et al., 2024)	$+0.77$	$+0.70$	$\mathbf{-0.56}$	0.085

Table 10: Aggregation variants for each CANONSCORE leg, over that leg's sub-scores as Table 2 defines them. ρ is the Spearman correlation between the variant's change under prompt augmentation and the change in the corresponding human rating, over the five T2I systems in the human survey; *sign agree* counts how many of those five moved in the same direction under both. SD and Range describe the variant's sixteen model-level means over the 549 base prompts, one per T2I system, and say how far apart the variant places the roster. The aggregation chosen for the CanonScore design is marked †.

Leg	Aggregation	ρ vs. human	Sign agree	SD	Range
CS_{anchor}	product†	$+0.90$	$4/5$	0.076	0.299
	arithmetic mean	$+0.80$	$4/5$	0.044	0.173
	geometric mean	$+0.50$	$4/5$	0.062	0.241
	minimum	$+0.40$	$4/5$	0.069	0.271
CS_{resem}	average†	$+1.00$	$5/5$	0.079	0.278
	channel only	$+1.00$	$5/5$	0.078	0.280
	relation only	$+1.00$	$5/5$	0.080	0.277
CS_{bound}	$1 - \max(\text{literals})$†	$+0.80$	$4/5$	0.077	0.254
	$1 - \text{mean}(\text{literals})$	$+0.90$	$5/5$	0.053	0.169

changes, and a *sign agreement* count, the number of those five systems for which the variant and the raters moved in the same direction.

The second is *discrimination across the roster*. Each of the sixteen T2I systems gets one value per variant, its mean over all 549 base prompts, and we report the standard deviation and the range of those sixteen values. Both say how far apart the variant places the systems, which is what a leg must do to profile one system against another.

The two readings check different things. A variant that agrees with raters but compresses the roster into a narrow band cannot profile a T2I system, and a variant that separates systems while disagreeing with raters is not measuring what it claims to.

The product is the strongest aggregation for CS_{anchor}. It leads on both readings at once, at $\rho = +0.90$ against the human ratings and a range of 0.299 across the roster. The arithmetic mean is the informative comparison: it compresses the roster to a range of 0.173, a little over half the product's, which is the behavior Section 4 predicts from a sum. A T2I system that renders $\langle A \rangle$ faithfully while also placing $\langle B \rangle$ beside it scores well on two factors and poorly on one, and averaging lets the two carry it. The product suppresses the leg whenever any factor is low, which is what a soft conjunction is for. The minimum and the geometric mean are the same conjunction taken harder and softer respectively, and both fall to $\rho = +0.40$–0.50: the minimum discards everything except the weakest factor, and with it the evidence the other two carry.

The two resemblance probes are interchangeable, and the average is taken for robustness. All three variants reach $\rho = +1.00$ and $5/5$, and their roster ranges fall within 0.003 of one another. The average is not preferred here because it measures better, but because it depends on neither probe alone. Section 4 argues that the two fail in opposite ways. They are also not redundant, agreeing at

Table 11: Each judge scored separately on all 549 base prompts. The leg columns are means over the sixteen T2I systems. The ρ columns give the Spearman correlation between the ordering of those sixteen systems under a single judge and their ordering under the three-judge average the paper reports.

	Leg mean			ρ vs. three-judge average		
Judge	CS_{anchor}	CS_{resem}	CS_{bound}	CS_{anchor}	CS_{resem}	CS_{bound}
CLIP-FlanT5-XXL	0.369	0.776	0.303	+0.93	+0.98	+0.95
InternVL3-14B	0.607	0.765	0.494	+0.91	+0.96	+0.99
Qwen2.5-VL-32B	0.545	0.559	0.572	+0.99	+0.99	+0.99
Leave-one-out: the remaining two judges, against the full average						
without CLIP-FlanT5-XXL				+0.99	+1.00	+1.00
without InternVL3-14B				+1.00	+0.99	+0.98
without Qwen2.5-VL-32B				+0.99	+0.97	+0.98

$r = 0.73$–0.85 across the three judges, below the 0.9 at which we would treat one as restating the other. Averaging is insurance against either failing on an entry rather than an improvement on this evidence.

The worst case is what makes CS_{bound} diagnostic. One unmistakable literal particular of $\langle B \rangle$ is enough to read an image as $\langle B \rangle$ over-literalized, however cleanly the others are left out. Only the maximum reports that. Averaging admits the particulars that were left out as evidence of abstraction, so an image carrying a mast is credited for having omitted the planks and the water, and the leg no longer names the failure it is defined on.

B.3 Per-Judge Scores and Leave-One-Judge-Out Robustness

CanonScore is LVLM-judge-based, and every value the paper reports is averaged over three judge model architecture lineages. This section asks what that average rests on: whether any one lineage carries the result, and what changes if it is removed.

The judges disagree on level and agree on order. The leg means in Table 11 differ widely between lineages: CS_{anchor} spans 0.369 to 0.607 across the three, and CS_{bound} spans 0.303 to 0.572. Those gaps are larger than most of the differences between T2I systems that the paper reports. The orderings nonetheless coincide. Every single-judge ranking of the sixteen systems correlates with the three-judge ranking at $\rho \geq +0.91$ on all three legs, and the pairwise agreements between lineages run $+0.83$ to $+0.97$. This is the concrete form of the reading the paper asks for throughout: a leg value is interpreted against other conditions and other systems, never as an absolute level, because the level is the part that moves with the judge.

No single lineage carries the result. Dropping any one judge and averaging the other two leaves the ordering essentially intact, at $\rho \geq +0.97$ on every leg, with the largest shift in any system's leg mean being 0.118. The resemblance–abstraction tradeoff of Section 5.1 is likewise reproduced by each lineage on its own, at $\rho = -0.91$, -0.89 and -0.89 against the three-judge -0.89, so the opposition the decomposition is built to expose does not depend on which judge observes it.

What varies is where the worst leg falls. The claim that abstraction is the mode T2I systems fail worst is the one reading that is judge-dependent: CS_{bound} is a system's lowest leg for 14 and 13 of the sixteen under CLIP-FlanT5-XXL and InternVL3-14B, but for 6 under Qwen2.5-VL-32B, which scores that leg highest of the three lineages in absolute terms. The averaged figure reported in the main text, nine of sixteen, sits between them. The tradeoff and the orderings survive the disagreement; the count of systems whose worst leg is CS_{bound} should be read as a property of the three-judge average rather than as a judge-independent fact.

B.4 Item-Level Coupling of the Resemblance and Abstraction Legs

Section 5.2 takes each of the 16 evaluated T2I systems as one observation, with its score on a leg the mean over all benchmark items, and finds CS_{resem} and CS_{bound} strongly opposed at Pearson

Table 12: Default load-time configurations and estimated model sizes in (B)illions regarding the fourteen open-weight T2I models used in this study, grouped by model category and ordered as in Table 3. For the unified models, which generate from a single language backbone instead of a frozen encoder, the Text Encoder column names that backbone. Each T2I model's pipeline is moved to a single NVIDIA A100 GPU after instantiation.

Category	Model	Text Encoder		Visual (B)	Total (B)
		Name	Size (B)		
CoT	Think-Then-Generate	Qwen2.5-VL-7B	8.3	20.0	28.3
	Bagel-Zebra-CoT	Qwen2.5-7B-Instruct	7.6	7.0	14.0
Unified	Qwen-Image	Qwen2.5-VL-7B	8.3	20.0	28.3
	Lumina-DiMOO	LLaDA-8B-Base	8.0	8.0	8.0
	Show-o2-7B	Qwen2.5-7B-Instruct	7.6	0.9	8.5
	OmniGen2	Qwen2.5-VL-3B-Instruct	3.0	4.0	7.0
	BAGEL-7B-MoT	Qwen2.5-7B-Instruct	7.6	7.0	14.0
	Janus-Pro-7B	DeepSeek-LLM-7B-base	7.0	7.0	7.0
Diffusion	SD 3.5 Large	T5-XXL + CLIP-L/G	5.7	8.1	13.8
	FLUX.2[klein]	Mistral-Small-3.1	24.0	9.0	33.0
	SANA1.5-4.8B	Gemma-2-2B-IT	2.0	$\sim$2.8	4.8
	Z-Image Turbo	Qwen2.5-VL-7B	8.3	6.0	14.3
	FLUX.1[dev]	T5-XXL + CLIP-L	4.9	11.9	16.8
	Kolors	ChatGLM3-6B	6.2	2.6	8.8

$r = -0.92$. That coupling is weaker than it looks. Recomputing the same coefficient over individual images rather than over model means, CS_{resem} and CS_{bound} correlate only at $r = -0.35$ to -0.41 ($n = 8{,}784$ images per judge): images that carry the abstraction without importing the instance are not rare, and the tight coupling appears only under aggregation, as item-level variation cancels and each model's systematic disposition remains. The two levels together locate the deficit in each model's disposition rather than in the prompts it was given, an attribution no single score can make.

C Additional Experimental Details

C.1 Model Configurations for Evaluation

We compare fourteen open-weight text-to-image pipelines, all instantiated through `HuggingFace Hub` and `diffusers`. For each model we keep the upstream defaults exposed by its `from_pretrained` loader and only override precision, the `fp16` weights variant where one is published. No CFG scale, scheduler, or sampler is overridden at load time; all images are generated at 1024×1024 pixels with `num_inference_steps=30` (except that Qwen-Image uses `num_inference_steps=50` by default). All hyperparameters not listed in Table 12 are inherited from the upstream Hugging Face checkpoint.

C.2 Human Survey Design and Full Results

Section 5.2 reports the headline agreement between CanonScore and the human survey. This section gives the survey's construction and the results omitted there.

Ninety-three raters scored the outputs of five representative T2I models on the 13 relation-matched prompts, under both the base and the Gemma-augmented condition. Each image received four ratings, which combine into three human quantities aligned with the legs: subject anchoring from whether $\langle A\rangle$ remains the main object and $\langle B\rangle$ does not appear separately, resemblance from whether $\langle A\rangle$ carries the licensed channel of $\langle B\rangle$, and abstraction from whether *at least one* of the three annotated literal particulars of $\langle B\rangle$ appears on or around $\langle A\rangle$, which is the form Figure 10 shows and the reason the human quantity is a worst case rather than an average. The two items that ask after a failure, separate depiction of $\langle B\rangle$ and leakage of its literal particulars, are reverse-coded, so that all three human quantities increase with canonicity in the same direction as the legs they are compared against. Raters saw only the original prompt; the condition was hidden, and the two images for a prompt were never shown together; Figure 10 shows one rating page as it was presented. The test is directional: for each model and each leg, does the human rating move the way CanonScore moved?

Table 13: CANONSCORE legs under the three prompt conditions, averaged over the same 14 open-weight models in every row, since the two proprietary systems have no augmented run. Scores are raw, not normalized, and parentheses give the change from base. Both conditions raise CS_{bound} while lowering CS_{resem}. Per-model scores are in Appendix C.3.

Condition	CS_{anchor}	CS_{resem}	CS_{bound}
Base	0.490	0.682	0.478
Gemma-augmented	0.568 (+0.078)	0.584 (−0.098)	0.569 (+0.091)
ViPE	0.230 (−0.260)	0.441 (−0.241)	0.631 (+0.152)

Table 14: Per-model CANONSCORE legs on the **Gemma-augmented** prompts; scores are raw, not normalized. Models are listed in the category order of Table 3; the two proprietary systems have no augmented run and are omitted. ± reports the mean per-judge standard deviation across items. Condition means are in Table 13.

Category	Model	CS_{anchor}	CS_{resem}	CS_{bound}
CoT	Think-Then-Generate	0.607 ± 0.266	0.643 ± 0.262	0.535 ± 0.335
	Bagel-Zebra-CoT	0.576 ± 0.273	0.582 ± 0.270	0.589 ± 0.317
Unified	Qwen-Image	0.583 ± 0.272	0.628 ± 0.258	0.539 ± 0.329
	Lumina-DiMOO	0.571 ± 0.270	0.583 ± 0.260	0.590 ± 0.314
	Show-o2-7B	0.592 ± 0.263	0.613 ± 0.258	0.545 ± 0.317
	OmniGen2	0.542 ± 0.285	0.545 ± 0.273	0.606 ± 0.316
	BAGEL-7B-MoT	0.574 ± 0.274	0.577 ± 0.266	0.583 ± 0.318
	Janus-Pro-7B	0.591 ± 0.264	0.605 ± 0.258	0.546 ± 0.322
Diffusion	SD 3.5 Large	0.556 ± 0.273	0.573 ± 0.271	0.577 ± 0.319
	FLUX.2[klein]	0.594 ± 0.269	0.607 ± 0.262	0.565 ± 0.331
	SANA1.5-4.8B	0.567 ± 0.267	0.575 ± 0.259	0.550 ± 0.320
	Z-Image Turbo	0.564 ± 0.276	0.576 ± 0.278	0.554 ± 0.330
	FLUX.1[dev]	0.551 ± 0.277	0.551 ± 0.268	0.577 ± 0.323
	Kolors	0.486 ± 0.288	0.517 ± 0.264	0.609 ± 0.295

Pooled across raters, the human quantities move by -0.115 on resemblance ($p = 7.8\text{e}{-10}$), $+0.186$ on abstraction ($p = 3.6\text{e}{-13}$), and -0.070 on anchoring ($p = 1.3\text{e}{-11}$). On CS_{bound} four of the five models agree in sign, and all four of those move up. The two legs that the perturbation condition moves in opposite directions are therefore the two the raters independently confirm, and they confirm the opposition rather than merely the existence of change.

On $\text{CS}_{\text{anchor}}$ the one disagreeing cell is Show-o2-7B, whose CANONSCORE change under augmentation is among those nearest zero, so the disagreement is a sign assigned to a quantity that barely moved rather than a reversal of a clear effect.

These three pooled deltas and the three rank correlations of Section 5.2 are the only quantities the paper reports a p-value for. Every other correlation it gives is descriptive. Each of the six was specified before it was run, one per leg, and none was selected post hoc on effect size. All six clear any conventional correction for six comparisons by several orders of magnitude.

C.3 PER-MODEL SCORES UNDER THE PERTURBATION CONDITIONS

Section 5.2 reports the three CANONSCORE legs averaged across models for each prompt condition. Tables 14 and 15 give the underlying per-model scores for the Gemma-augmented and ViPE conditions; base-condition scores are in Table 3. Both tables use the model ordering of Table 3 so that the three conditions can be read against one another row by row. The two proprietary systems were not run under either rewriting condition and are therefore omitted from both tables, and the condition means in Table 13 are computed over the 14 models common to all three conditions.

The ViPE operating point. The published ViPE rewriting baseline occupies a markedly different point from the Gemma-augmented condition. It pushes CS_{bound} higher still (0.631) while collapsing $\text{CS}_{\text{anchor}}$ to 0.230, a level at which the subject is largely no longer recoverable, and its between-model spread narrows to 0.053 on that leg. Read as an operating point, this shows that the three legs are

Table 15: Per-model CANONSCORE legs on the **ViPE-rewritten** prompts; scores are raw, not normalized. Models are listed in the category order of Table 3; the two proprietary systems have no rewritten run and are omitted. ± reports the mean per-judge standard deviation across items. ViPE has no counterpart in the human study, so these scores support no human-grounded claim. Condition means are in Table 13.

Category	Model	CS_{anchor}	CS_{resem}	CS_{bound}
CoT	Think-Then-Generate	0.248 ± 0.294	0.468 ± 0.324	0.617 ± 0.313
	Bagel-Zebra-CoT	0.219 ± 0.273	0.435 ± 0.306	0.635 ± 0.297
Unified	Qwen-Image	0.220 ± 0.274	0.450 ± 0.306	0.623 ± 0.297
	Lumina-DiMOO	0.238 ± 0.290	0.445 ± 0.312	0.634 ± 0.306
	Show-o2-7B	0.243 ± 0.288	0.467 ± 0.310	0.612 ± 0.294
	OmniGen2	0.219 ± 0.275	0.429 ± 0.305	0.653 ± 0.289
	BAGEL-7B-MoT	0.226 ± 0.278	0.434 ± 0.311	0.638 ± 0.297
	Janus-Pro-7B	0.250 ± 0.281	0.459 ± 0.306	0.604 ± 0.300
Diffusion	SD 3.5 Large	0.239 ± 0.285	0.465 ± 0.323	0.621 ± 0.299
	FLUX.2[klein]	0.234 ± 0.287	0.443 ± 0.310	0.634 ± 0.298
	SANA1.5-4.8B	0.242 ± 0.288	0.455 ± 0.315	0.614 ± 0.301
	Z-Image Turbo	0.227 ± 0.278	0.419 ± 0.312	0.649 ± 0.296
	FLUX.1[dev]	0.220 ± 0.281	0.415 ± 0.307	0.646 ± 0.298
	Kolors	0.197 ± 0.259	0.387 ± 0.291	0.648 ± 0.294

Table 16: Scores under the three prompt conditions, averaged over the 14 open-weight models evaluated in all three conditions. Every condition is scored against the same original prompt; CaRVE uses the same Qwen3-4B visual elaborations in every condition. VIEScore is the overall score divided by 10, and DSG uses dependency zero-out. ± is the standard deviation across items (549 prompts), averaged over the 14 models. Gemma augmentation leaves all four scores roughly unchanged, whereas ViPE lowers all of them.

Score	Base	Gemma-augmented	ViPE
CaRVE	0.635 ± 0.077	0.639 ± 0.077	0.606 ± 0.084
VIEScore	0.756 ± 0.086	0.766 ± 0.072	0.647 ± 0.159
TIFA	0.707 ± 0.264	0.695 ± 0.252	0.514 ± 0.307
DSG	0.706 ± 0.348	0.714 ± 0.303	0.462 ± 0.371

displaceable independently and across a wide range. ViPE was not included in the human survey, so its scores are read against CANONSCORE alone.

D MODEL LICENSES

Our study uses a mix of proprietary and openly released models. Table 17 lists every model together with the license under which it is distributed. All models were used in accordance with their respective licenses and terms of use, and solely for non-commercial research purposes.

E ADDITIONAL SHOWCASES AND DIAGRAMS

Table 17: Licenses of all models used in this study.

Model	License
Text-to-image systems	
GPT Image 2	Proprietary (OpenAI)
Gemini 3.1 Flash Image	Proprietary (Google)
Think-Then-Generate	Apache 2.0
Bagel-Zebra-CoT	Apache 2.0
BAGEL-7B-MoT	Apache 2.0
Qwen-Image	Apache 2.0
Lumina-DiMOO	Apache 2.0
Show-o2-7B	Apache 2.0
OmniGen2	Apache 2.0
Janus-Pro-7B	DeepSeek Model License
Stable Diffusion 3.5 Large	Stability AI Community License
FLUX.2 [klein] (9B)	FLUX Non-Commercial License
FLUX.1 [dev]	FLUX.1 [dev] Non-Commercial License
SANA1.5-4.8B	NVIDIA Source Code License (NSCL)
Z-Image Turbo	Apache 2.0
Kolors	Kolors Model License
Judges, evaluators, and auxiliary models	
Qwen2.5-VL-32B-Instruct	Apache 2.0
InternVL3-14B	MIT
CLIP-FlanT5-XXL (VQAScore)	MIT
BLIP-VQA (LAVIS)	BSD-3-Clause
Qwen2.5-7B-Instruct	Apache 2.0
Qwen3-4B	Apache 2.0
Gemma-3-4B-it	Gemma Terms of Use
Gemma-4-31B	Apache 2.0

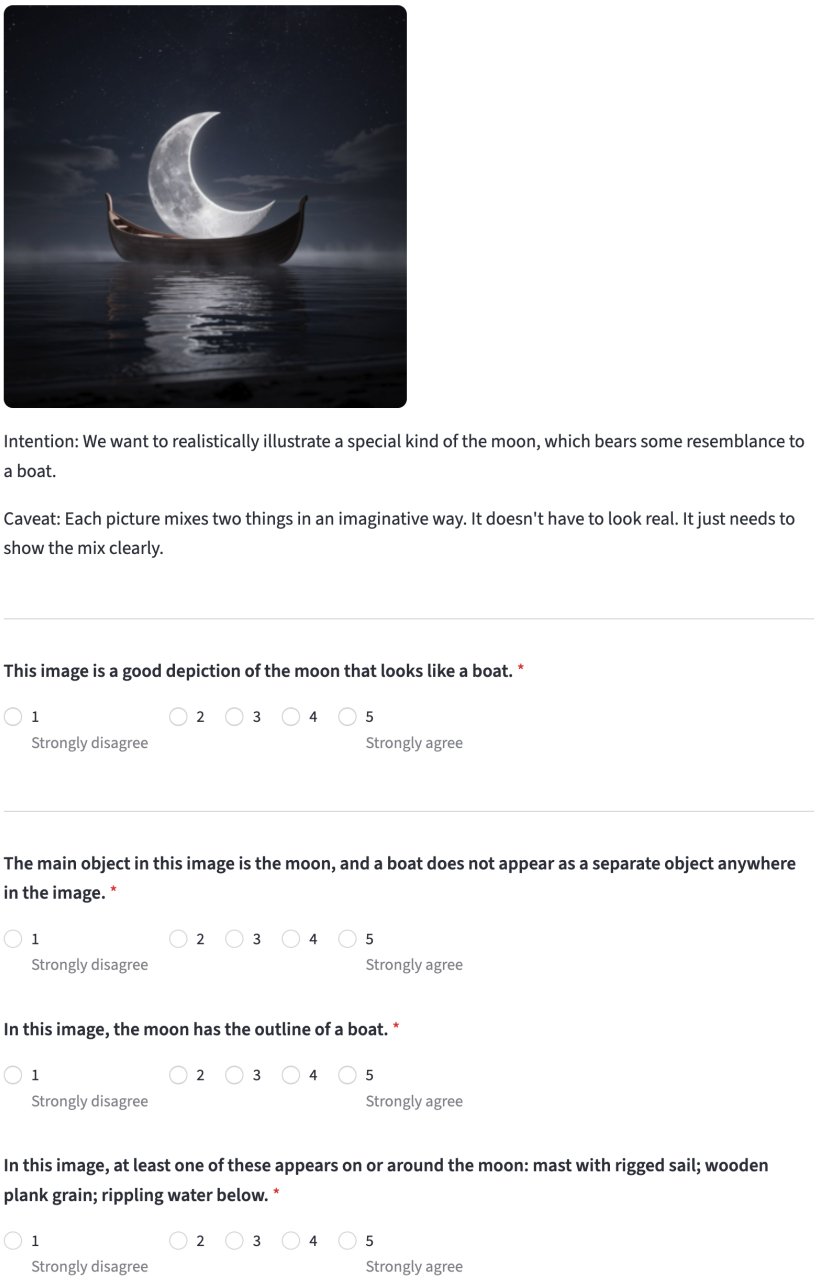


Figure 10: One rating page of the human survey, as raters saw it. The prompt is stated above the image and the four items below it, each on a five-point agreement scale. The first item asks after the generation as a whole; the remaining three correspond to the legs in order, subject anchoring, licensed-channel resemblance, and leakage of the annotated literal particulars, the last two of which name that entry's own annotations. The third and fourth items are phrased so that agreement indicates a failure, and are reverse-coded accordingly. Raters saw the base prompt only, never the condition or the paired image.

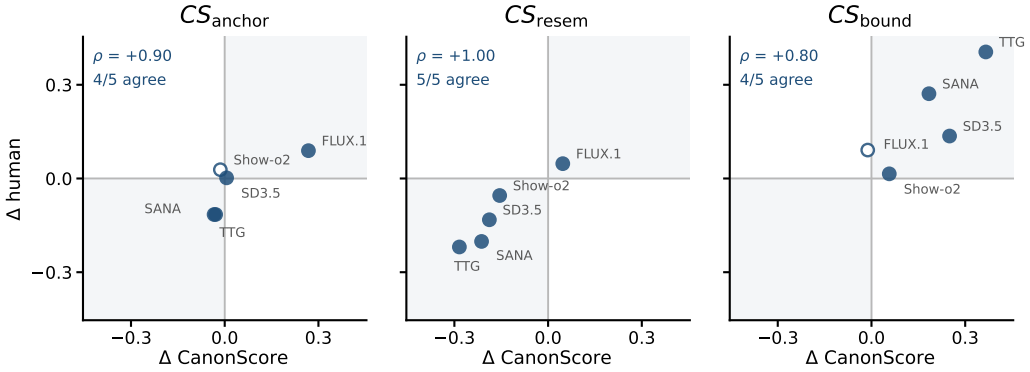


Figure 11: CANONSCORE change against human change under prompt augmentation, one panel per leg. Each point is one of the five T2I models rated in the human survey; the horizontal axis is the CANONSCORE change from the plain to the Gemma-augmented condition on the 13 relation-matched prompts, and the vertical axis is the change in the corresponding rating from 93 human raters on the same images. Scores are raw, not normalized. Shaded quadrants mark sign agreement, so a point inside one indicates that CANONSCORE reproduces the direction humans reported; an open marker is a model whose two axes disagree in sign. Spearman ρ over the five models and the sign-agreement count appear in each panel; thirteen of the fifteen model-leg cells agree overall. The CS_{anchor} points cluster near the origin, three of them within 0.03 of zero on the horizontal axis: augmentation leaves subject anchoring largely unmoved, so that panel reports agreement over a narrow range.



Figure 12: The three CANONSCORE legs for each of the sixteen T2I systems on the base VSG-BENCH prompts, one panel per architecture family and one shade per system. Scores are raw, not normalized, and every panel runs 0 to 1 on the radial axis. The characteristic shape is the same throughout: extended toward CS_{resem} and drawn in toward CS_{bound}, so that no family produces a system whose triangle is balanced. The lean is most pronounced for the two proprietary systems, which reach furthest on resemblance and least far on abstraction.

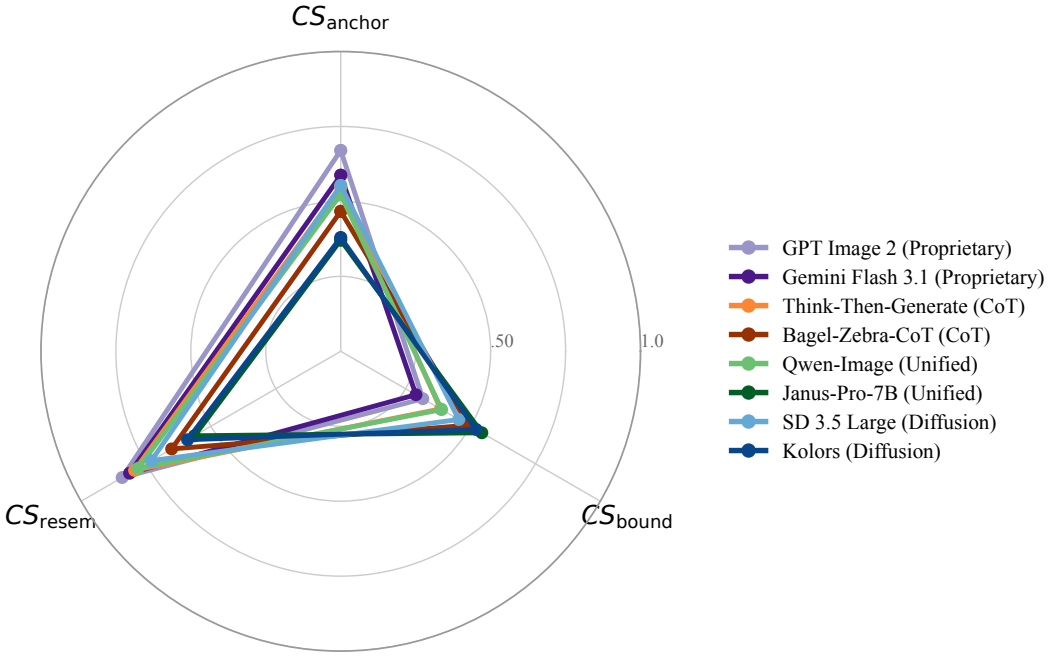


Figure 13: Two T2I systems from each category of Table 3, on one set of axes and drawn unfilled so the orderings can be read against each other. Within Unified and Diffusion the two chosen are the pair furthest apart across the three legs, so each category shows its own internal spread. Scores are raw, not normalized. The systems fan out along one direction: those reaching furthest on CS_{resem} sit closest to the center on CS_{bound}, and the ordering reverses between the two legs rather than varying independently.

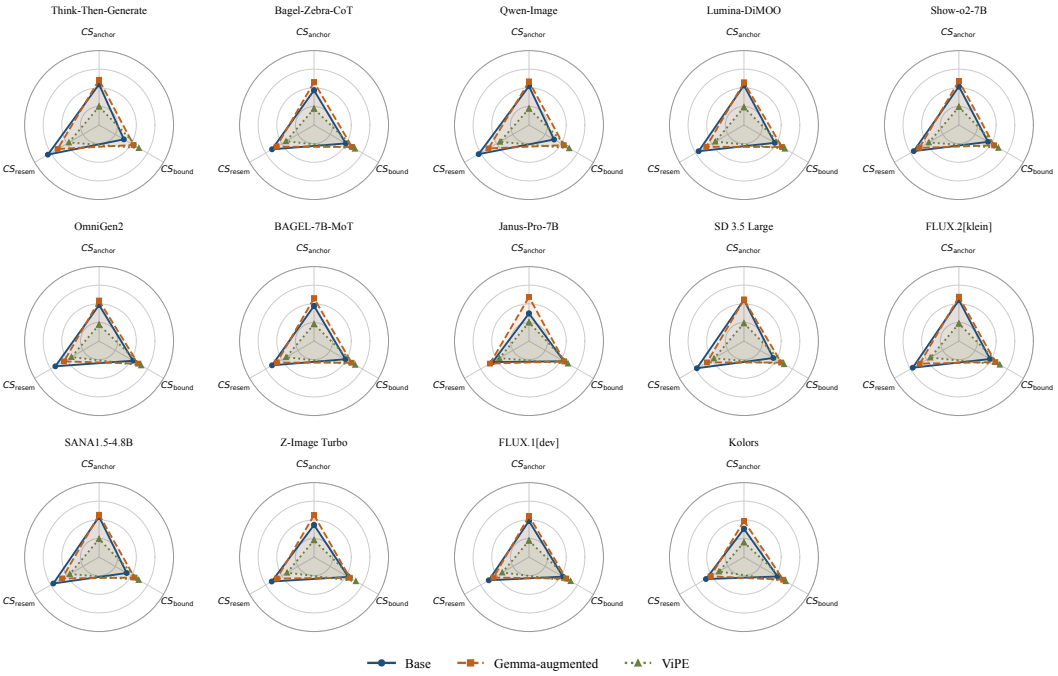


Figure 14: The same three legs for each of the fourteen open-weight T2I systems under the three prompt conditions, one panel per system. Scores are raw, not normalized, and every panel runs 0 to 1 on the radial axis. Gemma augmentation rotates the triangle rather than growing it, gaining on CS_{bound} against CS_{resem} in the same direction for nearly every system. ViPE moves further along that same rotation while collapsing CS_{anchor}, which is visible as the flattened top vertex in every panel.

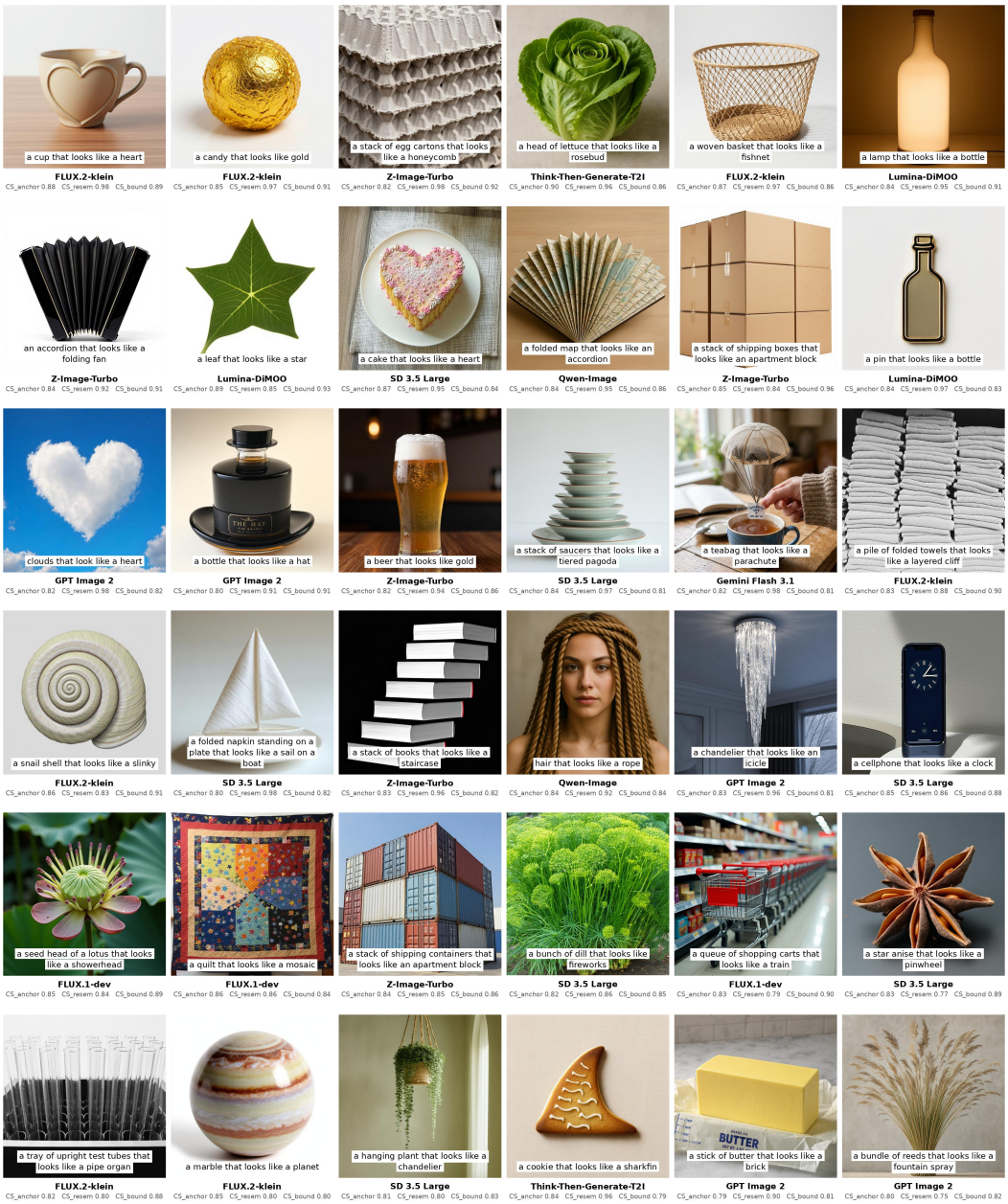


Figure 15: Generations scoring high on all three CANONSCORE legs, with each image's prompt, generating T2I system, and leg scores.