THE SCENE, NOT THE SHOT: MEASURING INFERRED COMPOSITION IN TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Photographers stage emotion geometrically: awe by dwarfing a figure against a vast scene, dominance by tilting the camera upward. These conventions are rarely spelled out—no photographer is told to shrink the subject—yet they are encoded in the photographs text-to-image models are trained on. Models follow explicit framing instructions well; what they do by default, when the framing is not specified, has not been measured. We ask whether models infer the composition an intent calls for, through a benchmark of six axes—awe, romance, social group, power, vulnerability, and relative scale—whose prompts name the intent but never how to frame it. Each axis extracts a geometric quantity from the image and scores it against human photographic practice or an established convention, with bootstrap uncertainty throughout. The reference is a range of human practice, not a single correct answer or a recipe for the emotion.

Across twelve recent models we find one consistent split, holding within every model: models get the scene right but not the camera. What lies within the scene—the depth of an awe landscape, the spacing of couples and groups—stays near human practice. How the camera captures it does not: the awe subject is framed too large or too uniformly, couples are shot from too far, and although power calls for a low camera angle and vulnerability a high one, every model keeps the camera at eye level in at least seven of ten generations, tilting it barely more often than when no intent is named. Relative scale compresses: small things render too large, vast things too small. No model dominates across axes. Models stage the scene; they do not shoot it.

1 INTRODUCTION

Text-to-image (T2I) models have shown striking progress (Labs, 2025; Wu et al., 2025; Cai et al., 2025b; Seedream et al., 2025), as reflected in the increasingly sophisticated benchmarks built to evaluate them—from visual quality and prompt-following (Ghosh et al., 2023; Huang et al., 2023) to reasoning, commonsense, and physical or world knowledge (Qiu et al., 2026; Fu et al., 2024; Niu et al., 2025). Each wave has pushed models to be more faithful to the prompt and more correct about the world—the right objects and attributes, respecting physical and commonsense knowledge.

Yet a primary and fast-growing use of these models is creative—social media content, advertising, cover art, film pre-visualization (Hartmann et al., 2025; Møller et al., 2026; Tsiavos & Kitsios, 2025). Here, an image's value lies less in what it contains than in what it *communicates*, and the same content can be arranged to mean very different things–an emotion, a relation, a physical fact: a vast landscape dwarfing a lone figure conveys awe (Piff et al., 2015), a low camera angle makes a subject loom with power (Giessner et al., 2011; Kraft, 1987; Sevenants & d'Ydewalle, 2006), correct proportions make a scene read as physically real. This is the work of *scene composition*–turning content into meaning by arranging the subject and the frame–yet, despite being a central use of these models, it has received very little attention in evaluation.

The reason is not hard to see: composition is a creative act, and creative choices look subjective. But directors and photographers do not compose at random. They follow recurrent conventions, many of them backed by perception research (Arnheim, 1974; Kress & van Leeuwen, 2021)–the angle–dominance effect above is not a stylistic preference but an experimentally established response—and these conventions are *geometric*: scale, viewpoint, vertical placement, inter-subject distance,

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: **Qualitative examples of four of the proposed evaluation axes.**

physical proportion are all quantities one can measure. What looks subjective can thus be measured in part: not the emotion itself, but how photographers arrange the frame to convey it.

We focus on this *geometric* layer of composition because it can be measured factually: each axis is measured by a physical quantity of the scene—a camera pitch, a depth, an inter-subject distance—that can be estimated, validated against human annotation, and checked. No model in our pipeline judges whether an image conveys the intent; each only measures where things are and how large they appear. To our knowledge, no existing benchmark measures this layer for generation. Spatial benchmarks for text-to-image models check relations stated in the prompt ("a dog to the left of a cat"), and cinematography benchmarks test whether multimodal models can *recognize* techniques such as a low-angle shot in existing film and video (Wang et al., 2025; Liu et al., 2025), not whether a generator knows when to use one. How a model *composes* the frame when given only an emotional or narrative intent–without being told the framing–has not been measured.

The test is one of *inference*. The prompt "a powerful king" specifies no camera angle, yet a skilled photographer would choose one—a low angle, which makes a subject read as more dominant (Mandell & Shaw, 1973; Sevenants & d'Ydewalle, 2006). A model that renders the king at eye level has satisfied the letter of the prompt; whether it has read its meaning is what we want to know. We do *not* claim a unique ground truth: a single eye-level shot may be a deliberate choice. Instead, each axis follows a convention from perception research or photographic practice: the literature gives the direction (for awe, a small figure in a deep scene), human photographs give the magnitude (how small, and how much this varies across photographs), and we measure where the model lands. A deviation is a fact, not a verdict—but a model that *never* leans in the conventional direction, across many prompts, reveals not a style but an insensitivity.

Contributions. We introduce GEOCOMP, a benchmark that measures whether T2I models compose the geometric layer of a scene the way human authors do to convey intent (Fig. 1 showcases four axes). It spans six axes in two registers, *expressive* composition and *physical* coherence:

- **Awe** (expressive): does the subject shrink against a vast scene, as photographers frame it?
- **Romance and social group** (expressive): are couples framed close, and groups spaced as humans stage them?
- **Power and vulnerability** (expressive): does the camera drop or rise to match the intent?
- **Relative scale** (physical): do objects keep real-world proportions—is a person rendered far taller than a soda can?

Section 3 details each axis. Across twelve recent models, we find one consistent split: models stage the scene as human practice does; they do not shoot it.

2 RELATED WORK

From prompt fidelity to expressive intent. Most T2I benchmarks share one structure: the property to produce is named in the prompt, and the test verifies compliance. This holds across attribute binding, counting, and spatial relations (Ghosh et al., 2023; Huang et al., 2023)—where a relation such as "a dog to the left of a cat" is stated and then checked from detected boxes—and extends to physical, commonsense, and world knowledge (Qiu et al., 2026; Fu et al., 2024; Niu

et al., 2025). Relative size sits in this commonsense family: Liu et al. (Liu et al., 2022) probe it *through* image synthesis, showing early models order object scales correctly; whether modern models *render* scales metrically—our relative-scale axis—has not been measured. These evaluations measure whether objects sit where the prompt says, not how the frame is composed around them: the same named relation can hold in a wide shot or a close-up, from above or below. A smaller body of work looks past fidelity toward what an image *expresses*—combining concepts into novel visual hybrids (Li et al., 2024; Yang et al., 2026), conveying an emotion through semantic content, color, and style (Yang et al., 2024; Dang et al., 2025), or folding affect, aesthetics, and storytelling into broad multi-dimensional suites (Aghazadeh & Kovashka, 2025; Li et al., 2026; Fang et al., 2025; Bakr et al., 2023). Here the expressive side has largely meant generating *abstract* or affective imagery, not composing a concrete scene to carry an intent; and whatever the target, these evaluations score the outcome—an MLLM rates a rubric, an emotion classifier reads the image—attributing any effect to global appearance cues such as color, brightness, and semantic content. None asks *which* compositional choices produced the effect. We measure those choices directly: the geometric levers of composition read off the image itself.

From understanding to generation. A separate line of work does target cinematographic language, but mostly on the *understanding* side: CineTechBench (Wang et al., 2025) and ShotBench (Liu et al., 2025) ask whether multimodal models can recognize shot scale, camera angle, or lighting in existing footage, and CameraBench (Lin et al., 2025b) probes camera-motion understanding in video. These assess how well a model *reads* the language of film, not whether it can *write* it: recognizing a low-angle shot in a still is not the same as choosing one when generating an image.

On the generation side, camera geometry is treated as an *explicit* input throughout: PreciseCam (Bernal-Berdun et al., 2025) conditions still-image generation on pitch, roll, and field of view, CameraCtrl and MotionCtrl (He et al., 2025; Wang et al., 2024) steer camera trajectories in video, and Qwen-Image Bench (Li et al., 2026)—the closest evaluation to ours—scores whether a named framing ("a low-angle shot") is obeyed, a capability its low-variance compliance scores suggest is largely in hand. All of it, control and evaluation alike, presupposes a caller who already knows the framing they want. We remove that premise: our prompts never name the geometry, and given only an emotional or narrative intent—"a powerful figure"—we measure geometrically whether the model produces the composition that intent implies. Following an order and sensing what an emotion asks for are different abilities, and mastering the first need not grant the second. It is this quieter, inferential skill that we measure, rooting each axis in a perceptual or social principle and the geometric signature it leaves in an image.

3 METHODOLOGY

Each axis has three parts: a convention that gives the expected direction, a geometric quantity measured on the image, and a target—the range of human photographs for the expressive axes, and for the others a threshold from the experimental literature or from real-world dimensions. More extreme is never better: the target is what humans do, not the far end of the direction.

3.1 AWE: THE VANISHING SUBJECT

Psychologists describe awe as an emotional response to "perceptually vast stimuli that transcend current frames of reference" (Piff et al., 2015), stimuli so far beyond ordinary experience that they demand a mental "need for accommodation" (Keltner & Haidt, 2003) and evoke a diminished, "small self" (Piff et al., 2015). In photographs, its signature is one of proportion: a lone figure set against an overwhelming environment, the frame itself enacting the smallness of the self before the vastness of the world. We capture this signature with two geometric measurements. Let M_s be the subject's segmentation mask and A the image area. *Subject occupancy*,

$$o = \frac{|M_s|}{A},$$

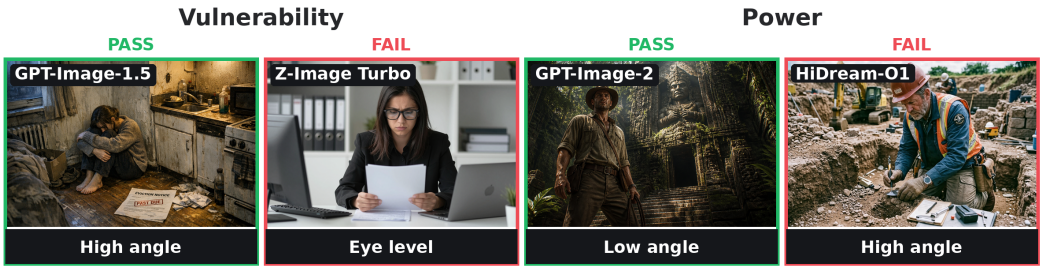

Figure 2: **Qualitative examples of the Vulnerability and Power evaluations.** Passing generations use a high-angle view for Vulnerability and a low-angle view for Power.

measures how much of the frame the subject claims: a small o[1] is a figure lost in its surroundings. Vastness is also a deep scene, so from a monocular depth map D we measure the *far-scene depth*

$$z_{\text{far}} \;=\; P_{98}(\{D(p) \mid p \in \Omega \setminus M_{\text{sky}},\; D(p) > 0\}),$$

the 98th percentile of depth over the visible, non-sky scene (Ω is the image domain and M_{sky} the sky mask; robustness rationale in App. B.1). A large z_{far} is a scene that falls away far behind the figure. The two measurements come apart—a model may place a small subject in a shallow scene, or render a deep scene dominated by a looming subject—and reporting both reveals which half of the signature a model has grasped. Neither quantity has a "more is better" direction. A subject shrunk to a few pixels has not become more awe-inspiring; it has simply vanished, and with it the small-self contrast the convention rests on. The target is the range human photographers use, taken from the *awe*-tagged subset of EmoSet (Yang et al., 2023).

3.2 Romance: The Closing Distance

Romance is written in proximity. Hall's proxemics keeps the innermost *intimate* zone—within roughly half a metre of the body, the range of "unmistakable involvement"—for our closest relations (Hall, 1966). We do not measure intimacy itself but its geometric trace: how close the two figures appear in the image. A photograph composes this closeness twice: it brings the two people together, and it brings the camera in on them, shot scale being a primary lever of a viewer's engagement with on-screen characters (Rooney & Bálint, 2018). The two come apart—a model may render a large pair held at a distance, or a small pair pressed together—so we measure both, and reporting both shows which half a model has grasped.

The first measurement, *subject occupancy*, is taken from the two person masks M_1 and M_2 with A being the area of the image:

$$o \;=\; \frac{|M_1 \cup M_2|}{A},$$

the union ensuring that overlap is not counted twice. The second, proximity, cannot use body size as its unit: tight framing truncates the body and embracing figures occlude it, so a body-normalized gap would be biased by the very closeness we measure. Heads escape both. Given each subject's head centroid c_i and head mask area m_i, with $\sqrt{m_i}$ as the characteristic head size, the *normalized interpersonal distance* is

$$d \;=\; \frac{\|c_1 - c_2\|_2}{\frac{1}{2}\left(\sqrt{m_1} + \sqrt{m_2}\right)},$$

the head-to-head distance in head-size units, the sizes averaged for stability under foreshortening. This follows Gallagher & Chen (2009), who normalize inter-person distances in group photographs by face size, calibrated on anthropometric proportions (Farkas & Munro, 1987).

3.3 Social Group: The Shape of Togetherness

A gathering is legible in its spacing. Hall's proxemics assigns each social register its distance—personal for friends, social for acquaintances (Hall, 1966)—and in group photographs, Gallagher &

[1] A generation with no detected or a frame-truncated subject yields no o and is excluded from the model's distribution (1.4% of generations; see App. B.1). It is not scored as $o = 0$: a missing subject is a prompt-following failure, not a composition.

Chen (2009) show that who stands close to whom reflects who belongs with whom. In photographs, the signature of a social group is therefore cohesion: people spaced as companions, not as strangers who happen to share the frame.

Spacing needs a relative unit: two figures a hundred pixels apart are touching in a wide shot and separated in a close-up. As for romance, the body is an unreliable unit, and more so in groups, where tight shots crop bodies and people occlude one another. As in the romance case, heads escape both: photographers arrange a group so that every face stays visible. Given each participating person's head centroid c_i and mask area m_i, with $\sqrt{m_i}$ as the characteristic head size, we measure the distance to the nearest other head in head-size units:

$$d_i \;=\; \frac{\|c_i - c_j\|_2}{\frac{1}{2}\big(\sqrt{m_i} + \sqrt{m_j}\big)}, \qquad j = \arg\min_{k \neq i} \|c_i - c_k\|_2,$$

the two sizes averaged so the measure remains stable when the pair straddles depth layers. The image-level score is the median over members—robust to a single stray detection—$d = \mathrm{median}_i(d_i)$: low d is a cohesive group, high d individuals who merely share a frame. Neither extreme is the target: $d \to 0$ signals overlapping or fused masks—a failed decomposition, not a closer group—while beyond a few head widths a set of people ceases to read as a group at all.

3.4 Power / Vulnerability: the angle of dominance

Vertical camera angle is among the most robustly replicated cues in visual communication research: five decades of experiments confirm that a low angle—the camera looking up—confers visual authority and dominance on the subject, while a high angle diminishes it (Kraft, 1987; Mandell & Shaw, 1973; Sevenants & d'Ydewalle, 2006). Giessner et al. (2011) corroborate this across four archival studies: "variations in vertical picture angle cause the subject to appear more powerful when depicted from below and less powerful when depicted from above." Their photographic signatures are therefore the two poles of a single signed parameter: camera pitch relative to the subject. We estimate this pitch θ with Puffin (Liao et al., 2025), a multimodal camera-understanding model that returns the global camera orientation. A generation passes when $\theta > +9^\circ$ (low angle) for power and $\theta < -9^\circ$ (high angle) for vulnerability (See Fig. 2), a threshold calibrated in App. C.1.

3.5 Relative Scale: the size of the world

The preceding axes test whether a model turns an emotional or social intent into compositional convention. Relative scale tests a more literal coherence: two named entities in one image must respect their real-world size relationship—a person is smaller than a lighthouse and larger than a key. Apparent pixel size conflates real size with camera distance, so we correct it with depth. Each entity is localized with a GroundingDINO (Liu et al., 2023) box of apparent size e_i, and its depth D_i is estimated with Depth Anything 3 (Lin et al., 2025a). Under a pinhole model (Hartley & Zisserman, 2003), the size ratio of the two entities is

$$R \;=\; \frac{e_s}{e_r} \cdot \frac{D_s}{D_r},$$

in which the focal length and any global depth scale cancel. A generation passes when R falls within the interval spanned by the pair's real-world heights, widened by $\pm 15\%$; generation, detection, or depth failures count as failures (pairs and full protocol in App. B.5).

4 Experimental Setup

4.1 Prompt construction

Five of the six axes share a two-part template: a *neutral sentence*–a subject in an everyday context, with no photographic or emotional vocabulary–followed by a *suffix* naming the target construct (App. A.2), e.g. "a lone hiker pausing at the edge of the Grand Canyon at sunrise, *an awe-inspiring scene*." Neither part names the geometry we score: nothing says the hiker should look small, so the composition cannot come from instruction-following. Models are scored on the full prompt (L2); the neutral sentence alone (L1) serves as a no-intent baseline, and alternative suffixes are tested

in App. D.2 and D.3. *Relative scale*, whose target is a fact about the world rather than an intent, instead pairs two entities of markedly different real-world size, across four size regimes and three depth arrangements. Additional details are in App. A.2.

4.2 Reference data and geometric extraction

Awe is referenced against EmoSet (Yang et al., 2023), romance and social group against Open Images V7 (LLC, 2020), each curated into a per-axis set of human photographs. Power and vulnerability need no reference set: their direction comes from the angle–dominance literature (Giessner et al., 2011; Kraft, 1987). Relative scale relies on documented real-world dimensions. Geometric quantities are extracted with four off-the-shelf perception models: SAM3 (Carion et al., 2025) for instance masks, GroundingDINO (Liu et al., 2023) for named-entity boxes, Depth Anything 3 (Lin et al., 2025a) for depth, and Puffin (Liao et al., 2025) for camera pitch.

4.3 Evaluated models

We evaluate twelve recent text-to-image systems spanning proprietary and open-weight models: OpenAI's GPT-Image-1/1.5/2 (OpenAI, 2026), Google's Nano Banana 2 (Google DeepMind, 2026), FLUX.2 Klein 4B (Labs, 2025), Z-Image Turbo (Cai et al., 2025a), Qwen-Image (Wu et al., 2025), ERNIE-Image Turbo (Liu et al., 2026), HiDream-I1/O1 (Cai et al., 2025b), HunyuanImage-3.0 (Cao et al., 2025), and Cosmos3 Super T2I (NVIDIA, 2026). Generation and default sampling settings for all models are given in App. A.1.

4.4 Scoring protocol

The six axes fall into two families that ask different questions of a model.

Distributional axes (awe, romance, social group). For the expressive axes there is no single correct composition: human photographers stage awe, romance, or a gathering across a whole *range* of framings, and that range is itself the signature of the convention. Matching only a summary statistic (e.g. the median) would reward a model that collapses to a single average framing; the reference must be the distribution. We therefore use the Wasserstein-1 distance W_1 (Peyré & Cuturi, 2019) between model and human distributions, which compares them as wholes, normalized by the human interquartile range (IQR): the dimensionless W_1/IQR is 0 at a perfect match and 1 at a gap on the order of the human spread. **Lower is better**, with departures in either direction—too extreme or too narrow—penalized alike. Awe and romance each measure two quantities (subject size and scene depth; subject size and interpersonal distance), and their axis score is the mean of the two W_1/IQR values. On the human reference the two quantities are at most weakly correlated (Spearman $\rho = -0.056$ for awe, -0.141 for romance; App. B.1 and B.2), so averaging does not count one quantity twice.

Directional axes (power, vulnerability, relative scale). The remaining axes have a reference direction–a documented convention for power and vulnerability, a physical fact for relative scale–so each generation passes or fails, and we report the *pass rate* (**higher is better**). Power passes when the camera pitch exceeds $+9^\circ$ (low angle), vulnerability when it falls below -9° (high angle); the threshold is calibrated against human annotation and validated on a model outside the evaluated panel (App. C.1). Relative scale passes when the depth-corrected size ratio falls within the pair's real-world interval, widened by $\pm 15\%$ (Sec. 3.5).

Cross-axis comparison and uncertainty. A W_1/IQR and a pass rate are not commensurable, so models are compared across axes by *average rank*: each axis ranks the twelve models, and the ranks are averaged with equal weight. All scores carry 95% confidence intervals—a two-level bootstrap over prompts and reference images for distributional axes (1,000 replicates), Wilson intervals (Wilson, 1927) for pass rates—and all rankings, including the average rank, come with bootstrap rank-stability analyses. The main tables report point estimates; intervals are in App. D.1, and the text notes where they do not support a comparison.

Model	Distance to human ref. (W_1/IQR, ↓)							Pass rate (↑)			Avg. Rank
	Awe			Romance			Social Group	Power	Vulnerability	Relative Scale	
	Subj. Ratio	Depth	Avg.	Subj. Ratio	Inter. Distance	Avg.					
Cosmos3 Super	0.608	**0.064**	**0.336**	0.749	2.841	1.795	0.274	0.153	**0.060**	0.357	**4.92**
GPT Image 2	0.726	0.319	0.522	**0.355**	**0.283**	**0.319**	0.531	**0.280**	0.021	0.456	5.08
GPT Image 1	0.924	0.141	0.532	0.494	0.911	0.703	0.279	0.067	0.041	0.456	5.17
Z-Image Turbo	0.629	0.237	0.433	0.735	3.718	2.227	0.464	0.087	0.047	**0.460**	5.25
FLUX.2 Klein 4B	**0.511**	0.501	0.506	0.507	1.760	1.134	0.356	0.073	0.033	0.438	5.67
GPT Image 1.5	0.821	0.562	0.691	0.592	0.402	0.497	0.275	0.127	0.021	0.408	6.17
Qwen Image	0.597	0.423	0.510	0.840	0.450	0.645	0.502	0.113	0.013	0.434	6.42
HunyuanImage 3.0	0.599	0.399	0.499	0.562	2.324	1.443	0.292	0.087	0.040	0.404	6.50
Nano Banana 2	0.620	0.455	0.537	0.937	1.002	0.969	0.385	0.013	0.021	0.426	7.67
ERNIE Image	1.398	0.260	0.829	0.795	4.951	2.873	0.238	0.000	**0.060**	0.412	7.67
HiDream-O1	1.675	0.141	0.908	0.694	2.148	1.421	**0.224**	0.007	0.013	0.412	8.50
HiDream I1 Full	0.679	0.403	0.541	0.928	1.619	1.273	0.649	0.067	0.027	0.404	9.00

Table 1: **Results across the six axes.** W_1/IQR to the human reference (↓) or pass rate (↑); sorted by average rank. **Red**: best per column; orange: second best.

5 Experimental Results

5.1 No model dominates across axes

Table 1 summarizes all six axes. Ranked by average rank (Sec. 4.4), the top is barely a podium: Cosmos3 finishes first, with GPT Image 1/2 and Z-Image practically tied just behind—a hierarchy so compressed that no model finishes first in more than a third of bootstrap replicates (App. D.1). The crowded top itself makes the point: Cosmos3 is closest to human practice on awe yet places tenth on romance and last on relative scale; GPT Image 2, just behind, leads romance and power in every replicate yet sits eleventh on social group. A single-axis benchmark would rank these models in almost any order: the axes measure distinct abilities, learned—or not—independently.

Where models approach human practice and where they miss it, however, is not arbitrary: what the scene contains follows human practice; how the camera frames it does not.

5.2 What the camera looks at: scenes, bodies, spacings

Three of our measurements probe what the generated scene contains, independently of where it is seen from—and all three align closely with human practice.

Scene depth. Under awe prompts, all twelve depth medians sit inside the human band (Fig. 4, right): medians span 48–74,m against 50,m for humans—nine models opening the landscape slightly *deeper* than the reference—and per-model W_1/IQR stays between 0.06 and 0.56 (Table 1). The subject-size component of the *same* axis tells the opposite story (Sec. 5.3): every model scores better on depth than on subject size, and for all but one by a margin well beyond sampling noise (App. D.1).

Interpersonal distance. Physical closeness is learned—by half the models. Under romance prompts, the median couple distance lies within 0.98–1.13× the human median for six models, and at 1.23–1.83× for the remaining six, with no model in between (Fig. 3, right). In contrast, the framing of these same images fails across the panel (Fig. 3, left): when models depart from human practice on romance, the departure lies in the shot, not in the couple

Social Group. Group spacing discriminates least among models: all twelve fall within W_1/IQR 0.22–0.65, the narrowest range of the three distributional axes. What remains is a shift, not a loss of spread: eleven of twelve models place people slightly farther apart than human photographs do (median spacing 1.56–1.83 vs. 1.44, i.e. 8–27% wider, See Fig. 13), while their interquartile ranges stay close to the human one (0.42–0.82 vs. 0.56). It is also the least demanding test in the benchmark: spacing is visible in any photograph of a group, and our head-normalized measure is invariant to camera distance and framing—the axis requires no viewpoint decision at all.

5.3 Where the camera looks from: the frame and the angle

For the camera, the same generations tell the opposite story. Three quantities depend on where the viewpoint is placed—how much of the frame the subject occupies (awe), how tightly the couple is

framed (romance), and whether the view tilts (power, vulnerability)—and none approaches human practice.

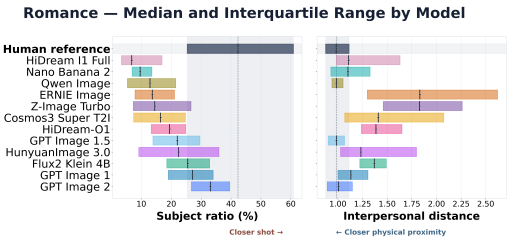

Figure 3: **Romance:** Bars show interquartile ranges and black ticks show medians and dashed lines mark the human medians.

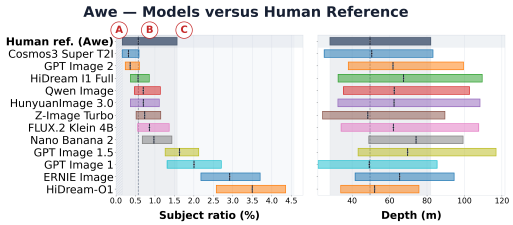

Figure 4: **Awe:** Bars show interquartile ranges and black ticks show medians and dashed lines mark the human medians.

Subject size under awe. No model reproduces the human distribution of subject sizes, and they miss it in two ways (Fig. 4, left). Four models render the figure too large: their medians fall above the human band (1.6–3.5%, i.e. 3–6× the human median of 0.57%; zone C). The other eight have medians inside the band (zone B), but their interquartile ranges cover at most roughly half the human spread: they produce one typical size where photographers vary. And the gap is one-sided (zone A): a quarter of human awe photographs place the figure below 0.16% of the frame—nearly lost in the scene—a register only Cosmos3 reproduces (25.2% of its generations, against the human 25%); no other model exceeds 10%. Framing the figure is the hard part; building the scene is not.

Couple framing under romance. The same asymmetry structures the romance axis. Half the models place the two people as close to each other as human photographs do (median interpersonal distance within 0.98–1.13× the human one; Fig. 3, right), yet none frames the couple as humans do: the human reference devotes a median 42% of the frame to the couple, and the highest model median—GPT Image 2 at 33%–stays well below it, while nine of twelve fall under even the human lower quartile of 25%(Fig. 3, left). The couple stands at the right distance; the camera stands too far.

Camera angle under power and vulnerability. The convention is unambiguous—power is staged from below, vulnerability from above (Giessner et al., 2011)—yet models keep the camera at eye level in around 90% of power generations and 96% of vulnerability generations on average (Tab. 15); even the best model, GPT Image 2, tilts upward in only 28% of generations. And not all of these low angles come from the intent: without the suffix, the same prompts already yield low-angle shots 4% of the time (App. D.2). Naming power raises the median model's low-angle rate by only 4 points (from 4% to 8%) and the increase is significant for only seven of twelve models (Tab. 21). Vulnerability fares worse: no model exceeds 6% (Tab. 15), and no ranking on the axis is statistically meaningful (Appendix D.1). Models do not tilt the camera the wrong way; they do not tilt it at all.

5.4 Neither scene nor shot: knowing how large things are

The last axis probes neither the scene's arrangement nor the camera's placement, but the knowledge both depend on: how large things are relative to one another. Models are asked for two entities "at realistic relative physical scale"—the sizes themselves are never stated—across four regimes of true size ratio, from moderate (person vs. door) to micro (insect vs. hand).

Relative scale. Success tracks the regime, not the model. Moderate ratios are rendered well (62–93% pass), but performance declines toward both ends of the spectrum: 45–70% in the large regime, 12–51% in the extreme one, and 0–25% in the micro one, where two models never succeed at all (Tab. 18). Overall, even the best model resolves fewer than half its pairs (46%). The errors have a direction: rendered ratios are pulled toward unity–in the micro regime the small entity comes out about seven times too large, in the extreme regime the large one about twice too small–and all twelve models follow the same monotone pattern, with a stronger bias on small entities than on large ones (Fig. 6 and Fig. 5). No model is simply better at scale: each regime has a different leader–GPT Image 2 on micro, Z-Image on large, HunyuanImage on extreme, HiDream-O1 on moderate–and each falls to the bottom half of at least one other regime.

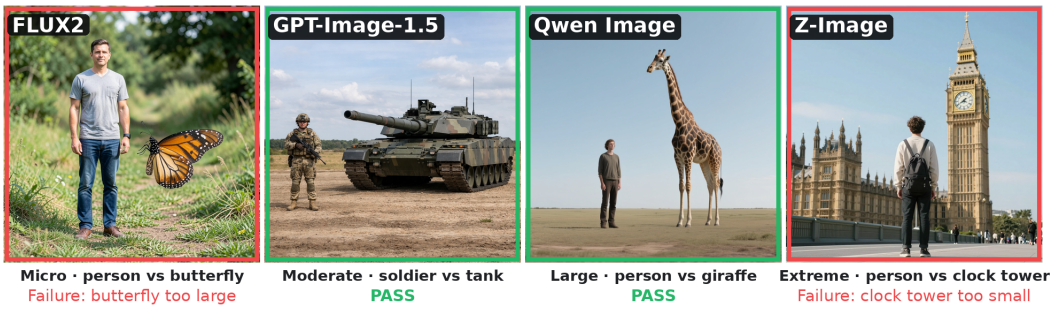

Figure 5: Boxplot showing distribution of subject ratio

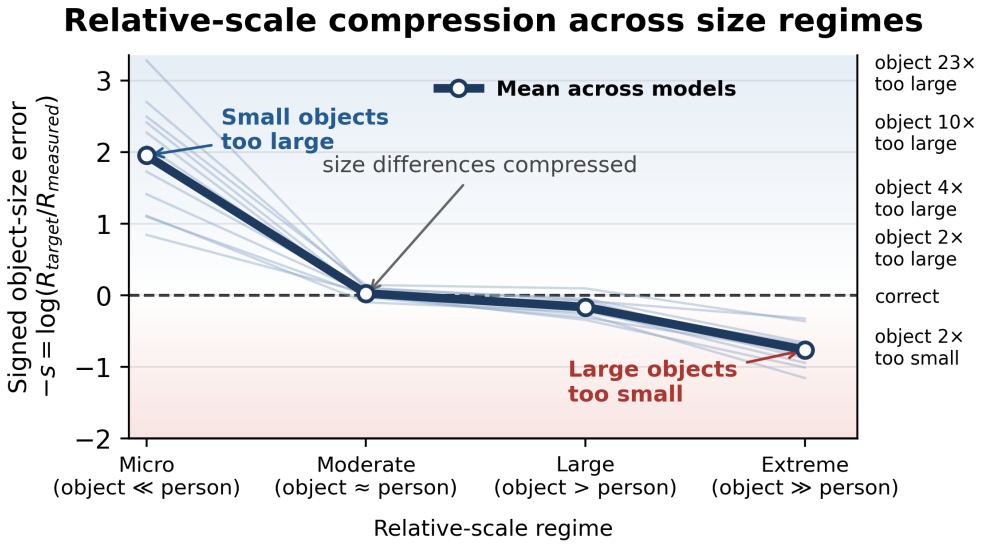

Figure 6: **Signed size error by regime.** $-s = \ln(R_{\text{target}}/R_{\text{measured}})$, with R the person-to-object size ratio; positive values mean the object is rendered too large. Thin lines are individual models, the thick line their mean.

Summary. Across six axes, the pattern is consistent: what lies within the scene—its depth, the distance between two people, the spacing of a group—stays close to human practice for most models, while how the camera captures it—how large the subject is framed, how tight the couple, whether the camera tilts—does not for any. Relative scale fails differently: success depends on the size gap between the two objects more than on the model.

6 Conclusion

We introduced GeoComp, a benchmark of six axes that asks whether text-to-image models infer the composition an emotional or relational intent calls for, when the framing is never named. Each axis compares a geometric quantity measured on the image with human photographic practice or an established convention, with bootstrap uncertainty throughout.

Across twelve recent models, one split holds within every model: models get the scene right but not the camera. Landscapes reach human depths, and couples and groups stand near human distances; but awe subjects are framed too large or all at much the same size, couples are shot from too far, and the camera stays at eye level, tilting barely more often when power or vulnerability is named than when nothing is. When composition requires knowing real-world proportions, rendered sizes are pulled toward each other, and success depends more on the size gap between the two objects than on the model. No model dominates across axes: the leader on one axis sits near the bottom of another. Current models seem to have learned what photographs contain, but not how photographers place the camera to convey an intent. Until they do, the framing that gives an image its meaning must still be spelled out in the prompt.

LIMITATIONS

Our references are built from two public image collections—EmoSet for awe and the emotion contrast sets, Open Images for romance and social group—and inherit their scope: predominantly Western, contemporary web photography, which we argue is the right yardstick for web-trained models but is not a universal norm (Hall, 1966). Robustness to alternative suffix phrasings is tested on power and awe (App. D.3); since each additional phrasing multiplies generation cost across all twelve models, the other axes use a single suffix each. Each axis measures one common way photographers stage its intent (awe, for instance, as a figure before a vast scene), not all of them. Finally, the benchmark measures what models do, not why: explaining the missing concept-to-camera link is a natural next step.

AI USE STATEMENT

We used a generative AI assistant to draft and edit parts of the manuscript for clarity and readability, to create and edit analysis and plotting code, to double-check the literature survey conducted by the authors, and to verify and format references. The research question, methodology, experimental design, and prompt design were developed by the authors. We used generative AI tools only to generate alternative prompt phrasings that were subsequently reviewed, selected, and evaluated by the authors. We did not use generative AI tools to generate synthetic datasets. We take responsibility for the final content of this work.

ETHICS STATEMENT

Reference images come from public datasets with their own licenses (EmoSet, Open Images V7) and are redistributed only as image identifiers and derived measurements, never as image files. Generated images shown in the paper are outputs of the evaluated models, labeled as such, and were screened for identifiable real people and brands. The benchmark measures conformity to conventions of a particular photographic culture—predominantly Western, contemporary web photography—which is the right yardstick for web-trained models but not a universal norm; we discuss this scope in Limitations and do not claim that departures from it are failures of expression. All evaluated and perception models were used under their respective licenses.

REPRODUCIBILITY STATEMENT

Prompt construction, templates, and lexical exclusions are described in Appendix A.2; the full prompt sets are released with the benchmark. Human references are built from public datasets, EmoSet (Yang et al., 2023) and Open Images V7 (LLC, 2020), and each axis's appendix section (Appendix B.1–B.5) documents its reference filtering, quality-control thresholds, and validation; the retained image IDs and excluded generations are released. All measurements use fixed off-the-shelf perception models—SAM3, GroundingDINO, Depth Anything-3 and Puffin—whose details are given in Section 4 and Appendix C.2, with the camera-pitch threshold calibration in Appendix C.1. Generation settings for every evaluated model are listed in Table 2, and the bootstrap protocol behind all confidence intervals and rank-stability analyses in Appendix D.1. Scoring code is provided in the supplementary material.

REFERENCES

Aysan Aghazadeh and Adriana Kovashka. Cap: Evaluation of persuasive and creative image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16970–16980, 2025.

Rudolf Arnheim. *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press, Berkeley, CA, 1974. ISBN 9780520026131.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20041–20053, 2023.

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. PreciseCam: Precise camera control for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2724–2733, 2025.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025a.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025b.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15218–15228, 2025.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025.

Leslie G. Farkas and Ian R. Munro. *Anthropometric Facial Proportions in Medicine*. Charles C. Thomas, Springfield, Illinois, USA, 1987.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024.

Andrew C. Gallagher and Tsuhan Chen. Understanding images of groups of people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 256–263, Miami, Florida, USA, 2009.

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.

Steffen R. Giessner, Michelle K. Ryan, Thomas W. Schubert, and Niels van Quaquebeke. The power of pictures: Vertical picture angles in power pictures. *Media Psychology*, 14(4):442–464, 2011.

Google DeepMind. Nano banana 2: Google's latest ai image generation model (gemini 3.1 flash image). https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2026. Released February 26, 2026. Accessed July 2026.

Edward T. Hall. *The Hidden Dimension*. Doubleday & Company, Garden City, NY, 1966.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2003.

Jochen Hartmann, Yannick Exner, and Samuel Domdey. The power of generative marketing: Can generative ai create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1):13–31, 2025. ISSN 0167-8116. doi: https://doi.org/10.1016/j.ijresmar.2024.09.002. URL https://www.sciencedirect.com/science/article/pii/S0167811624000843.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Dacher Keltner and Jonathan Haidt. Approaching awe, a moral, spiritual, and aesthetic emotion. *Cognition & Emotion*, 17(2):297–314, 2003. doi: 10.1080/02699930302297.

Robert N. Kraft. The influence of camera angle on comprehension and retention of pictorial events. *Memory & Cognition*, 15(4):291–307, 1987.

Gunther Kress and Theo van Leeuwen. *Reading Images: The Grammar of Visual Design*. Routledge, London, 3 edition, 2021. ISBN 9780415672573. doi: 10.4324/9781003099857. URL https://doi.org/10.4324/9781003099857.

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

Jun Li, Zedong Zhang, and Jian Yang. Tp2o: Creative text pair-to-object generation using balance swap-sampling. In *European Conference on Computer Vision*, pp. 92–111. Springer, 2024.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026.

Kang Liao, Size Wu, Zhonghua Wu, Linyi Jin, Chao Wang, Yikai Wang, Fei Wang, Wei Li, and Chen Change Loy. Thinking with camera: A unified multimodal model for camera-centric understanding and generation. *arXiv preprint arXiv:2510.08673*, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Rushikesh Zawar, Xue Bai, et al. Towards understanding camera motions in any video. *Advances in Neural Information Processing Systems*, 38, 2025b.

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Weichao Chen, Yu Qiao, et al. Shotbench: Expert-level cinematic understanding in vision-language models. *Advances in Neural Information Processing Systems*, 38:129987–130019, 2025.

Jiaxiang Liu, Zhida Feng, Pengyu Zou, Zhenyu Qian, Tianrui Zhu, Jun Xia, Yuehu Dong, Yanzheng Lin, Honglin Xiong, Anqi Chen, et al. Ernie-image technical report. *arXiv preprint arXiv:2605.25347*, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2365–2376, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.168.

Google LLC. Open images dataset. `https://storage.googleapis.com/openimages/web/index.html`, 2020. Accessed: 2020-09-01.

L. M. Mandell and D. L. Shaw. Judging people in the news—unconsciously: Effect of camera angle and bodily activity. *Journal of Broadcasting*, 17(3):353–362, 1973. doi: 10.1080/08838157309363698.

Anders Giovanni Møller, Daniel M. Romero, David Jurgens, and Luca Maria Aiello. The impact of generative ai on social media: An experimental study. *Scientific Reports*, 16(1):9376, 2026. ISSN 2045-2322. doi: 10.1038/s41598-026-40110-8. URL `https://doi.org/10.1038/s41598-026-40110-8`.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL `https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf`.

OpenAI. Introducing chatgpt images 2.0 (gpt-image-2). `https://openai.com/index/introducing-chatgpt-images-2-0/`, 2026. Released April 21, 2026. Accessed July 2026.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Paul K. Piff, Pia Dietze, Matthew Feinberg, Daniel M. Stancato, and Dacher Keltner. Awe, the small self, and prosocial behavior. *Journal of personality and social psychology*, 108 6:883–99, 2015. URL `https://api.semanticscholar.org/CorpusID:992237`.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Zhang Haoxu, Yi Hu, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Brendan Rooney and Katalin E. Bálint. Watching more closely: Shot scale affects film viewers' theory of mind tendency but not ability. *Frontiers in Psychology*, 9:261, 2018. doi: 10.3389/fpsyg.2018.00261.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

An Sevenants and Géry d'Ydewalle. Does it matter to be pictured from below? effects of camera angle on the perception of film characters. *Psychologica Belgica*, 46(3):199–210, 2006. doi: 10.5334/pb-46-3-199.

Vasilis Tsiavos and Fotis Kitsios. The digital transformation of the film industry: How artificial intelligence is changing the seventh art. *Telecommunications Policy*, 49(8):103021, 2025. ISSN 0308-5961. doi: https://doi.org/10.1016/j.telpol.2025.103021. URL https://www.sciencedirect.com/science/article/pii/S0308596125001181.

Xinran Wang, Songyu Xu, Shan Xiangxuan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Kongming Liang, Zhanyu Ma, et al. Cinetechbench: A benchmark for cinematographic technique understanding and generation. *Advances in Neural Information Processing Systems*, 38, 2025.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2276774.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Huzheng Yang, Katherine Xu, Andrew Lu, Michael D Grossberg, Yutong Bai, and Jianbo Shi. Vibe spaces for creatively connecting and expressing visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21912–21921, 2026.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *ICCV*, 2023.

Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6358–6368, 2024.

APPENDIX

This supplement is organized in four parts.

A. Setup and materials.

- **Models** (A.1): version, access route, generation settings, and licensing terms for all twelve evaluated systems
- **Prompts Construction** (A.2) : the prompt templates for the six axes, verbatim intent suffixes and neutral controls

B. The six axes.

- **Awe** (B.1): reference construction, subject occupancy and far-scene depth
- **Romance** (B.2): reference construction, subject occupancy and interpersonal distance
- **Social Group** (B.3): reference construction and nearest-neighbor spacing
- **Power & Vulnerability** (B.4): camera-pitch estimation and sign convention
- **Relative Scale** (B.5): reference pairs across the four size regimes, measurement policy, and per-model signed log-ratio results

C. Estimator validity.

- **Pitch threshold calibration** (C.1): the $\pm 9^{\circ}$ rule, its human-annotated calibration and cross-validation, and its replication on an out-of-panel generator
- **Estimators on generated images** (C.2): training domains, quality-control filters, and per-model rejection rates

D. Uncertainty and controls.

- **Bootstrap analyses** (D.1): confidence intervals and rank stability for all six axes and for the average rank
- **Neutral baselines, L1 prompts** (D.2): pass rates without intent suffixes, and the paired suffix effect per model
- **Suffix robustness** (D.3): impact of different suffixes and the rank correlations between them
- **Reference specificity** (D.4): awe among photographed emotions

A SETUP

A.1 MODEL SETTINGS

All models were evaluated under the same protocol. The five expressive axes use a 3:2 landscape format at each model's closest native resolution: subject occupancy is a fraction of image area, so the frame must be fixed across models, and 3:2 is the format of the reference corpora and of the calibration set on which the $\pm 9^\circ$ pitch threshold was fitted (C.1) . Relative scale uses a square 1024×1024 format: its size ratio compares two entities to each other and does not depend on the frame, so the only constraint is that both entities fit, and pairs disparate in either direction (a person and a key, a person and a monument) fit a square equally well. Generation settings and access details per model are given in the next section.

Comparability across models. Two aspects of this protocol deserve comment. First, generated resolution is not uniform: most models render at 1536×1024 for the 3:2 axes, while each remaining model uses its closest native 3:2 resolution, from 1008×672 (GPT Image 2) to 2496×1664 (HiDream-O1); relative scale is rendered at 1024×1024 by all models. This does not affect the measurements. Subject occupancy is a fraction of image area and inter-subject distances are normalized by in-image head size, so both are resolution-independent by construction. Scene depth is estimated after resizing every image to a long side of 504 px with aspect preserved, so all 3:2 generations reach the depth model at 504×336 and reference images at the same scale; camera pitch is likewise estimated at Puffin's fixed input resolution (App. C.2). Second, sampler configurations differ by necessity: distilled models operate at few steps and low guidance, and closed endpoints expose no settings at all. We therefore evaluate each model as it is intended to be used rather than under a common sampler; comparisons concern deployed model behavior, not architectures under matched inference budgets.

Table 2 summarizes the parameters used for each model. Parameter counts are taken from the official model cards; counts marked with an asterisk refer only to the image generator, the associated text encoders adding further parameters. Closed-source endpoints do not disclose their sampling configurations, which are reported as unavailable.

Model	Access	Params	Steps	CFG	Weights (HuggingFace)
ERNIE Image	open	8B	50	4.0	`baidu/ERNIE-Image`
FLUX.2 Klein 4B	open	4B	4	1.0	`black-forest-labs/FLUX.2-klein-4B`
Qwen Image	open	20B	50	4.0	`Qwen/Qwen-Image`
HiDream I1 Full	open	17B*	50	5.0	`HiDream-ai/HiDream-I1-Full`
Z-Image Turbo	open	6B*	8	0.0	`Tongyi-MAI/Z-Image-Turbo`
Cosmos3 Super T2I	open	64B	50	4.0	`nvidia/Cosmos3-Super`
HunyuanImage 3.0	open	80B†	50	3.0	`tencent/HunyuanImage-3.0`
HiDream-O1	open	8B	50	5.0	`HiDream-ai/HiDream-O1`
GPT Image 1	endpoint	—	—	—	—
GPT Image 1.5	endpoint	—	—	—	—
GPT Image 2	endpoint	—	—	—	—
Nano Banana 2	fal-ai	—	—	—	—

Table 2: **Generation settings for the twelve evaluated models.** *Parameter count for the image generator only; the associated text encoders add further parameters. †HunyuanImage 3.0 has 80B total parameters, with 13B active parameters per token. Closed-source endpoints do not disclose parameter counts or sampler settings.

A.2 PROMPT CONSTRUCTION

Template. Five of the six axes share a two-level template:

$$\text{L1} = \{\text{SUBJECT}\} + \{\text{CONTEXT}\}, \qquad \text{L2} = \text{L1} + \{\text{SUFFIX}\}.$$

L1 is a neutral sentence naming only who is present and where; L2 appends a comma and a short phrase naming the target construct. The two levels differ by that phrase alone, character for character, so that the contrast between them isolates the effect of the stated intent. Table 3 lists the subject, suffix, and an example for each axis.

Axis	Subject	Suffix
Awe	a lone figure in front of	"an awe-inspiring scene"
Romance	two people in /at a	"a romantic scene"
Social group	a group	"a friendly scene"
Power	a {role}	"a powerful figure"
Vulnerability	a {role}	"a vulnerable figure"

Table 3: **Prompt templates by axis.** Each L2 prompt is its L1 sentence followed by the suffix. Example (awe): L1 "a lone figure in front of a sandstone arch", L2 "a lone figure in front of a sandstone arch, an awe-inspiring scene".

Lexical exclusions. Neither part of a prompt may name the geometry we score. Beyond explicit camera vocabulary (*angle*, *shot*, *view*, *from below/above*), we exclude four families of implicit cues:

- **Vertical prepositions and relations**—"towering over", "looking down on", "at the feet of": these encode the camera's elevation in the scene description itself.
- **Architectural elevation**—thrones on daises, balconies, podiums, ramparts: the setting imposes the viewpoint regardless of intent.
- **Postural cues**—"head held high", "cowering", "hunched": body configuration correlates with camera angle in photographic practice, and would let a model reach the target geometry through content rather than composition.
- **Dimensional and atmospheric adjectives**—"imposing", "vast", "looming": these describe the intended rendering rather than the subject.

The same principle governs the other axes: no prompt states that the awe subject should appear small or distant, that the couple should be close, or that the group should be tightly spaced.

Relative scale. This axis names two entities of markedly different real-world size and asks for their true proportion, with no emotional or relational intent to infer. Unlike the expressive axes, its templates may name the framing: the scored quantity is the size ratio, which no prompt states, and the depth placement of the reference entity is the controlled variable. Three templates place it at a different depth:

> *Same depth.* A wide-shot photograph of exactly one {subject} and exactly one {reference}, side by side at the same distance from the camera, shown at realistic relative physical scale. Both entities are fully visible with space around them.
>
> *Midground.* A wide-shot photograph of exactly one {subject} in the foreground and exactly one {reference} in the midground, shown at realistic relative physical scale. Both entities are fully visible with space around them.
>
> *Background.* A wide-shot photograph of exactly one {subject} in the foreground and exactly one {reference} far in the background, shown at realistic relative physical scale. Both entities are fully visible with space around them.

Entity pairs span four regimes of true size ratio $H_{\text{subject}}/H_{\text{reference}}$—micro ($\geq 6$), moderate (0.6–2.2), large (0.15–0.6), and extreme (0.04–0.15)—for 272 instances in all ($69/69/69/65$). Each pair's documented real-world dimensions and accepted ratio interval live in a manifest used only for scoring, never in the prompt.

B THE SIX AXES

B.1 AWE

Data collection. EmoSet is a large-scale visual-emotion dataset whose images were assigned emotion labels by human annotators Yang et al. (2023). We began with all 15,037 EmoSet images labeled *awe*. During curation, images for which no visible person was identified were excluded, and we retained images judged at selection time to contain one to three people watching, facing, or otherwise attending to a landscape. The image itself was not required to be a landscape-only photograph, but the people had to be oriented toward a visible landscape. This is the construct of Sec. 3.1—a figure before a vast scene—to which the reference is curated so that model and human compositions are compared like for like. Filtering and manual review produced a final quantitative set of 319 images.

Data filtering. The reference is built from the 15,037 images labelled *awe* in EmoSet (Yang et al., 2023) through a sequential filter chain, each stage applying one criterion:

1. **Metadata screening.** Images carrying a `human_action` or `object` annotation are removed, retaining scene-level photographs (15,037 → 7,595).
2. **Scenic filtering (CLIP ViT-B/32 (Radford et al., 2021)).** Wide scenic views are retained; portraits, close-ups, indoor scenes, and images dominated by foreground objects or animals are removed (7,595 → 6,131).
3. **Person detection (GroundingDINO).** Images where one to three persons are detected are retained, matching the benchmark's subject-in-a-scene construct (6,131 → 2,878).
4. **Screening and manual review.** GroundingDINO false positives, crowds, and off-context images are removed with VLM assistance (GPT-5.5), followed by a full manual review against a written criteria sheet (2,878 → 470).
5. **Single-subject validation.** A SAM3 person mask is extracted for each image and manually validated: the reference retains images with exactly one person, fully in frame, whose mask and depth estimate are usable (470 → 319).

Stages 1–3 are automatic and reproducible from released thresholds and model checkpoints; stages 4 and 5 involve manual judgment against a written criteria sheet. Image IDs, per-stage counts, the criteria sheet, and both annotators' verdicts are released with the benchmark.

Inter-annotator agreement. A second annotator, blind to the selection decisions, re-classified 100 images (50 accepted, 50 rejected, drawn from both judgment stages and presented in random order) against the written criteria, with Cohen's $\kappa = 0.70$ [95% CI 0.56–0.83]. Excluding the 12 contested images leaves the depth quartiles unchanged and shifts the median subject ratio by $+5\%$.

Subject occupancy (SAM3 person mask). We compute subject occupancy from the SAM3 person *mask*, not its bounding box, so that it reflects the pixels the subject actually covers rather than a loose enclosing rectangle. Because the final Awe human reference contains one validated person per image, let M_s denote that person's mask and A the image area. We define occupancy as

$$o \;=\; \frac{|M_s|}{A},$$

the fraction of the frame occupied by the subject. No averaging or union across multiple people is required.

Far-scene depth (z_{far}). Scene depth is estimated with DA3Nested-Giant-Large-1.1 (Depth Anything 3 (Lin et al., 2025a)), a nested model that combines an any-view geometry model with a monocular metric estimator to output a per-pixel depth map D directly in metres (no focal-based rescaling is applied on our side; the model also returns estimated camera intrinsics and a sky mask, which we use for M_{sky}). We use the `-1.1` checkpoint, retrained after a training-bug fix. Let Ω be the image domain and M_{sky} the estimated sky mask. We define far-scene depth as the 98th percentile of depth over the visible, non-sky scene:

$$z_{\text{far}} \;=\; P_{98}(\{\, D(p) \mid p \in \Omega \setminus M_{\text{sky}},\; D(p) > 0 \,\})\,.$$

We take the 98th percentile rather than the maximum to remain robust to stray depth errors, and exclude the sky so that its unbounded predicted depth does not dominate the measure. Unlike a subject-relative ratio, z_{far} does not depend on the subject's position in the frame.

Aggregation. The two descriptors are essentially uncorrelated in the Awe human reference (Spearman $\rho = -0.056$, bootstrap 95% CI $[-0.168, 0.059]$; log–log Pearson $r = -0.088$, CI $[-0.196, 0.025]$; $R^2 \approx 0.78\%$). A small subject therefore does not imply a deep scene, nor the converse: subject scale and scene depth are independent compositional choices, and scoring them separately does not double-count a single underlying quantity. For a model, we compute the Wasserstein-1 distance between its distribution and the human reference separately for each descriptor, divide by the corresponding human IQR, and report the awe score as the mean of the two normalized components.

Validity of generated images. Generated images are held to the same subject criteria as the reference. An automatic check flags generations whose subject mask touches the frame border; flagged images are reviewed manually against the reference criteria sheet, and only subjects actually truncated by the frame are excluded—border contact without truncation (e.g. feet at the lower edge of a fully visible figure) is retained. Excluded generations yield no valid occupancy and are removed from the model's distribution rather than scored as $o = 0$. Table 4 gives the per-model counts; the most affected model retains 142 of 150 generations.

Model	Excluded	Rate
GPT Image 1	8/150	5.3%
FLUX.2 Klein 4B	5/150	3.3%
GPT Image 1.5	4/150	2.7%
Cosmos3 Super T2I	3/150	2.0%
Qwen Image	3/150	2.0%
HiDream I1 Full	1/150	0.7%
Nano Banana 2	1/150	0.7%
ERNIE Image, GPT Image 2, HiDream-O1, HunyuanImage 3.0, Z-Image Turbo	0/150	0%
Total	25/1,800	1.4%

Table 4: Awe generations excluded after manual adjudication of frame truncation. Excluded image IDs are released with the benchmark.

Metric	Mean	Median	IQR $[Q_1, Q_3]$
Subject Occupancy o (%)	1.55	0.57	1.408 [0.16, 1.57]
Far-scene depth z_{far} (m)	58.98	49.51	53.87 [28.10, 81.97]

Table 5: Summary statistics for the EmoSet awe reference set (N=319). The interquartile range is the normalizing quantity in W_1/IQR.

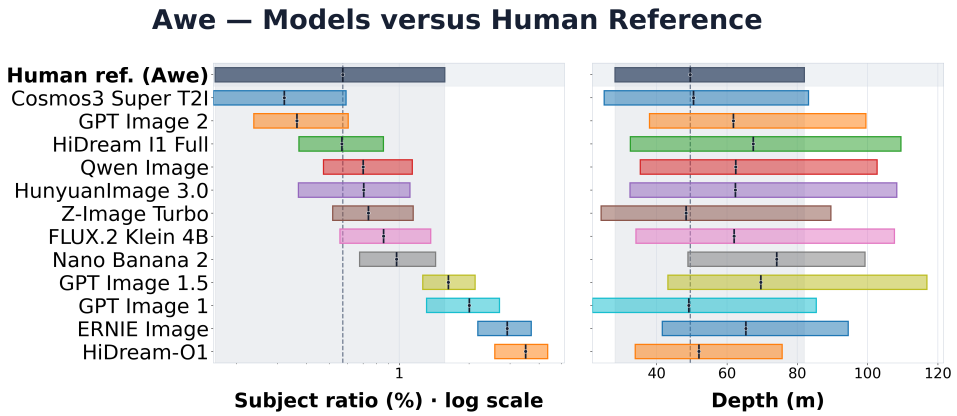

Figure 7: **Awe: subject occupancy and far-scene depth by model.** Bars are interquartile ranges, ticks are medians; top row and shaded band are the human reference. Subject occupancy is on a log scale.

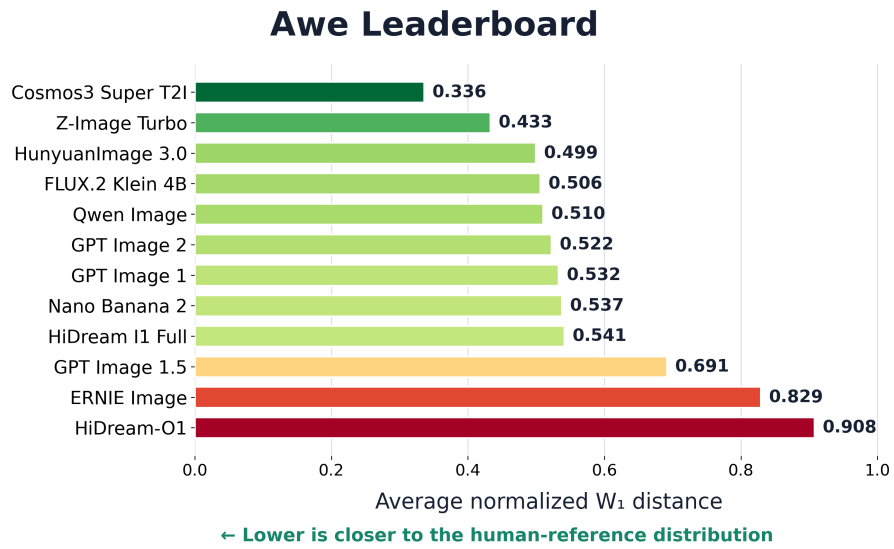

Figure 8: **Awe leaderboard.** Average normalized W_1 distance to the human reference.

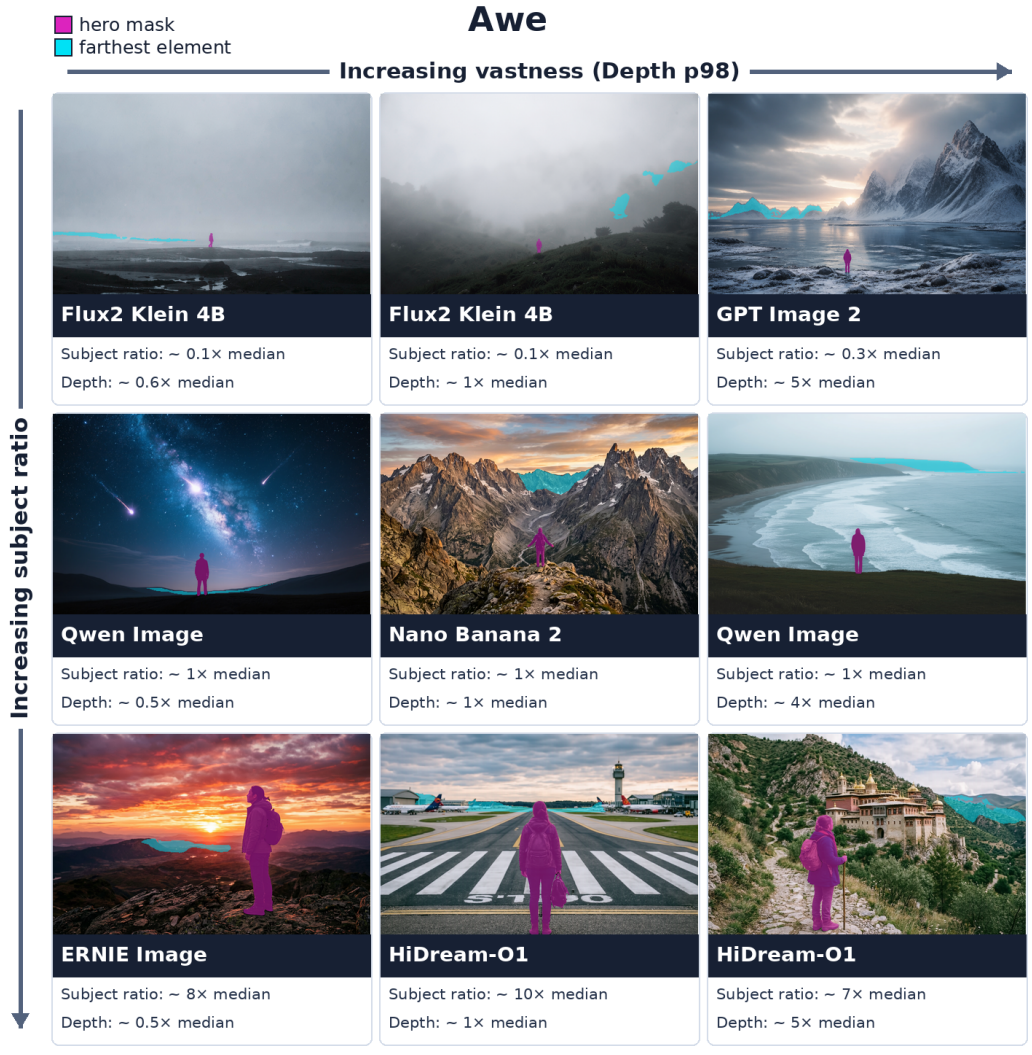

Figure 9: **The awe measurement space.** Generations arranged by subject occupancy (increasing downward) and far-scene depth (rightward), in multiples of the human median; overlays show the subject mask (magenta) and farthest scene element (cyan). Bottom row: subjects 9–10× the human median, yet with the full iconography of awe—models learn the motifs without the proportions.

B.2 ROMANCE

Data collection. We constructed the reference set from images carrying the positive, human-verified Open Images V7 label *Romance* (LLC, 2020). After retaining photographic content and reviewing the person detections and masks, the final quantitative reference set contained $N = 535$ images with two usable SAM3 masks and valid head extractions for both subjects. The Open Images label is broad: it records that human annotators associated an image with romance, but does not by itself guarantee a particular pose, interaction, or photographic composition. The measurements below therefore characterize the visual staging observed in the curated two-person set rather than define romance a priori.

Pair occupancy. Occupancy is computed from the union of the two SAM3 person masks, not their bounding boxes, so that it reflects the pixels the pair covers (Sec. 3.2). Subject scale was not used as a curation criterion.

Head extraction and quality control. Each subject's head mask is the union of the *face*, *head*, *hair*, and *head-including-hair* segmentations, intersected with the person mask so that no head region originates from the background. A head is discarded when its mask is empty, duplicated, anomalously large, or below 1% of its person mask; a pair is discarded when the ratio of the two head areas exceeds $\tau_{\mathrm{QC}} = 6.14$, the 99.5th percentile of that ratio on the human reference, which removes cases where one "head" is a spurious background fragment. On the reference, 3 of 538 pairs are discarded by this criterion, leaving $N{=}535$. The same procedure and threshold apply to generated images.

Head-normalized interpersonal distance. Let c^L, c^R be the head centroids and m_L, m_R the head mask areas. Taking $\sqrt{m}$—a length, homogeneous to the numerator—as the characteristic head size, we define

$$d = \frac{\|c^L - c^R\|_2}{\frac{1}{2}\left(\sqrt{m_L} + \sqrt{m_R}\right)},$$

the distance between the two heads in head-size units, the two sizes averaged so that the measure remains stable when the subjects lie at different depths. This is an image-space operationalization of proximity, not a reconstruction of physical distance. On the human reference (Table 7), the median corresponds to subjects staged at or near facial contact—two faces in contact are separated by roughly one head size—consistent with Hall's intimate zone (Hall, 1966), which at $0.46\,\mathrm{m}$ amounts to about two head sizes. The distribution is narrow: human photographers stage romantic pairs within a tight band of apparent proximity.

Aggregation. The two descriptors are weakly correlated in the human reference (Spearman $\rho = -0.141$, 95% CI $[-0.222, -0.059]$; $R^2 \approx 2\%$): larger pairs tend very slightly toward smaller distances, but the association is far from redundancy, so scoring the two separately and averaging does not double-count a single quantity. The romance score is the mean of the two W_1/IQR components, as for awe (Sec. 4.4).

Validity of generated images. A generation is scored when two head masks pass the quality-control criteria above. When more than two people are rendered, the two heads of the romantic pair are selected; in five Cosmos3 generations flagged during audit, this selection was corrected manually. Generations in which no single reliable pair could be identified are marked unusable and excluded, as are those the scorer rejects under the head criteria; generations that the model or its API did not return are counted as missing. Excluded generations yield no valid measurement and are removed from the model's distribution rather than scored. Table 6 gives the per-model counts; the most affected model retains 142 of 150.

Model	Scored	Rejected	Missing	Unscored
Cosmos3 Super T2I	142	7	1	5.3%
ERNIE Image	144	6	0	4.0%
HunyuanImage 3.0	145	5	0	3.3%
GPT Image 2	145	0	5	3.3%
FLUX.2 Klein 4B	149	1	0	0.7%
GPT Image 1.5	149	0	1	0.7%
Z-Image Turbo	149	1	0	0.7%
GPT Image 1, HiDream I1 Full, HiDream-O1, Nano Banana 2, Qwen Image	150	0	0	0%
Total	1,773	20	7	1.5%

Table 6: Romance generations scored, rejected by the head criteria or as unusable, and missing, out of 150 per model. Excluded image IDs are released with the benchmark.

Metric	Median	IQR	$[Q_1,\ Q_3]$
Subject ratio m (%)	42.34	35.48	$[25.37,\ 60.85]$
Interpersonal distance d	0.983	0.237	$[0.876,\ 1.112]$

Table 7: Summary statistics for the Open Images *Romance* reference set (N=535). The interquartile range is the normalizing quantity used in the W_1/IQR score.

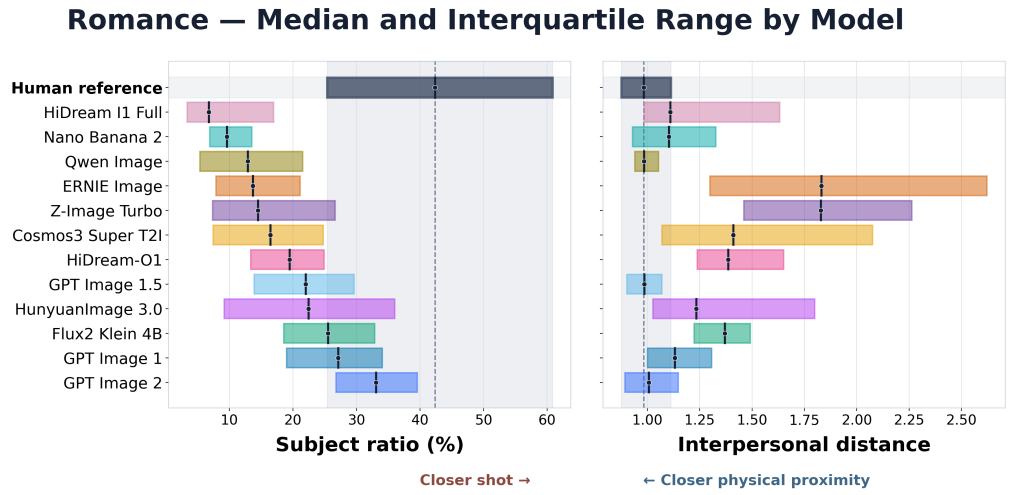

Figure 10: **Romance: pair occupancy and interpersonal distance by model.** Bars are interquartile ranges, ticks are medians; top row and shaded band are the human reference.

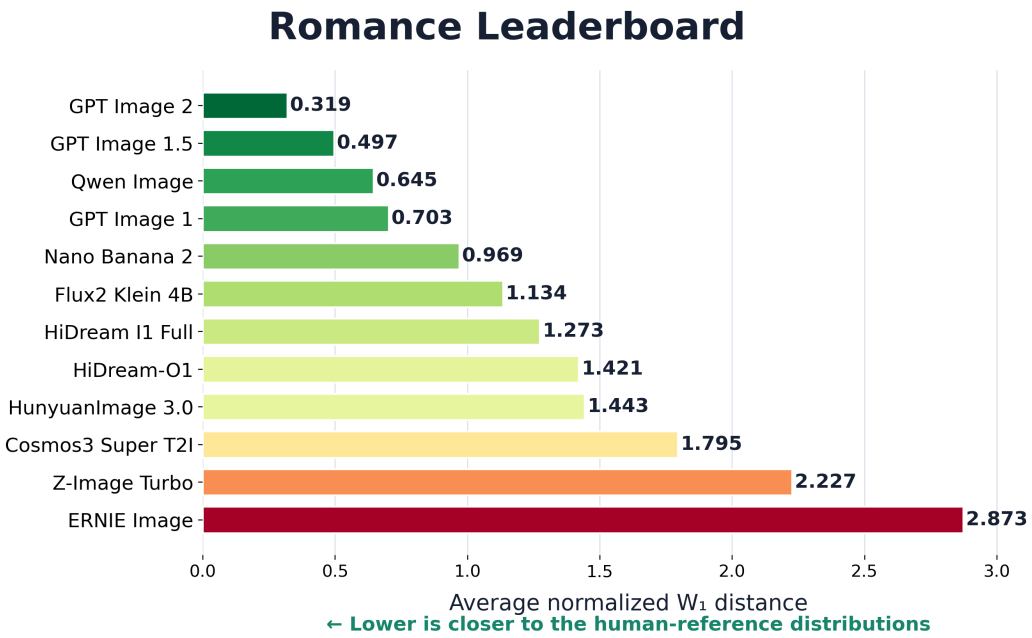

Figure 11: **Romance leaderboard.** Average normalized W_1 distance to the human reference; lower is closer.

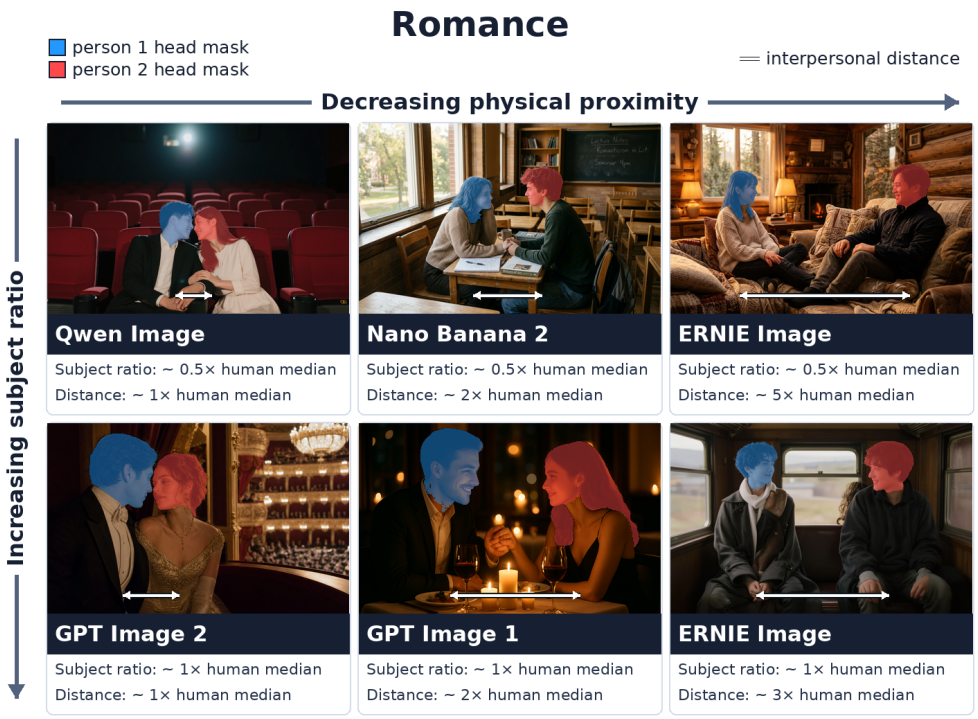

Figure 12: **Romance: qualitative examples.** Generated images arranged by the two descriptors, as multiples of the human median: rows by pair occupancy (top $\approx 0.5\times$, bottom $\approx 1\times$), columns by interpersonal distance ($\approx 1\times$, $2\times$, and 3–$5\times$). Head masks are overlaid (blue, red) and the white segment marks the head-to-head distance. Each image is one L2 generation from the model named.

B.3 SOCIAL GROUP

Data collection. The reference is built from images carrying the positive, human-verified Open Images V7 label *Social group* (LLC, 2020). After retaining photographic content and reviewing person detections, the reference contains $N{=}1{,}006$ images with at least two participating subjects. Person masks are foreground SAM3 instances, retained according to their area, height, and vertical location relative to the primary person candidate; they are not exhaustive annotations of every visible person, nor definitive annotations of group membership. The same selection rule is applied to generated images. The Open Images label is broad—posed groups, teams, public events, crowds, small gatherings—and the measurements below characterize the staging of this reference set rather than define a social group a priori.

Head extraction and quality control. Head masks are built as for romance (App. B.2): union of the *face*, *head*, *hair*, and *head-including-hair* segmentations, intersected with the person mask. A head is discarded when its mask is empty, duplicated, above 60% or below 1% of its person mask; a nearest-neighbor pair is discarded when the ratio of the two head areas exceeds $\tau_{\mathrm{QC}} = 6.14$, the 99.5th percentile of that ratio on the human reference. On the reference, $2{,}518$ of $15{,}152$ head contributions are discarded (16.6%), leaving $12{,}634$ valid nearest-neighbor distances; no image is removed, all $1{,}006$ retaining at least two valid heads. The same procedure and threshold apply to generated images, where the discard rate is 17.4%.

Head-normalized nearest-neighbor distance. Let c_j and m_j be the centroid and mask area of person j's head. With $\sqrt{m_j}$—a length, homogeneous to the numerator—as the characteristic head size, each person's distance to their nearest neighbor in head-size units is

$$d_j = \frac{\|c_j - c_k\|_2}{\frac{1}{2}\left(\sqrt{m_j} + \sqrt{m_k}\right)}, \qquad k = \arg\min_{l \neq j} \|c_j - c_l\|_2,$$

the two sizes averaged so the measure remains stable when the pair straddles depth layers. Each person contributes one nearest-neighbor ratio; we do not average all pairwise distances. The image-level score is the median over valid contributions, $d = \mathrm{median}_j(\rho_j)$, rather than the mean: the within-image distribution is right-skewed, and a single peripheral member or residual segmentation error would displace a mean while leaving the typical spacing unchanged.

Aggregation. For a model, we compute the Wasserstein-1 distance between its distribution of d_i and the human-reference distribution, and divide by the human IQR. Lower W_1/IQR indicates a compositional distribution closer to human practice.

Metric	Mean	Median	IQR $[Q_1, Q_3]$
Nearest-neighbor distance d (head-size units)	1.610	1.438	0.561 [1.231, 1.792]

Table 8: Summary statistics for the Open Images social group reference ($N{=}1{,}006$, all images retained). d is the image-level median of head-normalized nearest-neighbor distances; the interquartile range is the normalizing quantity in W_1/IQR.

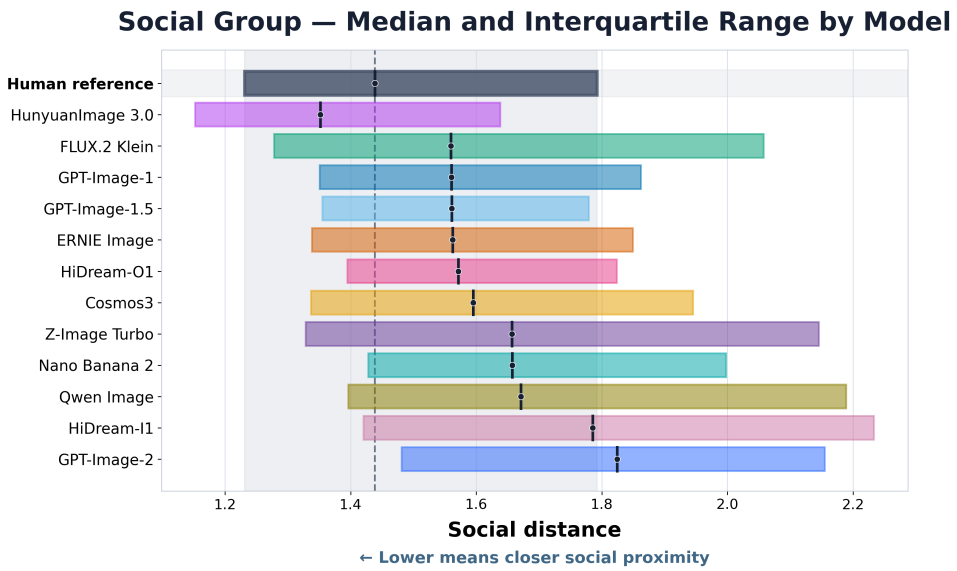

Figure 13: **Social group: head nearest-neighbor distance by model.** Bars are interquartile ranges, ticks are medians; top row and shaded band are the human reference.

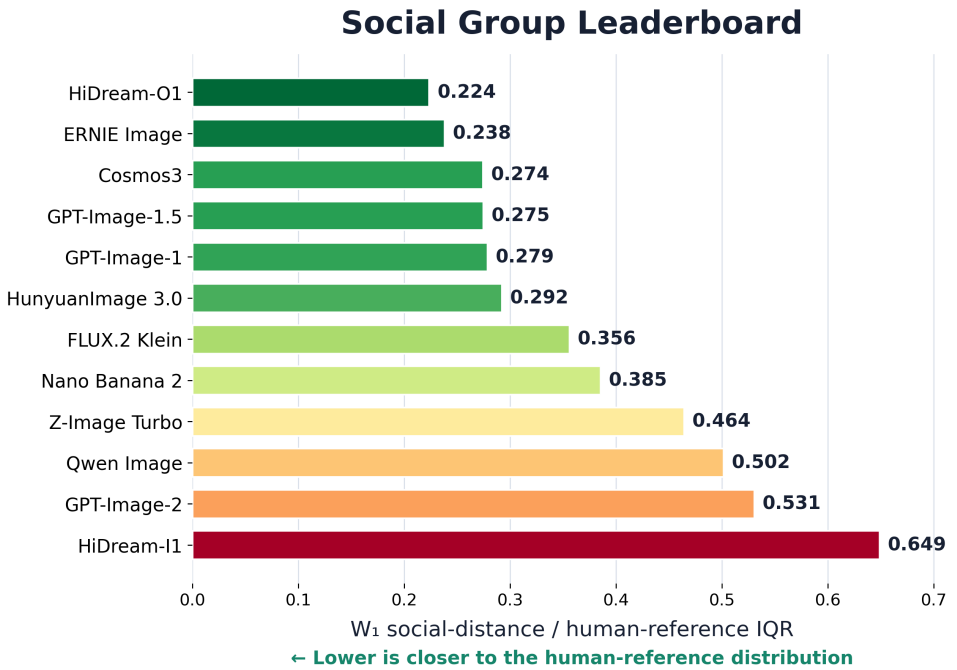

Figure 14: **Social group leaderboard.** Normalized W_1 distance to the human reference; lower is closer.



Figure 15: **Social group: qualitative examples.** Generated images ordered by head nearest-neighbor distance ρ, as multiples of the human median ($0.5\times$, $1\times$, $2\times$, $5\times$). Bottom row: the same images with head masks overlaid and each head linked to its nearest neighbor. Each image is one L2 generation from the model named.

B.4 Power and Vulnerability

No reference distribution. Power and vulnerability are not scored against a human distribution but against a direction established across decades of visual-communication research: a low angle confers dominance, a high angle vulnerability, straight-on reads as neutral (Kraft, 1987; Giessner et al., 2011; Mandell & Shaw, 1973; Sevenants & d'Ydewalle, 2006). Kraft (1987) states the mechanism directly—a low-angle shot "forces the viewers to look up at the actors, placing the actors in a position of visual authority and dominance." The photographic signatures of the two constructs are thus the two poles of one signed parameter, camera pitch, and a generation is scored by whether its pitch has the conventional sign. No reference corpus is needed: the direction is supplied by the literature.

From pitch to a pass. Camera pitch θ is estimated with Puffin's understanding branch (Liao et al., 2025), chosen over a geometric calibrator (GeoCalib) and a VLM judge on agreement with human annotation (App. C.1). Puffin returns the global camera orientation; for our single, centered subjects this coincides with the subject-relative angle. With the sign convention of Sec. 3.4 ($\theta > 0$ looking up), a generation is classed `low_angle` if $\theta > +9°$, `high_angle` if $\theta < -9°$, and `eye_level` otherwise. Power passes on `low_angle`, vulnerability on `high_angle`; eye-level and the opposite class both count as failures. The pass rate is the fraction of a model's total generations that pass. The $\pm 9°$ threshold is calibrated once on a held-out set and frozen—applied identically to all twelve models and both poles, never re-optimized per model (App. C.1).

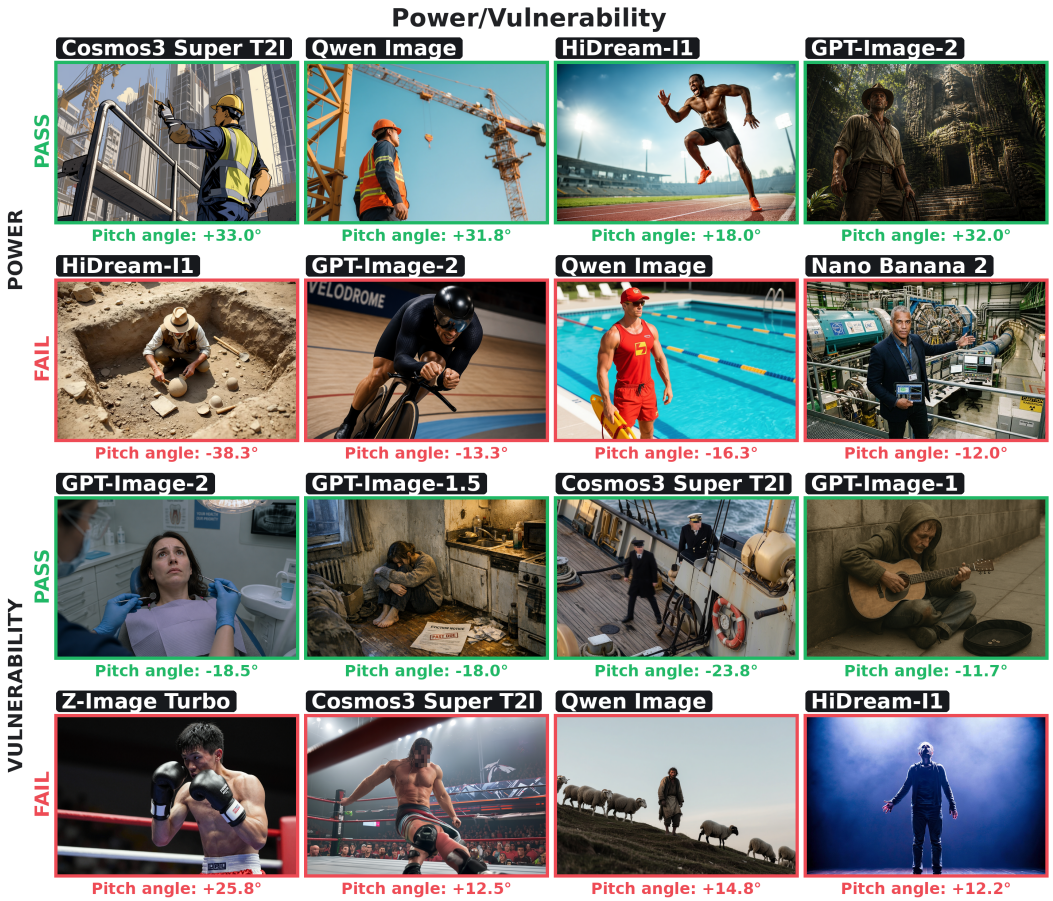

Figure 16: **Power and vulnerability: qualitative examples.** L2 generations with their estimated camera pitch θ (Puffin). Rows: power (top two) and vulnerability (bottom two); green frames pass ($\theta > +9°$ for power, $\theta < -9°$ for vulnerability), red frames fail.

B.5 RELATIVE SCALE

Overview. Prompts, regimes, and counts are given in App. A.2 (272 instances over N entity pairs and three depth arrangements). Every prompt constrains both entities to be fully visible and ground-supported, so that the measured ratio compares standing heights. For each generated image we recover a depth-corrected size ratio and compare it to the interval spanned by the entities' documented real-world dimensions (Sec. 3.5).

Detection and apparent size. Each named entity is localized with a GroundingDINO (Liu et al., 2023) bounding box. Its apparent size e_i is the box height, except for entities defined by their length (e.g. a key, pen, smartphone, or USB drive), for which we use the longer side $\max(w_i, h_i)$. A generation for which either entity is not detected yields no valid measurement and is counted as a failure.

Depth estimation. Apparent size alone conflates real-world size with distance: a small nearby object and a large distant one project to the same pixel size. We estimate per-object depth with DA3Nested-Giant-Large-1.1 (Depth Anything 3 (Lin et al., 2025a), 1.1 checkpoint), which outputs metric depth in metres, at the working resolution of App. C.2. A median over the whole box would mix the entity with background pixels, especially for thin or tall objects, so depth is read over a ground-contact region—the bottom 5% of the box height, restricted to its central 50% width—and D_i is the median valid prediction there. Because every prompt is ground-supported, this region reflects true ground contact, comparable across entities sharing the ground plane. The same convention applies to subject and reference in all three depth arrangements.

Depth-corrected ratio. Under a pinhole camera model (Hartley & Zisserman, 2003), an object's real-world size is proportional to its apparent size scaled by its distance from the camera. The size ratio of the two entities is therefore

$$R \;=\; \frac{e_s}{e_r} \cdot \frac{D_s}{D_r},$$

where (e_s, e_r) are the apparent sizes and (D_s, D_r) the ground-contact depths of the subject and reference. The factor D_s/D_r removes the perspective foreshortening introduced when the two entities lie at different distances from the camera, so that R recovers their true size relationship rather than an artifact of placement. Because only the depth *ratio* enters, the unknown focal length and any global multiplicative scale in the depth predictions both cancel; the score is thus invariant to the absolute scale of the depth model.

Expected interval and scoring. For each pair we assign plausible real-world height ranges $[H_{s,\min}, H_{s,\max}]$ and $[H_{r,\min}, H_{r,\max}]$ for the subject and reference, and take the physical ratio at its widest—smallest subject over largest reference, and largest subject over smallest reference. We then apply a 15% tolerance, giving the accepted interval

$$R_{\text{accepted}} \;=\; \left[\, 0.85\,\frac{H_{s,\min}}{H_{r,\max}},\;\; 1.15\,\frac{H_{s,\max}}{H_{r,\min}} \,\right].$$

The tolerance absorbs residual variation—within-category size differences (e.g. dog breeds, car models), imperfect box localization, and depth-estimation noise—so a generation is not failed for a plausible ratio lying just outside the nominal physical range. A generation passes when $R \in R_{\text{accepted}}$ and fails otherwise; because generation, detection, or depth failures all yield no valid R, the reported pass rate is a lower bound on the true pass rate.

Diagnostic segmentation. For the Micro regime we additionally run SAM3 (Carion et al., 2025) and retain its masks for inspection, but the final $v8$ scores use GroundingDINO boxes for both apparent size and depth; SAM3 is not used in the reported ratio.



Figure 17: **Relative scale: qualitative examples by regime.** Columns are the four regimes of true size ratio; the top two rows pass, the bottom two fail, each image one generation from the model named with its entity pair. Consistent with Fig. 6, rendered ratios are pulled toward unity: in the micro regime the small entity comes out too large, in the large and extreme regimes the large entity too small.

Table 9: **Relative-scale entity pairs and accepted ratio intervals.** Subject is a standing person unless noted in parentheses. Regimes are assigned on the nominal ratio of typical heights; intervals span the plausible height ranges of both entities with the $\pm 15\%$ tolerance and may extend past the regime boundaries. Extreme pairs use the midground and background arrangements only.

Reference	Interval	Reference	Interval
Micro (ratio ≥ 6)		**Large** (ratio 0.15–0.6)	
strawberry	[27.2, 85.1]	small airplane, parked	[0.389, 0.925]
USB flash drive	[19.4, 53.2]	standing camel (child)	[0.388, 0.839]
key	[18.1, 47.3]	school bus	[0.412, 0.734]
chicken egg	[20.9, 38.7]	standing elephant	[0.340, 0.709]
ant (open hand)	[14.4, 60.4]	helicopter, on the ground	[0.302, 0.788]
tennis ball	[19.7, 32.7]	fire truck	[0.349, 0.645]
ladybug (open hand)	[14.4, 48.3]	standing Golden Retriever (cat)	[0.315, 0.586]
apple	[16.0, 30.4]	steam locomotive	[0.296, 0.560]
standing mouse	[13.6, 35.5]	double-decker bus	[0.309, 0.532]
butterfly	[11.3, 42.5]	standing giraffe	[0.247, 0.473]
light bulb	[11.3, 21.3]	street lamppost	[0.227, 0.473]
pen	[9.07, 16.4]	wooden telephone pole	[0.194, 0.532]
smartphone	[8.24, 15.2]	town clock tower	[0.194, 0.473]
spoon	[7.56, 14.2]	stone obelisk	[0.170, 0.532]
toothbrush	[7.16, 13.3]	classical stone column	[0.170, 0.532]
banana	[6.80, 13.3]	tall Christmas tree	[0.170, 0.532]
fork	[6.80, 11.8]	maypole	[0.170, 0.425]
screwdriver	[5.44, 11.8]	spruce tree	[0.170, 0.425]
coin (open hand)	[4.82, 13.4]	totem pole	[0.151, 0.425]
paperclip (open hand)	[4.38, 8.62]	standing person (cat)	[0.106, 0.201]
Moderate (ratio 0.6–2.2)		**Extreme** (ratio 0.04–0.15)	
standing Great Dane	[1.58, 3.00]	utility pole	[0.097, 0.236]
pool table	[1.66, 2.73]	saguaro cactus	[0.091, 0.236]
washing machine	[1.43, 2.50]	flagpole	[0.068, 0.213]
bicycle	[1.24, 2.13]	palm tree	[0.068, 0.177]
office water cooler	[1.13, 2.13]	oak tree	[0.054, 0.142]
standing emperor penguin	[1.05, 2.13]	windmill	[0.045, 0.142]
motorcycle	[1.09, 2.03]	grain silo	[0.045, 0.142]
upright piano	[1.01, 1.85]	large obelisk	[0.045, 0.106]
standing brown bear	[0.907, 2.13]	minaret	[0.036, 0.106]
standing donkey	[0.907, 1.77]	church steeple	[0.036, 0.106]
parking meter	[0.907, 1.77]	factory chimney	[0.034, 0.085]
standing kangaroo	[0.756, 1.64]	cathedral spire	[0.034, 0.085]
standing gorilla	[0.756, 1.52]	high-rise apartment tower	[0.034, 0.076]
standing llama	[0.756, 1.42]	radio tower	[0.034, 0.076]
floor lamp	[0.756, 1.42]	stadium floodlight tower	[0.034, 0.076]
refrigerator	[0.716, 1.25]	cooling tower	[0.034, 0.071]
standing mannequin	[0.716, 1.25]	redwood tree	[0.034, 0.071]
doorway	[0.648, 1.07]	water tower	[0.030, 0.085]
vending machine	[0.680, 1.18]	lighthouse	[0.027, 0.106]
tank (standing soldier)	[0.539, 0.967]	grain elevator	[0.030, 0.085]
standing ostrich	[0.504, 1.01]	clock tower	[0.027, 0.071]
humanoid robot, $\sim$2.5 m	[0.523, 0.886]	electricity pylon	[0.027, 0.071]
phone booth	[0.523, 0.886]		
standing adult (child, $\sim$6 y)	[0.505, 0.898]		
shipping container	[0.525, 0.821]		
tractor	[0.425, 0.818]		

C ESTIMATORS VALIDITY

C.1 PITCH THRESHOLD CALIBRATION

Why calibration is needed. Power and vulnerability both reduce to a single geometric quantity: the vertical camera angle relative to the subject. Scoring them requires deciding whether the camera looks upward at the subject (`low_angle`, the framing associated with power), downward (`high_angle`, associated with vulnerability), or approximately straight ahead (`eye_level`). We estimate the camera's vertical orientation with Puffin (Liao et al., 2025), which returns a continuous pitch in degrees rather than the three classes. Converting it into a categorical decision requires boundaries that materially affect the reported pass rates: too close to zero, and perceptually insignificant tilts count as genuine compositions; too far, and visually meaningful angles are absorbed into the neutral class. We therefore calibrate the boundary ourselves, on a balanced set of generations annotated by hand, rather than fixing it by intuition or importing a value from another domain.

From continuous pitch to three classes. Under our sign convention, a negative pitch indicates a camera directed downward, a pitch near zero an approximately horizontal camera, and a positive pitch a camera directed upward. We convert the continuous pitch θ, in degrees, into the three classes using a single symmetric threshold τ:

$$\text{class}(\theta) = \begin{cases} \texttt{high_angle} & \theta < -\tau, \\ \texttt{eye_level} & -\tau \le \theta \le \tau, \\ \texttt{low_angle} & \theta > \tau. \end{cases} \tag{1}$$

The half-width τ is the only free parameter. A symmetric boundary preserves zero pitch as the natural centre of the eye-level class and applies the same angular tolerance to upward and downward deviations.

Calibration protocol. The threshold is calibrated on a generator outside the evaluated panel, so that no evaluated model's output statistics can influence the decision boundary. We generated a balanced set of 300 images with Seedream-v4 (Seedream et al., 2025) (100 per class) using explicitly instructed viewpoints; because instruction compliance is imperfect, every image was manually annotated for its *actual* angle class, and this annotation—not the instruction—serves as ground truth. Sweeping τ over integer values from 5° to 10° (Table 10) gives a maximum accuracy of 91.3% at $\tau = 9^\circ$. The trade-off is readable across the sweep: raising τ trades recall on the tilted classes for recall on eye-level, with 9° at the balance point. Accuracy varies by less than 2.5 points over the whole range, so no reasonable choice of threshold would alter our conclusions.

Cross-validated selection. To verify that this optimum is not an artefact of selecting on the same data, we repeat the procedure under five-fold cross-validation, stratified by class (240 calibration / 60 held-out images per fold). Within each fold the threshold is selected on the calibration split alone and applied once to the held-out split, so no test image contributes to the choice. The selected threshold is 9° in three folds and 8° in two, with an out-of-fold accuracy of 90.7% over the 300 concatenated held-out predictions. We adopt $\tau = 9^\circ$ and freeze it across all models and both axes.

τ	Accuracy	Recall High	Recall Eye	Recall Low
5°	89.0%	99%	71%	97%
6°	89.0%	99%	73%	95%
7°	89.7%	99%	77%	93%
8°	91.0%	99%	86%	88%
9°	**91.3%**	**99%**	**88%**	**87%**
10°	89.7%	96%	90%	83%

Table 10: **Symmetric threshold sweep** on the balanced 300-image Seedream-v4 calibration set, human-annotated for actual camera angle. The decision rule classifies pitch below $-\tau$ as high-angle, above $+\tau$ as low-angle, and values within $[-\tau, +\tau]$ as eye-level. Accuracy peaks at $\tau = 9^\circ$ and varies by less than 2.5 points across the range.

C.2 DEPTH-ESTIMATOR VALIDITY

Depth is estimated with Depth Anything 3 (DA3) (Lin et al., 2025a) rather than annotated by hand. Its teacher–student pipeline combines large-scale synthetic supervision with diverse real-world data, so the model has seen synthetic imagery during training and the domain gap to generated images is smaller than for estimators trained on real photographs alone. This makes it a suitable fixed estimator for cross-model comparison: every image, generated or reference, passes through the same model at the same resolution (long side 504 px, aspect preserved), so any estimator bias is shared. We nevertheless treat its outputs as model-based estimates, not ground-truth depth.

D UNCERTAINTY AND CONTROLS

D.1 BOOSTRAP ANALYSES

All uncertainty estimates follow a common two-level bootstrap design (1,000 replicates): each replicate resamples both the prompt set and the human reference, with draws shared across models so that rank comparisons remain paired (Sec. 4.4). The composite six-axis analysis pairs the per-axis bootstrap replicates using seed 42. This section reports the resulting intervals: the average-rank stability of the overall leaderboard (Table 11), followed by the per-axis score and rank intervals (Tables 12–18).

The overall ranking is best read through Table 11, whose two summaries deliberately dissociate the point ranking from the bootstrap winning probability. Cosmos3 Super T2I has the best point estimate (average rank 4.92), yet finishes first in only 9.1% of replicates: its per-axis ranks are comparatively stable—leading on Awe while remaining near the bottom on Romance and Relative Scale across most draws—which keeps even its most favorable replicates near 4.5. GPT Image 2 and GPT Image 1, whose ranks fluctuate more, top the ranking most often (32.5% and 32.0% of replicates, a negligible difference), and GPT Image 2 is also the most reliable podium finisher (top three in 75.1% of replicates). No model finishes first in more than one third of replicates. Which model "leads" thus depends on whether one asks for the best point estimate or the most frequent bootstrap winner—the quantitative form of the paper's central negative result.

Rank	Model	Avg. Rank	95% CI	P(rank 1)	P(top 3)
1	Cosmos3 Super T2I	4.92	[4.50, 6.59]	9.1%	54.1%
2	GPT Image 2	5.08	[3.58, 6.25]	32.5%	75.1%
3	GPT Image 1	5.17	[3.75, 6.42]	32.0%	71.6%
4	Z-Image Turbo	5.25	[4.75, 7.42]	3.7%	20.5%
5	FLUX.2 Klein 4B	5.67	[4.33, 7.58]	9.0%	27.4%
6	GPT Image 1.5	6.17	[4.92, 7.33]	0.4%	8.5%
7	Qwen Image	6.42	[4.92, 7.75]	1.1%	13.0%
8	HunyuanImage 3.0	6.50	[4.33, 8.00]	6.4%	19.5%
9	Nano Banana 2	7.67	[5.92, 8.83]	0.0%	0.4%
9	ERNIE Image	7.67	[6.92, 9.25]	0.0%	0.0%
11	HiDream-O1	8.50	[7.33, 9.75]	0.0%	0.0%
12	HiDream-I1 Full	9.00	[6.50, 9.42]	0.0%	0.1%

Table 11: **Average-rank stability** over 1,000 paired bootstrap replicates (seed 42; vulnerability ranks on the 139-prompt intersection). Avg. Rank is the deterministic six-axis point estimate of Table 1; its confidence interval is computed from the mean of the six bootstrap rank vectors, without re-ranking. P(rank 1) and P(top 3) are obtained by re-ranking the twelve models within each replicate. The Awe bootstrap uses the final audited 319-image human reference and excludes the 25 manually confirmed cropped Awe L2 generations.

Awe. Table 12 gives the per-model intervals for the awe composite. The leader is settled, the middle is not: Cosmos3 Super T2I reaches the top three in 98.5% of replicates (rank CI $[1, 3]$), Z-Image Turbo follows (77.7%), and ranks 3–9—six models within 0.04 of one another—share overlapping rank CIs of $[2, 9]$ or $[3, 9]$: mid-field positions on this axis are representative draws from an interchangeable set, which is why the sensitivity analyses of Appendix B.1 reorder precisely this region and nothing else. The three-model tail is settled ($P(\text{top } 5) = 0$ throughout).

Table 12 gives the per-model intervals for the awe composite. The leader is settled, the middle is not: Cosmos3 Super T2I reaches the top three in 99.3% of replicates (rank CI $[1, 2]$), Z-Image Turbo follows (76.3%), and ranks 3rd to 9th –seven models within about 0.04 of one another–share overlapping rank CIs of $[2, 9]$ or $[3, 9]$: mid-field positions on this axis are representative draws from an interchangeable set, which is why the sensitivity analyses of Appendix B.1 reorder precisely this region and nothing else. The three-model tail is settled ($P(\text{top } 5) = 0$ throughout).

		Score		Rank stability		
Rank	**Model**	**Avg.** W_1/IQR	**95% CI**	**Rank CI**	**P(top 3)**	**P(top 5)**
1	Cosmos3 Super T2I	0.336	[0.274, 0.473]	[1, 2]	99.3%	100.0%
2	Z-Image Turbo	0.433	[0.337, 0.589]	[1, 8]	76.3%	91.5%
3	HunyuanImage 3.0	0.499	[0.352, 0.648]	[2, 9]	32.3%	69.6%
4	FLUX.2 Klein 4B	0.506	[0.351, 0.678]	[2, 9]	27.3%	56.8%
5	Qwen Image	0.510	[0.375, 0.680]	[2, 9]	24.6%	55.5%
6	GPT Image 2	0.522	[0.403, 0.660]	[2, 9]	13.6%	42.8%
7	GPT Image 1	0.532	[0.423, 0.703]	[2, 9]	12.7%	27.7%
8	Nano Banana 2	0.537	[0.411, 0.692]	[3, 9]	8.1%	29.3%
9	HiDream I1 Full	0.541	[0.403, 0.678]	[3, 9]	5.8%	26.8%
10	GPT Image 1.5	0.691	[0.520, 0.883]	[9, 11]	0.0%	0.0%
11	ERNIE Image	0.829	[0.614, 1.079]	[10, 12]	0.0%	0.0%
12	HiDream-O1	0.908	[0.656, 1.165]	[11, 12]	0.0%	0.0%

Table 12: **Awe axis: bootstrap uncertainty** (1,000 replicates), after excluding the 25 manually confirmed cropped L2 generations. Score intervals resample each model's retained paired subject-ratio/depth observations and the final 319-image human reference; rank stability uses a shared prompt-key and reference resample across models, a confirmed crop being omitted only for its affected model.

Romance. Table 13 gives the per-model intervals for the romance composite. This is the most decided axis of the six: four ranks are exact or nearly so—GPT Image 2 first and Z-Image Turbo eleventh in every replicate, GPT Image 1.5 second and ERNIE Image last in essentially all—and the top four are top-five in 100% of draws. The scale of the scores does the separating: the panel spans an order of magnitude (0.32 to 2.87), gaps the resampling noise cannot bridge. Residual uncertainty concentrates in the 5–10 region, where rank CIs overlap without ever reaching the podium.

		Score		Rank stability		
Rank	**Model**	**Avg.** W_1/IQR	**95% CI**	**Rank CI**	**P(top 3)**	**P(top 5)**
1	GPT Image 2	0.319	[0.289, 0.449]	[1, 1]	100%	100%
2	GPT Image 1.5	0.497	[0.441, 0.622]	[2, 2]	99.9%	100%
3	Qwen Image	0.645	[0.611, 0.777]	[3, 4]	64.4%	100%
4	GPT Image 1	0.703	[0.568, 0.886]	[3, 4]	35.4%	100%
5	Nano Banana 2	0.969	[0.759, 1.267]	[5, 7]	0.3%	82.6%
6	FLUX.2 Klein 4B	1.134	[0.964, 1.373]	[5, 8]	0%	10.8%
7	HiDream I1 Full	1.273	[0.961, 1.746]	[5, 10]	0%	5.9%
8	HiDream-O1	1.421	[1.189, 1.668]	[7, 9]	0%	0%
9	HunyuanImage 3.0	1.443	[1.047, 1.857]	[6, 10]	0%	0.7%
10	Cosmos3 Super T2I	1.795	[1.409, 2.208]	[9, 10]	0%	0%
11	Z-Image Turbo	2.227	[1.860, 2.585]	[11, 11]	0%	0%
12	ERNIE Image	2.873	[2.285, 3.423]	[12, 12]	0%	0%

Table 13: **Romance axis: bootstrap uncertainty** (two-level, 1,000 replicates, prompts and reference resampled jointly for all models). Left: scores with 95% CIs; right: rank stability under shared resampling.

Social group. Table 14 gives the per-model intervals. This axis barely ranks: the panel is compressed into 0.22–0.65—every model relatively close to the human spacing—and seven of the twelve rank CIs include both rank 1 and rank 7 or worse. Only the bottom pair separates reliably from the top, and no adjacent pair separates at all. The axis is best read as the benchmark's positive control: models broadly *can* match human interpersonal spacing when the composition requires no viewpoint decision, and the axis verifies it rather than ranking them.

		Score		Rank stability		
Rank	**Model**	**Avg.** W_1/IQR	**95% CI**	**Rank CI**	**P(top 3)**	**P(top 5)**
1	HiDream-O1	0.224	[0.177, 0.395]	[1, 7]	64.9%	91.5%
2	ERNIE Image	0.238	[0.168, 0.353]	[1, 7]	69.6%	90.3%
3	Cosmos3 Super T2I	0.274	[0.163, 0.510]	[1, 8]	38.2%	62.4%
4	GPT Image 1.5	0.275	[0.206, 0.366]	[1, 8]	40.7%	79.1%
5	GPT Image 1	0.279	[0.209, 0.411]	[1, 8]	25.4%	73.5%
6	HunyuanImage 3.0	0.292	[0.173, 0.423]	[1, 10]	40.2%	60.7%
7	FLUX.2 Klein 4B	0.356	[0.177, 0.623]	[1, 11]	15.8%	27.2%
8	Nano Banana 2	0.385	[0.260, 0.608]	[4, 11]	2.3%	9.1%
9	Z-Image Turbo	0.464	[0.264, 0.740]	[4, 12]	2.4%	5.1%
10	Qwen Image	0.502	[0.332, 0.743]	[7, 12]	0.3%	0.9%
11	GPT Image 2	0.531	[0.392, 0.772]	[8, 12]	0%	0%
12	HiDream I1 Full	0.649	[0.426, 0.939]	[9, 12]	0.2%	0.2%

Table 14: **Social group axis: bootstrap uncertainty** (1,000 replicates; 1,006 human references, approximately 140–150 valid observations per model). Left: scores with 95% CIs; right: rank stability under shared resampling.

Power and vulnerability. Table 16 gives pass rates, ranks, and paired lifts; Table 15 the full angle-class distributions behind them. The two axes behave very differently under resampling. On power, the ranking is reliable at both ends: GPT Image 2 is first in every replicate (rank CI [1, 1]), the bottom three never leave the tail, and naming power raises the low-angle rate significantly for seven of twelve models–the effect is small, but we can tell reliably which models respond most. On vulnerability, no ranking is possible: all models fail about equally (every pass-rate interval lies within [0.4%, 11%], and all overlap), so the axis shows that the whole panel fails, not which model fails least. The class distributions explain both results: the camera stays at eye level in around 90% of power generations and 96% of vulnerability generations, and when models do tilt, they almost never tilt the wrong way–at most 3% of generations are shot from above for power, or from below for vulnerability. The failure is an absence of tilt, not a wrong one.

	Power (pass = low)			**Vulnerability** (pass = high)		
Model	High	Eye	**Low**	**High**	Eye	Low
Cosmos3 Super T2I	2.7	81.3	**15.3**	**6.0**	92.7	1.3
ERNIE-Image	2.0	98.0	**0.0**	**6.0**	94.0	0.0
FLUX.2 Klein 4B	0.0	92.7	**7.3**	**3.3**	96.7	0.0
GPT-Image-1	0.0	93.3	**6.7**	**4.1**	95.2	0.7
GPT-Image-1.5	0.0	87.3	**12.7**	**2.1**	97.9	0.0
GPT-Image-2	0.7	71.3	**28.0**	**2.1**	97.9	0.0
HiDream-I1 Full	2.0	91.3	**6.7**	**2.7**	96.7	0.7
HiDream-O1	1.3	98.0	**0.7**	**1.3**	98.7	0.0
HunyuanImage-3.0	0.7	90.7	**8.7**	**4.0**	94.0	2.0
Nano-Banana-2	2.7	96.0	**1.3**	**2.1**	97.9	0.0
Qwen-Image	0.7	88.0	**11.3**	**1.3**	96.0	2.7
Z-Image Turbo	0.0	91.3	**8.7**	**4.7**	94.0	1.3

Table 15: **Camera-angle class distribution per model** (% of valid generations), estimated by Puffin at the calibrated $\pm 9^\circ$ threshold. The **bold** column is the target class for each axis; **red** the best, orange the second.

	Power (pass = low angle)			**Vulnerability** (pass = high angle)	
Model	**Pass rate [95% CI]**	**Rank [CI]**	**Lift L2−L1**	**Pass rate [95% CI]**	**Rank [CI]**
GPT-Image-2	28.0% [21.4, 35.7]	1 [1, 1]	+24.0*	2.1% [0.7, 5.9]	8.5 [3, 12]
Cosmos3 Super T2I	15.3% [10.4, 22.0]	2 [2, 4.5]	+10.7*	6.0% [3.2, 11.0]	1.5 [1, 8.5]
GPT-Image-1.5	12.7% [8.3, 18.9]	3 [2, 6.5]	+8.0*	2.1% [0.7, 5.9]	11 [5, 12]
Qwen Image	11.3% [7.2, 17.4]	4 [2, 7]	+7.3*	1.3% [0.4, 4.7]	11 [5, 12]
HunyuanImage-3.0	8.7% [5.1, 14.3]	5.5 [3, 9]	+3.3	4.0% [1.8, 8.5]	5.5 [1, 11]
Z-Image Turbo	8.7% [5.1, 14.3]	5.5 [3, 9]	+4.0*	4.7% [2.3, 9.3]	3.5 [1, 10]
FLUX.2 Klein 4B	7.3% [4.1, 12.7]	7 [4, 9]	+4.7*	3.3% [1.4, 7.6]	5.5 [1.5, 10]
GPT-Image-1	6.7% [3.7, 11.8]	8.5 [4.5, 9]	+2.7	4.1% [1.9, 8.6]	3.5 [1, 10]
HiDream-I1 Full	6.7% [3.7, 11.8]	8.5 [4, 9]	+4.0*	2.7% [1.0, 6.7]	7 [2, 12]
Nano Banana 2	1.3% [0.4, 4.7]	10 [10, 11.5]	+1.3	2.1% [0.7, 5.9]	8.5 [4, 12]
HiDream-O1	0.7% [0.1, 3.7]	11 [10, 11.5]	−0.7	1.3% [0.4, 4.7]	11 [5.5, 12]
ERNIE Image	0.0% [0.0, 2.5]	12 [11, 12]	−0.7	6.0% [3.2, 11.0]	1.5 [1, 8]

Table 16: **Camera-angle axes: uncertainty quantification.** Pass-rate CIs are Wilson score intervals over each model's available generations (power: $n = 150$; vulnerability: $n = 144$–150, as four API models decline 3–6 prompts on content-policy grounds). Rank CIs use a 1,000-replicate bootstrap with draws shared across models (vulnerability ranks on the 139-prompt intersection); the paired L2−L1 lift (percentage points, same contexts) is starred when its 95% CI excludes zero. Rows sorted by power pass rate.

Relative scale. Table 17 gives the per-regime pass rates, and Table 18 their Wilson intervals and the overall rank stability. The overall ranking is largely undecided: the top four are statistically tied (43.8–46.3%, mutually overlapping CIs), the middle is interchangeable, and only Cosmos3 separates at the bottom ([9.5, 12])—on this axis, the panel fails together more than it ranks. What the intervals do decide is the structure *within* the axis. The micro regime is a wall: even the best model (GPT Image 2) has an upper confidence bound of 36%, and two models fail on all 69 micro pairs. And regime competence dissociates: each regime has a different leader—GPT Image 2 on micro, HunyuanImage on extreme, Z-Image on large, HiDream-O1 on moderate—and each leader ranks in the bottom half of at least one other regime. The axis mirrors the benchmark as a whole: just as no model leads every axis, none leads every regime.

Model	Micro (n=69)	Moderate (n=69)	Large (n=69)	Extreme (n=65)	Overall (n=272)	Rank
Z-Image Turbo	0.058	0.696	**0.696**	0.400	**0.463**	1
GPT Image 1	0.014	0.870	0.623	0.308	0.456	2
GPT Image 2	**0.246**	0.783	0.507	0.277	0.456	2
FLUX.2 Klein 4B	0.087	0.812	0.551	0.292	0.438	4
Qwen Image	0.000	0.754	0.551	0.431	0.434	5
Nano Banana 2	0.130	0.754	0.565	0.246	0.426	6
ERNIE Image	0.043	0.855	0.522	0.215	0.412	7
HiDream-O1	0.130	**0.928**	0.449	0.123	0.412	7
GPT Image 1.5	0.014	0.797	0.623	0.185	0.408	9
HiDream I1 Full	0.043	0.667	0.565	0.338	0.404	10
HunyuanImage 3.0	0.000	0.623	0.493	**0.508**	0.404	10
Cosmos3 Super	0.058	0.652	0.478	0.231	0.357	12

Table 17: **Relative-scale pass rate by ratio regime.** Fraction of generations whose depth-corrected size ratio falls within the accepted interval, by regime (Micro: ratio ≥ 6; Moderate: 0.6–2.2; Large: 0.15–0.6; Extreme: 0.04–0.15); higher is better. Rows sorted by overall rank. **Red**: best per column; orange: second.

	Per-regime pass rate [Wilson 95% CI]				Overall	Rank
Model	Micro	Extreme	Large	Moderate	[95% CI]	Mean [95% CI]
Z-Image Turbo	5.8 [2.3, 14.0]	40.0 [29.0, 52.1]	69.6 [57.9, 79.2]	69.6 [57.9, 79.2]	46.3 [40.5, 52.3]	2.68 [1.0, 7.5]
GPT-Image-1	1.4 [0.3, 7.8]	30.8 [20.9, 42.8]	62.3 [50.5, 72.8]	87.0 [77.0, 93.0]	45.6 [39.8, 51.5]	3.09 [1.0, 7.5]
GPT-Image-2	**24.6 [16.0, 36.0]**	27.7 [18.3, 39.6]	50.7 [39.2, 62.2]	78.3 [67.2, 86.4]	45.6 [39.8, 51.5]	3.17 [1.0, 8.0]
FLUX.2 Klein 4B	8.7 [4.0, 17.7]	29.2 [19.6, 41.2]	55.1 [43.4, 66.2]	81.2 [70.4, 88.6]	43.8 [38.0, 49.7]	5.03 [1.0, 10.5]
Qwen Image	0.0 [0.0, 5.3]	43.1 [31.8, 55.2]	55.1 [43.4, 66.2]	75.4 [64.0, 84.0]	43.4 [37.6, 49.3]	5.28 [1.0, 11.0]
Nano Banana 2	13.0 [7.0, 23.0]	24.6 [15.8, 36.3]	56.5 [44.8, 67.6]	75.4 [64.0, 84.0]	42.6 [36.9, 48.6]	6.15 [2.0, 11.0]
ERNIE Image	4.3 [1.5, 12.0]	21.5 [13.3, 33.0]	52.2 [40.6, 63.5]	85.5 [75.3, 91.9]	41.2 [35.5, 47.1]	7.71 [2.0, 11.0]
HiDream-O1	13.0 [7.0, 23.0]	12.3 [6.4, 22.5]	44.9 [33.8, 56.6]	**92.8 [84.1, 96.9]**	41.2 [35.5, 47.1]	7.91 [3.0, 11.0]
GPT-Image-1.5	1.4 [0.3, 7.8]	18.5 [10.9, 29.6]	62.3 [50.5, 72.8]	79.7 [68.8, 87.5]	40.8 [35.1, 46.7]	8.21 [3.5, 12.0]
HiDream-I1 Full	4.3 [1.5, 12.0]	33.8 [23.5, 46.0]	56.5 [44.8, 67.6]	66.7 [54.9, 76.6]	40.4 [34.8, 46.4]	8.51 [3.0, 12.0]
HunyuanImage 3.0	0.0 [0.0, 5.3]	**50.8 [38.9, 62.5]**	49.3 [37.8, 60.8]	62.3 [50.5, 72.8]	40.4 [34.8, 46.4]	8.59 [3.0, 12.0]
Cosmos3 Super T2I	5.8 [2.3, 14.0]	23.1 [14.5, 34.6]	47.8 [36.5, 59.4]	65.2 [53.4, 75.4]	35.7 [30.2, 41.5]	11.69 [9.5, 12.0]

Table 18: **Relative-scale axis: uncertainty quantification** ($n = 272$ instances per model; per-regime $n = 69/69/69/65$; pass = depth-corrected size ratio within the regime's accepted range). Pass-rate CIs are Wilson score intervals; rank CIs are percentile intervals over a 1,000-replicate bootstrap stratified by regime, with draws shared across models. Bold marks the best value per regime.

D.2 Neutral baselines, Impact of naming the emotions (L1 vs L2 prompts)

A pass rate is a bare number: knowing that a model produces low-angle framings in 8.7% of its generations says nothing until one knows what it would have produced without being asked. This section isolates the effect of the stated intent by comparing each model's scored condition (L2, the full prompt) against its neutral baseline (L1, the same sentence without the intent suffix). The comparison separates two accounts of the same low rate—a model that responds weakly to the construct, and a model whose output simply reflects its default framing behaviour—and quantifies how much of the measured behaviour the intent is responsible for.

Power. The L1 condition removes the intent suffix and leaves the neutral sentence alone ("a judge in a courtroom"), which gives no reason to tilt the camera in either direction. Under L1, low-angle framings occur in 0–5.3% of generations (median 4.0%): eye-level is not a failure to comply with the prompt, it is the default state of the generator.

Against this baseline, naming the construct has a small but real effect. The paired lift L2 − L1, computed over matched contexts and reported in Table 19, is significantly positive for seven of the twelve models; for the remaining five it is statistically indistinguishable from zero. The median lift is under five percentage points—naming power moves the camera in fewer than one generation in twenty. The exception is GPT-Image-2, which rises from 4.0% to 28.0% ($+24.0$ points): the one model that responds substantially to the construct still leaves the camera level in seven generations out of ten.

	Power (pass = low angle)			
Model	**L1 pass rate** *(none)*	**L2 pass rate [95% CI]** *(powerful figure)*	**Rank [CI]**	**Lift L2−L1**
GPT-Image-2	4.0%	28.0% [21.4, 35.7]	1 [1, 1]	$+24.0^{*}$
Cosmos3 Super T2I	4.7%	15.3% [10.4, 22.0]	2 [2, 4.5]	$+10.7^{*}$
GPT-Image-1.5	4.7%	12.7% [8.3, 18.9]	3 [2, 6.5]	$+8.0^{*}$
Qwen Image	4.0%	11.3% [7.2, 17.4]	4 [2, 7]	$+7.3^{*}$
HunyuanImage-3.0	5.3%	8.7% [5.1, 14.3]	5.5 [3, 9]	$+3.3$
Z-Image Turbo	4.7%	8.7% [5.1, 14.3]	5.5 [3, 9]	$+4.0^{*}$
FLUX.2 Klein 4B	2.7%	7.3% [4.1, 12.7]	7 [4, 9]	$+4.7^{*}$
GPT-Image-1	4.0%	6.7% [3.7, 11.8]	8.5 [4.5, 9]	$+2.7$
HiDream-I1 Full	2.7%	6.7% [3.7, 11.8]	8.5 [4, 9]	$+4.0^{*}$
Nano Banana 2	0.0%	1.3% [0.4, 4.7]	10 [10, 11.5]	$+1.3$
HiDream-O1	1.3%	0.7% [0.1, 3.7]	11 [10, 11.5]	-0.7
ERNIE Image	0.7%	0.0% [0.0, 2.5]	12 [11, 12]	-0.7

Table 19: **Power camera-angle axis: baseline, prompted performance, and uncertainty.** L1 omits the power-related suffix, whereas L2 adds "a powerful figure." L2 pass-rate CIs are Wilson score intervals over $n = 150$ generations per model, and rank CIs use a 1,000-replicate bootstrap with draws shared across models. The paired L2−L1 lift is reported in percentage points for matched contexts and is starred when its 95% CI excludes zero. Rows are sorted by L2 pass rate.

Awe. For the distributional axes the neutral condition plays the same role, compared distribution-wise rather than through a threshold. Figure 18 shows, per model, the median subject ratio and scene depth under L1 (context only) and L2 (context plus suffix). Two asymmetries emerge. On subject size, naming the emotion mostly acts in the right direction: seven of twelve models move their median toward the human value, some substantially (GPT Image 2 drops from 0.9% to 0.4%; GPT Image 1 from 3.5% to 2.0%)–and under a paired per-prompt comparison, the subject shrinks in 42 of 50 contexts (Wilcoxon $p < 10^{-8}$). The movement is real but insufficient: the four largest-subject models remain 3–6× above the human median after the shift. On scene depth, by contrast, the suffix is not what opens the scene: depth medians already sit beyond the human value under L1 alone for nine of twelve models (52–77 m), and the suffix pushes nine of them deeper still. The deep scene is not a response to the emotion; it is the default these prompts elicit. Together the two panels sharpen the main-text split: what the suffix buys is a partial correction of the subject, while the scene was never the suffix's doing.

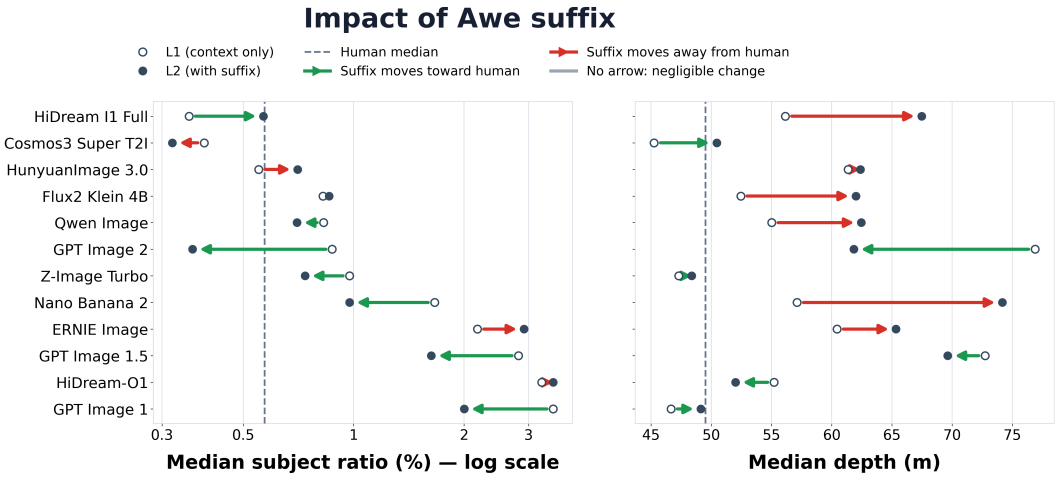

Figure 18: **Effect of the awe suffix (L1 → L2) on median subject ratio (left, log scale) and scene depth (right).** Arrows point from L1 to L2; green when the median moves toward the human reference (dashed line), red when it moves away; no arrow when the change is negligible.

	Median subject ratio (%)			Median depth (m)		
Model	**L1**	**L2**	**Suffix effect**	**L1**	**L2**	**Suffix effect**
HiDream-I1 Full	0.355	0.567	toward	56.1	67.5	away
Cosmos3 Super T2I	0.391	0.320	away	45.2	50.4	toward
HunyuanImage 3.0	0.551	0.703	away	61.4	62.4	away
FLUX.2 Klein 4B	0.824	0.856	negligible	52.4	62.0	away
Qwen Image	0.827	0.700	toward	55.0	62.5	away
GPT Image 2	0.874	0.363	toward	76.9	61.8	toward
Z-Image Turbo	0.974	0.736	toward	47.3	48.4	toward
Nano Banana 2	1.662	0.974	toward	57.1	74.2	away
ERNIE Image	2.175	2.917	away	60.4	65.3	away
GPT Image 1.5	2.814	1.629	toward	72.7	69.6	toward
HiDream-O1	3.257	3.498	away	55.2	52.0	toward
GPT Image 1	3.501	2.001	toward	46.7	49.1	toward
Human reference	0.57			49.5		

Table 20: **Median subject ratio and scene depth under L1 (context only) and L2 (context plus awe suffix), per model.** Values underlying Figure 18, rows in the same order. The suffix effect is toward when the L2 median is closer to the human reference than the L1 median, away otherwise; changes below 0.05 percentage points (subject) or 1 m (depth) are negligible.

D.3 SUFFIX ROBUSTNESS, TRYING DIFFERENT SUFFIXES

Each axis names its construct with a single phrase, which raises the question of whether the results characterise the construct or the particular string. We test this on power, the axis where the effect of the suffix is most directly measurable, by re-generating the full prompt set under two additional synonym suffixes: "an imposing figure" (L3) and "a dominant figure" (L4), each over the same 150 contexts and all twelve models.

The floor holds under every wording. Table 21 reports the four conditions side by side. No paraphrase lifts the panel out of the regime described in the main text: the median low-angle rate is 8.7% under the reference suffix, 11% under "imposing" and 7% under "dominant", against a 4.0% no-suffix baseline. The best model reaches 41.3% under its most favourable wording, still a minority of its generations, and the three lowest-scoring models remain at or below 2% throughout. The conclusion is a property of the construct, not of the string.

Wordings preserve the ordering. Rankings induced by the three suffixes correlate at $\rho = 0.77$–0.89 (Table 22), with the same model first and the same three last under every wording; the permutations that do occur fall in the middle of the field, which our bootstrap analysis independently shows to be interchangeable.

Model	L1 *(none)*	L2 *powerful figure*	L3 *imposing figure*	L4 *dominant figure*	L5 *commanding figure*	L11 *'low-angle' instructed*
Cosmos3 Super T2I	4.7%	15.3%	23.3%	8.7%	7.3%	71.3%
ERNIE Image	0.7%	0.0%	2.0%	0.0%	0.7%	49.3%
FLUX.2 Klein 4B	2.7%	7.3%	13.3%	10.0%	10.0%	90.7%
GPT-Image-1	4.0%	6.7%	8.0%	6.7%	6.7%	87.3%
GPT-Image-1.5	4.7%	12.7%	14.0%	10.7%	12.0%[†]	**97.3%**
GPT-Image-2	4.0%	**28.0%**	**41.3%**	**24.7%**	**16.7%**	96.7%
HiDream-I1 Full	2.7%	6.7%	10.0%	6.0%	6.0%	24.0%
HiDream-O1	1.3%	0.7%	0.7%	0.7%	0.7%	39.3%
HunyuanImage-3.0	5.3%	8.7%	8.7%	6.0%	4.0%	65.3%
Nano Banana 2	0.0%	1.3%	0.7%	1.3%	1.3%	29.3%
Qwen Image	4.0%	11.3%	12.0%	8.0%	8.0%	96.0%
Z-Image Turbo	4.7%	8.7%	16.7%	4.7%	7.3%	60.0%
Panel median	4.0%	8.0%	11.0%	6.4%	7.0%	68.3%

Table 21: **Power axis: low-angle pass rate under six prompt conditions** ($n = 150$ per cell; pass = Puffin pitch $> +9°$). L1 omits the suffix entirely; L2 ("a powerful figure") is the benchmark reference condition; and L3–L5 are semantic paraphrases. L11 explicitly requests "a low-angle shot of …" without the additional "viewed from below" instruction. Semantic wording alone produces little departure from the eye-level default, whereas the explicit camera instruction raises the panel median pass rate to 68.3%. [†]For GPT-Image-1.5 L5, one safety-blocked prompt used the minimally modified role "spokesperson" instead of "politician."

Comparison	**Spearman ρ**	**p-value**
Clean paraphrases (construct wording)		
L2 (powerful) vs. L4 (dominant)	0.840	< 0.001
L2 (powerful) vs. L5 (commanding)	0.843	< 0.001
L4 (dominant) vs. L5 (commanding)	0.931	< 0.001
Visually grounded probe		
L2 (powerful) vs. L3 (imposing)	0.893	< 0.001
L3 (imposing) vs. L4 (dominant)	0.772	0.003
L3 (imposing) vs. L5 (commanding)	0.856	< 0.001
Inferred vs. instructed		
L2 (powerful) vs. L11 (Ceiling)	0.737	0.006
L4 (dominant) vs. L11 (Ceiling)	0.844	< 0.001
L5 (commanding) vs. L11 (Ceiling)	0.846	< 0.001

Table 22: **Rank correlation between prompt conditions on the power axis** (Spearman ρ over the twelve models' low-angle pass rates, $n = 150$ generations per model and condition; ties assigned mean ranks). Correlations with L1 are not reported: floor-level rates produce extensive ties, making its ranking unstable.

D.4 Reference specificity: awe among photographed emotions

Could the pipeline, rather than the emotion, explain the reference's position? A natural objection to any curated reference is circularity: perhaps the awe reference occupies its position in compositional space not because of the emotion, but because of the chain that curated it—the scenic filtering, the single-person constraint, the geometric validity checks. If so, *any* set of photographs passed through the same chain would land near it, and a small W_1/IQR would mean little. We test this directly: we curate reference sets for other emotions through the identical filter chain—same automatic stages, same written acceptance criteria, same manual review, and no occupancy threshold at any point (subject size is measured, never selected on)—changing only the EmoSet label, and score them against the awe reference with the same statistic used for models. Of the remaining EmoSet categories, three survive with usable counts: contentment ($n = 141$), the principal contrast (same positive valence as awe, opposite arousal, and the opposite predicted composition—an intimate subject in a close scene); sadness ($n = 55$); and excitement ($n = 102$); amusement yields seven images and is discarded. The three are also pooled ($n = 298$) into a label-agnostic non-awe contrast, resampled with stratification by emotion.

Where does non-awe photography land? At the opposite end of compositional space (Table 23, Fig. 19). The pooled contrast scores $W_1/\mathrm{IQR} = 1.71$ [1.05–2.30] on subject occupancy—above every evaluated model (0.51–1.68, Table 1), and more than three times the best. The identical pipeline, applied to different emotion labels, does not reproduce the awe reference's position: that position is driven by the emotion, not by the curation.

Is awe merely different from the chosen control? No—it is an outlier among the emotions themselves. The three contrast emotions *converge*: median subject ratios of 2.2–3.0%, mutually within confidence intervals, despite covering opposite valences and arousal levels (positive-calm, negative, positive-aroused). The band they share is a property of neither valence nor arousal; awe alone sits outside it.

Which component carries the separation? The subject, not the scene. Subject occupancy separates every set from the reference (1.38–1.82); scene depth barely separates at all (0.07–0.46)—by construction, since the scenic filter constrains all sets to wide outdoor scenes, equalizing available depth. Sadness is the sharpest case: its scene depth *matches* the awe reference (median 51.9 vs. 49.5 m; $W_1/\mathrm{IQR} = 0.07$ [0.011–0.205], a distance indistinguishable from zero) yet its subjects remain four times too large—the small self, not the deep scene, is awe's distinctive mark. Within the domain the pipeline holds constant, the emotions diverge exactly where the photographer retains freedom: the placement of the figure. It is the same hierarchy the benchmark finds in models—scene properties shared, framing distinctive—now visible within human photography itself.

		Subject ratio		Depth p98 (m)	
Set	N	**med. [Q1–Q3]**	$\mathbf{W_1}$/IQR **[95% CI]**	**med. [Q1–Q3]**	$\mathbf{W_1}$/IQR **[95% CI]**
Awe reference ($K = 1$)	319	0.57 [0.16–1.57]	—	49.5 [28.1–82.0]	—
Human contrast pool	298	2.86 [1.38–5.62]	1.74 [1.09–2.32]	37.3 [19.8–64.4]	0.23 [0.14–0.32]
Contentment	141	2.90 [1.36–5.84]	1.82 [1.15–2.65]	41.5 [26.3–64.6]	0.17 [0.09–0.29]
Sadness	55	2.23 [1.48–4.22]	1.38 [0.75–2.18]	51.9 [26.8–78.0]	0.07 [0.011–0.205]
Excitement	102	3.02 [1.36–5.63]	1.82 [1.11–2.62]	25.0 [12.9–43.8]	0.46 [0.34–0.59]

Table 23: **Specificity of the Awe reference relative to other photographed emotions.** Wasserstein distances are normalized by the corresponding interquartile range of the Awe reference. Depth p98 is computed using DA3Nested-Giant-Large-1.1 after excluding the harmonized DA3/SAM3 sky-mask union for both reference and contrast images. Confidence intervals are percentile intervals from 1,000 bootstrap replicates (seed 42). Resampling of the pooled contrast is stratified by emotion, preserving its 141/55/102 composition.

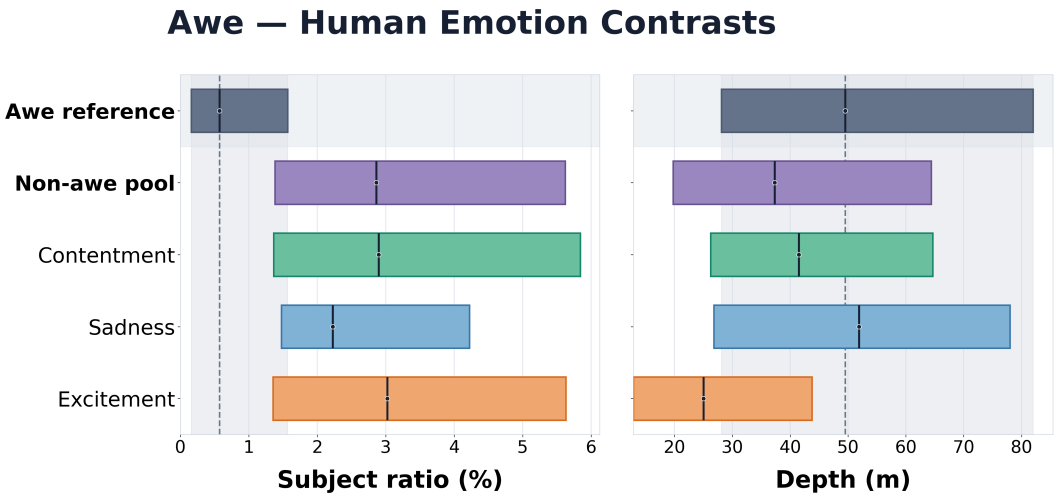

Figure 19: **Human emotion contrasts on the awe axes.** Interquartile ranges and medians of subject occupancy (left) and far-scene depth (right) for the awe reference, the pooled non-awe contrast, and its three component emotions, all curated by the identical pipeline. Dashed lines mark the awe reference medians; shaded bands its interquartile range.