NEGT2IBENCH: WHEN NEGATION CHANGES THE PICTURE. A POLARITY BENCHMARK FOR TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) models are judged by benchmarks that measure whether requested content appears, but these benchmarks largely overlook the complementary ability to satisfy negated constraints, for example, generating "a non-red cup." Measuring negation raises challenges not faced by affirmation-based benchmarks and requires careful prompt and evaluation design. We introduce NEGT2IBENCH, a benchmark of 4,800 prompts covering two attribute types and four relation categories. Prompts are organized by *polarity*: the number of *positive* statements that must hold and *negated* statements that must not, each ranging from 0 to 2. Varying the two independently separates the effect of negation from the effect of prompt complexity. Our detector-based scoring is reproducible, auditable, and pinpoints which requirement failed. On 600 images with three-annotator labels, it agrees with humans as closely as vision-language judges up to $30\times$ larger, while using only a fraction of their GPU memory. Across eleven T2I models and 211,200 images, nine score lower on a single negated statement than on a single positive one. Per-statement scoring reveals that the loss is largest for color and near zero for proximity, and that 41.5% of failed statements render exactly what the prompt forbids. Rendering what a prompt asks for and withholding what it forbids are distinct capabilities that an aggregate compositional score cannot distinguish. NEGT2IBENCH measures the latter directly, providing a controlled testbed for diagnosing negation failures and developing methods to overcome them.

1 INTRODUCTION

While text-to-image (T2I) models now render increasingly complex prompts accurately enough for professional (Shabani et al., 2023) and creative (Ko et al., 2022) applications, evaluating whether they respect what a prompt *forbids* is surprisingly difficult and underexplored. The frontier is judged by benchmarks like GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023), and DPG-Bench (Hu et al., 2024), which focus on challenging, dense prompts with multiple components that must each be correct. Though these benchmarks cover many important prompt-following aspects, they largely overlook an important capability: *negation*.

Measuring negation raises challenges not faced by affirmation-based benchmarks, making it more nuanced than simply adding negated prompts to existing benchmarks. First, success can be accidental: cups rarely fly, so "a cup that is not flying" can produce a valid image even if *not* was ignored. Second, if object presence is not verified, a negated attribute can be satisfied by omitting the object rather than the attribute: "a non-wooden bench" may simply omit the bench. Third, negation inherits the difficulty of dense prompts; adding objects, like "a red cup" vs. "a red cup with no spoon" introduces multiple potential failures, so failures from negation must be disentangled from those from prompt complexity. Fourth, the evaluator can also introduce errors: when checking for forbidden elements, vision-language models may report absent content (Li et al., 2023) or hallucinate (Xiao et al., 2026), making evaluator errors difficult to distinguish from image-generation errors.

We introduce NEGT2IBENCH, a benchmark for measuring negation in text-to-image generation, designed around these four challenges. To reduce accidental success, an LLM (Yang et al., 2025) selects plausible attributes and relations, and fixed templates keep wording consistent across con-

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.

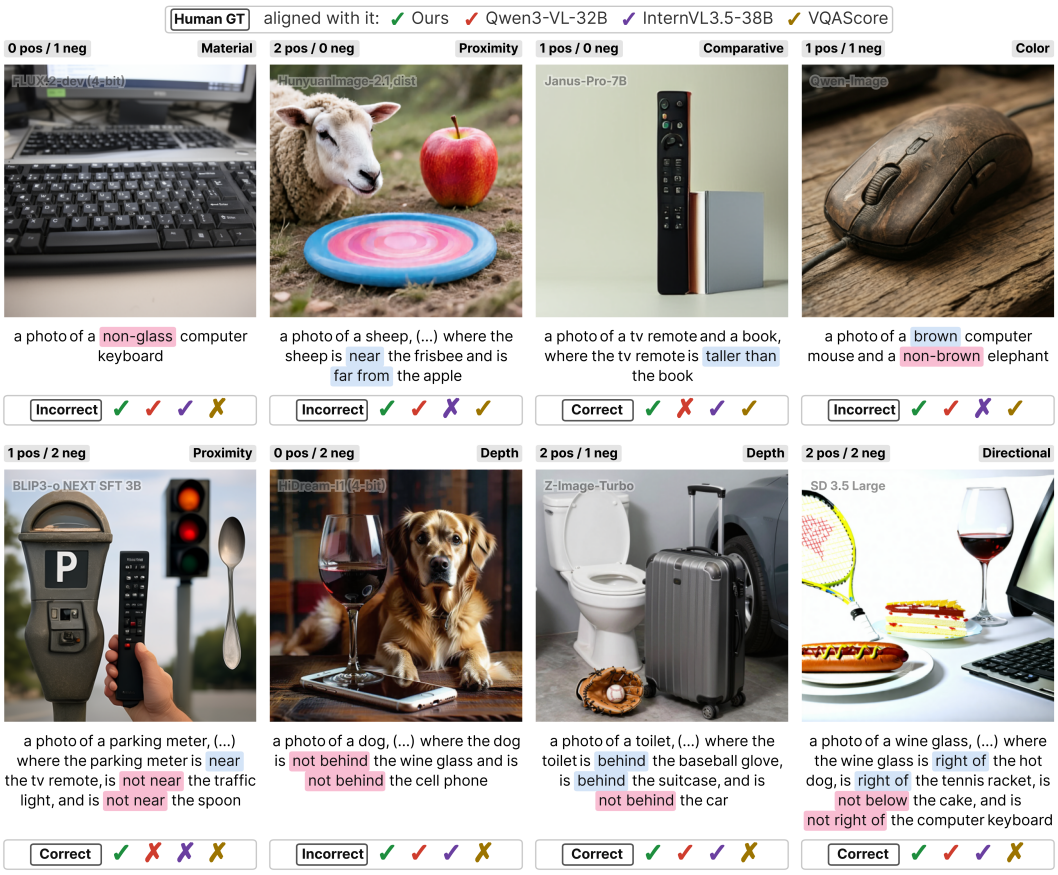


Figure 1: NEGT2IBENCH prompts mix positive (blue) and negated (pink) statements, spanning six categories (above image, right) and organized by the number of each statement (above image, left). Images come from the benchmarked T2I models. The box under each image gives the human verdict; marks show whether each evaluator agrees (✓) or disagrees (✗) with it.

ditions. To prevent success by omission, evaluation requires every named object to appear: i.e., "a non-wooden bench" still requires a bench. To isolate negation from prompt complexity, prompts are organized by *polarity*, the number of *positive* statements that must hold and *negated* statements that must not; prompts with the same statement count contain the same object count and differ only in polarity. Finally, to limit evaluator error, each requirement is checked by a dedicated perception component rather than by an LLM judge, revealing which requirement failed. NEGT2IBENCH covers two attribute categories (color and material) and four relation categories (directional, proximity, comparative, and depth).

We evaluate eleven T2I models on the full corpus, four samples per prompt totaling 211,200 images (Section 5). Nine of eleven models score lower on a single negated statement than on a single positive one, by up to 43.8% points. The drop differs by category: color falls from 85.0% to 31.3%, while proximity barely moves, 45.8% vs. 43.2%. At the highest tested complexity, no model exceeds 8.4%. More strikingly, 41.5% of failed statements are negated violations: the image shows the attribute or relation the prompt forbids. Ranking also changes with polarity: FLUX.2-dev leads on the single positive statement and Janus-Pro-7B on the negated one, so a single score hides the effect. Replacing SD 1.5's text encoder with negation-aware NegationCLIP (Park et al., 2025) improves accuracy only from 14.5% to 14.9%. Negation is therefore a failure mode of its own, and one that an aggregate score cannot reveal.

Contributions. (1) **NEGT2IBENCH**, a negation benchmark for T2I models with $4{,}800$ prompts over 79 object classes, six categories, and eight polarity settings, scored using only detection and geometry. Whereas prior negation evaluation for generation focuses on negated objects,

NEGT2IBENCH also negates attributes and relations while requiring the corresponding objects to remain present, and compares positive and negated prompts at matched prompt complexity. (2) **An evaluator validation study** against human annotation on 600 images, comparing NEGT2IBENCH with 14 alternative scoring methods separately for positive and negated statements. (3) **A fine-grained analysis of negation failures** across eleven T2I configurations, six categories, and eight polarity settings, including matched-load polarity comparisons and a statement-level failure taxonomy that reveals where and how models fail.

2 RELATED WORK

Text-to-image models and compositional evaluation. Diffusion, flow-based, and autoregressive T2I models (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Black Forest Labs, 2026; Wu et al., 2025; Cai et al., 2025b;a; Tencent Hunyuan Team, 2025; Chen et al., 2025b;a) and closed systems (Betker et al., 2023; Baldridge et al., 2024; OpenAI, 2025) follow increasingly complex prompts. However, advances mostly show results for whether requested content appears, not whether a specified attribute or relation is withheld. Distribution metrics such as FID (Heusel et al., 2017) and alignment scores such as CLIPScore (Hessel et al., 2021) give coarse quality and prompt following scores. Like our approach, GenEval (Ghosh et al., 2023) uses detectors to check objects, counts, colors, and positions, while PosEval (Bader et al., 2025) extends its position task; NEGT2IBENCH builds on GenEval's prompt schema and PosEval's spatial encoding. Other benchmarks cover more attributes, relations, and longer prompts (Huang et al., 2023; 2025; Hu et al., 2024; Bakr et al., 2023; Gokhale et al., 2022; Kamath et al., 2025). GenAI-Bench (Li et al., 2024) and UniGenBench++ (Wang et al., 2025b) include negation as a compositional skill or a grammar category, but neither separates attribute from relation negation or pairs negated prompts with positive ones. Detector-based benchmarks also face template contamination (Xie et al., 2026), weakening detector-human agreement as models improve (Kamath et al., 2025), and sensitivity to wording (Gan et al., 2025); we therefore report human agreement by category and keep wording fixed across polarity conditions.

Negation in text-to-image generation. T2I models still render what a prompt forbids, across several languages (Conwell et al., 2024; Nadeem et al., 2024). NegGenBench (Guo & Du, 2026) contains 200 prompt pairs and evaluates negative-prompt guidance: each prompt is paired with a separate negative prompt naming the content to exclude, instead of expressing the negation in the prompt itself. LMD (Lian et al., 2024) tests 100 scene-level prompts without a named object, scoring with an open-vocabulary detector and supplying the denied object via a negative-prompt input. NegationCLIP (Park et al., 2025) evaluates 107 negated prompts with an MLLM judge, while NEGTTOI (Cai et al., 2025c) pairs a wanted object with an unwanted one in 2,000 prompts scored by GPT-4o. These studies establish that exclusion is hard, but none compares positive and negated versions of attributes and relations while requiring every named object to remain. A separate line of work steers generation away from unwanted content with negative prompts or guidance (Ho & Salimans, 2021; Liu et al., 2022; Stirling et al., 2026; Chefer et al., 2023; Desai & Vasconcelos, 2025). While steering generation with negative prompts can exclude unwanted content, it does not test whether the model follows a negation written in the prompt, which is what NEGT2IBENCH measures.

Judging negation. Image-text and preference scores such as CLIPScore (Hessel et al., 2021), BLIP-Score (Li et al., 2022b), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and HPSv2 (Wu et al., 2023) measure overall alignment, but performance on ordinary prompts does not indicate negation reliability: CLIP-family models perform near chance on negated captions (Alhamoud et al., 2025), VQA models perform much worse on negated questions (Zhang et al., 2025), and CC-Neg (Singh et al., 2025) provides negated captions for training. VQA-based methods such as TIFA (Hu et al., 2023), which uses mPLUG (Li et al., 2022a), and VQAScore (Lin et al., 2024) verify prompt requirements directly, but VQA judges, including Qwen3-VL (Bai et al., 2025a) and InternVL3.5 (Wang et al., 2025a), can hallucinate absent details (Cho et al., 2024; Xiao et al., 2026), which matters when the evaluator must confirm that an attribute or relation is absent while the objects are present. Sammani et al. (2026) propose a decomposed MLLM judge for negated image retrieval, but in our validation such judges lose up to 18.8 κ points when negations are phrased as negations (Section 4). We therefore score each element with dedicated perception components and report human agreement by category and polarity.

Table 1: **NEGT2IBENCH categories.** Each category has 800 prompts (100 prompts per polarity setting, 8 settings, 4,800 prompts over 79 object classes). Each row shows a matched pair from the benchmark: the same objects and value, differing only in polarity (positive, negated). Examples omit the shared prefix: "a photo of" for attributes, and "a photo of a ⟨subject⟩ and a ⟨target⟩, where" for relations.

Category	Values	Positive example	Negated example
Attributes			
Color (10)	red, orange, yellow, green, blue, purple, pink, brown, black, white	a red cup	a non-red cup
Material (4)	wooden, metal, plastic, glass	a glass cup	a non-glass cup
Relations			
Directional (4)	left of, right of, above, below	the backpack is above the parking meter	the backpack is not above the parking meter
Proximity (2)	near, far from	the backpack is far from the parking meter	the backpack is not far from the parking meter
Comparative (4)	bigger than, smaller than, taller than, shorter than	the backpack is bigger than the parking meter	the backpack is not bigger than the parking meter
Depth (2)	in front of, behind	the backpack is in front of the parking meter	the backpack is not in front of the parking meter

3 THE NEGT2IBENCH BENCHMARK

Task formulation. NEGT2IBENCH measures how models respond when attributes and relations are negated. It tests both the effect of negation at a fixed number of statements and the effect of combining multiple statements in one prompt. Each prompt comes with metadata listing required objects and statements. A statement's polarity specifies whether an attribute or relation must hold (positive) or must not hold (negated); all named objects remain required in either case. The eight settings vary the number of positive and negated statements; we write a setting as P:N, so 1:1 contains one of each, and call P+N its load. To capture diverse forms of negation in prompts, we cover two attribute and four relation categories (Table 1), illustrated by the examples in Figure 1. Relation prompts link one shared subject to a different target for each statement, so M statements name $M + 1$ objects and changing polarity leaves both counts fixed. To support reliable automatic evaluation, we use COCO (Lin et al., 2014) object classes recognized by the detector and attributes and relations that our pipeline can verify in both polarities.

Benchmark construction. We use GenEval's color vocabulary and adapt the position rules (Ghosh et al., 2023), select material values from OVAD (Bravo et al., 2023), and follow PosEval's relation encoding (Bader et al., 2025). An LLM (Yang et al., 2025) selects combinations of objects, attributes, and relations from fixed lists. It keeps combinations that are plausible or unusual but possible to depict, so each negated statement denies something that could reasonably appear. Rule-based templates turn these selections into prompts with consistent wording. We balance different types of objects, attributes, and relations by favoring the ones that are used least often. The benchmark contains 4,800 prompts, with 100 for each combination of six categories and eight settings. Each category includes all 79 eligible object classes. Appendix A.1 gives the settings, selection rules, and templates; Appendices C.1 and A.2 give the full vocabularies and distributions.

3.1 DETECTOR-BASED EVALUATION

We check each prompt requirement separately to identify why an image fails. This could be a missing object, an incorrect count, or a violated attribute or relation. Each category uses the same verification rule (Appendix Table 14), making decisions reproducible and comparable across generators. Checking that every named object is present first prevents a model from satisfying a negated statement by omitting the object it refers to. Figure 2 illustrates this evaluation pipeline: both objects are present, but the horse is behind the computer mouse, violating the negated relation. We use RF-DETR-XL (Robinson et al., 2026) to detect objects, SAM 2 (Ravi et al., 2025) to refine their masks, and Depth-Anything V2 (Yang et al., 2024) to estimate depth for front-behind relations.

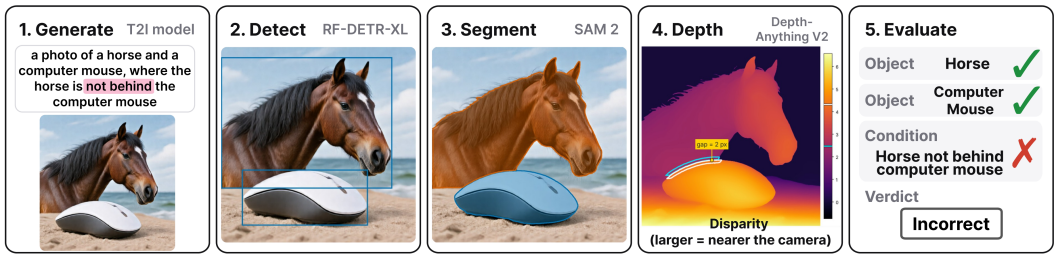


Figure 2: **NEGT2IBENCH scores a negated depth relation.** RF-DETR-XL detects the objects, SAM 2 refines their masks, and Depth-Anything V2 estimates depth. Both objects are present, so the image cannot pass by omission, but the horse is behind the computer mouse and the negated statement fails. Checking each requirement separately lets the verdict name which one failed.

Statement decisions. For color and material, SigLIP 2 (Tschannen et al., 2025) scores the masked object crop against the category's values. For color, a positive attribute passes if its score is within 0.5 of the highest score; a negated attribute passes if its score is more than 0.5 below it. This prevents a near-tied negated value from being treated as absent. For material, the top-scoring value decides both polarities. For relations, positive statements pass when their rule holds, and negated statements pass when it does not. Directional, comparative, and depth rules can return no relation when measurements are ambiguous (a dead zone). Positive statements then fail, while the negated one passes because the denied relation was not found. Table 14 defines the rules and final thresholds. Appendix C.1 gives the full equations and implementation details, and Appendix E.2 reports the frequency of ambiguous passes. Appendix D compares detectors, mask refinement, and relation thresholds on a small subset of 140 images with a clear human-majority verdict, from 600 annotated images.

Metrics. **Image accuracy** is the fraction of generated images that satisfy the full prompt: all required objects appear with valid counts, and every positive and negated statement passes. **Prompt accuracy** is the fraction of prompts for which at least one of four generated images is correct. The first measures success per image; the second measures success within four attempts. We report image accuracy by category and setting, with overall scores averaged equally across the six categories; prompt accuracy is released with the per-prompt results. Each failed statement is also assigned one of five failure causes, which Section 5 uses to diagnose where models fail; failure and metric definitions appear in Appendix C.3.

4 VALIDATION OF NEGT2IBENCH'S EVALUATION

Since NEGT2IBENCH scores images automatically, its verdicts matter only if they match human judgment. We test this with a human study and compare NEGT2IBENCH with other evaluators.

Human reference. A pool of 36 annotators labeled 600 generated images, with three labels per image, answering one yes/no question per object and per statement. We broke down each prompt into its objects and statements and phrased each statement as a question about whether it holds: for "a non-wooden bench", annotators say whether the bench is wooden, and the correct answer is "no". The majority vote on each element gives two references: a binary image label, true only if every element is correct, and a continuous score, the fraction of correct elements. Some images are ambiguous, so we also report an inter-annotator score (H. Inter-annotator), computed by scoring each annotator against the majority of the other two (Appendix B.1). Evaluators are scored against the majority of all three annotators, a less noisy reference, so they can exceed this value.

Evaluators. We compare NEGT2IBENCH (Section 3.1) with eight VQA judges: mPLUG (Li et al., 2022a), VQAScore (Lin et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3-8B (Zhu et al., 2025), Qwen3-VL-8B and 32B (Bai et al., 2025a), and InternVL3.5-8B and 38B (Wang et al., 2025a). Each judge answers yes/no questions in three modes: one question for the whole prompt (whole), the annotators' per-element questions (pos.), and the same questions with negated statements asked in negated form (neg.). The probability of the correct answer produces a continuous score, while an answer with a probability greater than 0.5 yields a binary verdict. In per-element

Table 2: **Cost and agreement with human labels** on 600 images, against the majority of three annotators. Judge modes: whole prompt (whole), per-element questions (pos), and per-element with negated statements asked in negated form (neg); Δ = neg − pos. Precision and recall use the per-element mode. H. Inter-annotator: agreement between annotators. Bold: best in column. †Timed through t2v-metrics, which scores every image against every question, so time is overestimated.

	Cost			Cohen's κ				Per element		Kendall τ			
Evaluator	Params ↓	GiB ↓	s/img ↓	whole ↑	pos ↑	neg ↑	Δ	Prec. ↑	Rec. ↑	whole ↑	pos ↑	neg ↑	Δ
InternVL3.5-38B	38B	73.0	3.29	48.4	69.5	52.6	-16.9	81.0	78.6	49.2	59.7	52.8	-6.9
Qwen3-VL-32B	32B	62.8	0.99	48.2	**74.3**	**66.5**	-7.8	**84.1**	82.0	48.6	**63.3**	**58.8**	-4.5
VQAScore†	11B	24.2	3.40	33.6	49.9	48.6	**-1.3**	68.4	65.0	38.8	42.3	42.8	**0.5**
Qwen3-VL-8B	8B	16.8	0.49	48.2	69.7	54.1	-15.6	81.7	78.2	46.4	61.0	54.8	-6.2
InternVL3-8B	8B	15.8	0.71	25.2	59.7	41.8	-17.9	74.7	71.8	46.2	56.3	51.1	-5.2
InternVL3.5-8B	8B	17.1	0.76	35.3	69.0	50.2	-18.8	77.9	82.0	43.2	55.3	49.8	-5.5
Qwen2.5-VL-7B†	7B	15.9	8.39	35.2	52.8	44.3	-8.5	79.6	56.8	44.6	43.7	44.2	**0.5**
mPLUG	**0.6B**	6.5	**0.32**	7.9	16.0	10.1	-5.9	67.9	17.5	21.1	-4.3	-5.5	-1.2
NEGT2IBENCH (ours)	1.24B	**6.3**	0.78	**72.0**	–	–	–	79.2	**85.0**	**65.1**	–	–	–
H. Inter-annotator	–	–	–	65.6	–	–	–	–	–	71.1	–	–	–

mode, an image passes only if every element receives the correct answer, similar to human labeling. NEGT2IBENCH runs one detector pass per image, so it has a single mode. We also test six image-text similarity and reward scores, which have no binary form (see Appendix B.3).

Agreement with humans. Table 2 reports Cohen's κ (Cohen, 1960), precision and recall against the binary label (per-element mode for judges), and Kendall τ (Kendall, 1938) against the continuous score. NEGT2IBENCH reaches $\kappa = 72.0$, above the human inter-annotator score (65.6), and $\tau = 65.1$, above the best judge, Qwen3-VL-32B (63.3; bootstrap intervals overlap, Table 7). On κ, only Qwen3-VL-32B scores higher (74.3), while the other seven judges score below NEGT2IBENCH on both measures. Size adds little: scaling Qwen3-VL from 8B to 32B gains 4.6 points of κ, and InternVL3.5 from 8B to 38B gains 0.5. Similarity and reward scores reach at most $\tau = 27.7$ and cannot tell whether a negated requirement holds (Table 7). In Table 2, NEGT2IBENCH has the highest recall of any evaluator (85.0 vs. 82.0 for the best judge) and trails it on precision (79.2 vs. 84.1), so it rejects fewer human-accepted images than any judge: the failures we report are not an artifact of a strict evaluator (Appendix B.5).

Negative phrasing effect. When we ask negated statements in negated form ("is the bench non-wooden?") and keep all other questions the same (Δ in Table 2), all judges lose 1.3 to 18.8 points of κ, and five of them lose 4.5 to 6.9 points of τ, including the largest models. Judge verdicts thus depend on wording, the very variable a negation benchmark changes. NEGT2IBENCH checks each requirement directly instead of answering a question, so wording does not affect it, and it beats every judge under whole prompt and negated phrasing (Appendix B.6).

Cost. The NEGT2IBENCH detector-based evaluator uses 1.24B parameters and scores an image in 0.78 s with 6.3 GiB of GPU memory, measured on one GPU with one process and no batching. Judges with similar agreement use 8B to 38B parameters, 2.7 to 11.6 times more memory (16.8 to 73.0 GiB), and 0.49 - 3.29 s per image, measured per question and multiplied by the mean of 4.66 questions per image. Their cost also grows with the number of elements, since they ask one question per element, while NEGT2IBENCH runs its detector once per image (Appendix B.4). Its components can also be replaced as better detectors appear (Appendix D).

5 BENCHMARKING TEXT-TO-IMAGE MODELS

We evaluate eleven T2I models on the full NEGT2IBENCH corpus: BLIP3-o NEXT SFT 3B (Chen et al., 2025a), FLUX.2-klein-9B and FLUX.2-dev (4-bit) (Black Forest Labs, 2026), HiDream-I1 (4-bit) (Cai et al., 2025b), HunyuanImage-2.1 (distilled) (Tencent Hunyuan Team, 2025), Janus-Pro-7B (Chen et al., 2025b), Qwen-Image (Wu et al., 2025), SD 3.5 Large (Esser et al., 2024), Z-Image-Turbo (Cai et al., 2025a), and SD 1.5 (Rombach et al., 2022) with and without negation-aware text encoder NegationCLIP (Park et al., 2025). Each configuration produces four samples for 4,800 prompts: 19,200 images per configuration, totaling 211,200 (Appendix E.1).

Table 3: **Accuracy (%) on NEGT2IBENCH, by polarity and statement load (left) and by semantic category (right).** Even the strongest models struggle with negation, and accuracy falls sharply as load increases. P:N denotes P asserted and N negated statements; category values are averaged over all eight settings. Bold marks the best value per column.

	Statement polarity and load								Semantic category						
	Load 1		Load 2			Load 3		Load 4	Attributes		Relations				
Model	1:0	0:1	2:0	1:1	0:2	2:1	1:2	2:2	Color	Material	Compar.	Depth	Direct.	Proxim.	Overall
U-Net diffusion															
SD 1.5	36.1	44.1	7.3	9.7	14.0	1.8	2.4	0.5	27.7	22.8	8.9	8.8	9.6	9.1	14.5
SD 1.5 + NegationCLIP	35.7	44.3	8.8	9.6	15.5	1.8	3.0	0.5	28.3	23.9	8.4	9.4	10.0	9.2	14.9
Diffusion / flow transformers															
HiDream-I1 (4-bit)	54.8	43.9	27.7	17.2	24.1	11.0	9.2	4.1	31.2	28.4	18.9	21.7	23.9	20.0	24.0
Z-Image-Turbo	63.7	31.4	41.5	19.9	14.3	12.4	9.0	4.9	24.9	25.8	21.4	23.0	28.6	24.0	24.6
SD 3.5 Large	57.9	47.5	29.3	17.8	23.8	8.7	8.2	3.8	31.4	36.4	18.5	19.7	22.5	19.3	24.6
Qwen-Image	67.5	35.2	38.3	21.0	19.1	11.9	10.1	5.2	24.2	28.1	24.3	25.5	30.3	23.8	26.0
FLUX.2-klein-9B	72.8	29.0	49.0	20.0	14.6	12.9	9.8	4.8	24.8	36.8	23.7	21.5	29.2	23.8	26.6
HunyuanImage-2.1 (distilled)	66.2	44.7	41.2	26.2	25.1	**17.9**	12.9	**8.4**	29.3	36.7	25.2	**27.6**	36.6	26.6	30.3
FLUX.2-dev (4-bit)	**76.3**	42.6	**50.1**	22.8	23.8	16.1	12.2	6.4	27.9	**44.2**	**25.4**	26.6	37.2	26.3	31.3
Autoregressive / hybrid															
Janus-Pro-7B	55.2	**54.8**	31.0	21.8	25.8	9.7	9.9	4.7	**39.8**	38.0	19.5	20.6	22.4	19.3	26.6
BLIP3-o NEXT SFT 3B	67.2	48.7	41.7	**27.8**	**28.1**	16.2	**15.0**	6.2	34.7	37.1	24.5	26.6	**38.1**	**27.1**	**31.4**

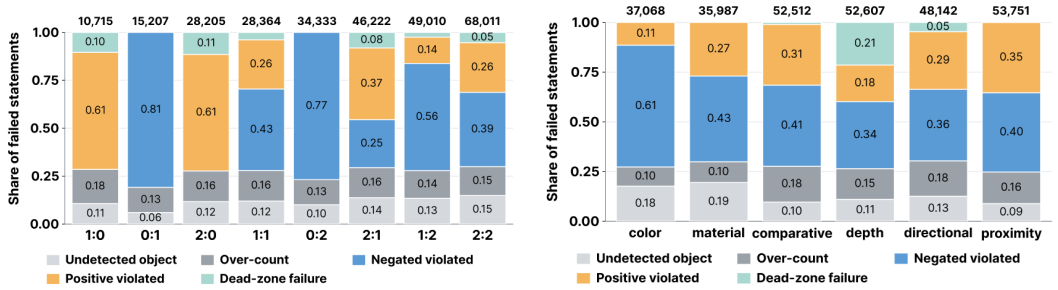
Scoring uses Section 3.1's evaluator: RF-DETR-XL (Robinson et al., 2026) at presence/counting thresholds $0.03/0.5$, SAM 2 masks (Ravi et al., 2025) for relations, SigLIP 2 (Tschannen et al., 2025) for color and material, and Depth-Anything V2 (Yang et al., 2024) for depth. Geometric cut points and attribute confusion buffer are fitted on 140 human-annotated images, then fixed for the whole benchmark (Appendix D). Prompts are sampled four times. Each model uses its native configuration: 4 to 50 steps, and default resolution from 384^2 to 1024^2. Appendix E.1 lists exact configurations.

Table 3 reports accuracy by polarity, statement load, and semantic category. The strongest model, BLIP3-o NEXT, reaches 31.4%. Scores vary by about ± 0.7 points overall and ± 2 points per subtask by chance alone (95% interval). The top three models lie within 1.1 points, so overall score cannot separate them, but subtasks can. Janus-Pro-7B ties FLUX.2-klein-9B fourth at 26.6% overall, yet first on negated single statement at 54.8%, ahead of BLIP3-o NEXT at 48.7%, and first on color at 39.8%. FLUX.2-dev (4-bit) is the opposite: first on the asserted single statement at 76.3%, and 42.6% once negated. Accuracy on one negated statement spans 29.0 to 54.8%, a wider range than overall score. Four models lead at least one subtask, and none leads more than three of eight polarity subtasks or two of six category subtasks.

Polarity and compositional load. At Load 1, nine of eleven configurations score higher on the asserted statement (1:0) than on the negated one (0:1); the exceptions are the two SD 1.5 variants. At Load 2, the same nine score highest on the all-positive setting (2:0). Across those nine, drop averages 22.7 points, from 0.4 to 43.8. Load matters on its own. Between 1:1 and 0:2 there is no consistent ordering, so after the first negation the number of statements matters as much as how many are negated. Accuracy then collapses as load grows: the best score in each group falls from 76.3% at Load 1 to 17.9% at Load 3 and 8.4% at Load 4. NEGT2IBENCH exposes two separate difficulties: negation is harder than assertion at equal load, and satisfying several constraints at once is hard whatever their polarity.

Penalty depends on category. At Load 1, the drop is largest for the attributes: 53.7 points for color, from 85.0% asserted to 31.3% denied, and 17.0 points for material. Relations lose less, 16.6 points for comparative, 14.6 for directional. Proximity barely moves, 45.8% vs. 43.2%. Depth is opposite, 41.7% vs. 44.0%. This reversal is largely explained by dead-zone passes rather than by better negation Categories with higher positive accuracy also lose more when negated (Spearman $\rho = 0.94$ across the six categories). A larger drop therefore does not by itself show that a category is harder to negate; it may partly reflect a higher starting point (see Appendix E.2 for more details).

Two caveats apply when reading these comparisons. First, the size of a model's drop does not rank its negation ability, because the drop is capped by the model's accuracy on the positive statement: FLUX.2-klein-9B scores 72.8% at 1:0 and loses 43.8 points at 0:1, while Janus-Pro-7B scores



(a) By benchmark setting **(b) By category**

Figure 3: **Statement-level failure decomposition, pooled over all models.** Each bar partitions the failed statements of one polarity setting (a) or category (b) into the five failure causes; the number above each bar is its count of failed statements. Bars show how failures divide among causes, not how often statements fail. Each failed statement is counted once; statements that were satisfied are not counted, even in failed images. Dead zones only affect positive statements in directional, comparative, and depth relations, so the other bars have no dead-zone failures.

55.2%, loses 0.4, and still reaches the highest negated accuracy. A model that is weak on the assertion has less to lose on the denial, so only the within-model comparison is safe to read. Second, NEGT2IBENCH measures whether a negated attribute or relation is absent from the image, not whether the model understood the negation. A negated statement is easier to satisfy by chance than a positive one, so a model that ignores the negation can still pass some negated statements. A control prompt that omits the tested attribute or relation would separate the two; it fits the current design and is left for future work (Appendix F).

Failure analysis. Every failed statement falls into one of five causes: undetected object, over-count, positive violation, negated violation, or dead-zone failure, defined in Appendix C.3. Of the 280,067 failed statements, negated violations are the largest group at 41.5%. Positive violations follow at 25.9%, then over-counts at 15.0%, undetected objects at 12.6%, and dead-zone failures at 5.0%. The most common failure is a model rendering what the prompt forbade. Dead zones, the only cause attributable to the evaluator alone, are the smallest at 5.0%. Within the same prompt, negated statements fail more often than positive ones (Figure 3(a)). At 1:1, the negated statement causes 43% of failures and the positive one 26%; at 1:2, the two negated statements cause 56% and the positive one 14%, twice as much per statement. Per statement, the imbalance holds in all four mixed settings, by a factor of 1.3 to 2.0. The share of negated violations also varies by category (Figure 3(b)): they account for 61% of failures for color and 43% for material, but 34 to 41% for the four relations. Negated colors are thus most often rendered, resulting in an incorrect image. Polarity therefore changes which constraint a model breaks, not only how often it breaks one. Appendix E.2 gives the per-model and row-level breakdowns.

Dead zones. Directional, comparative, and depth relations have deliberate dead zones. When the measured geometry is ambiguous, the evaluator returns no predicate rather than forcing one. This is often the correct outcome: two objects at similar depth are neither *in front of* nor *behind* each other. Proximity has no dead zone, because *near* and *far* are exact complements. Dead zones affect the two polarities differently. An asserted statement fails when its predicate is absent. A negated statement passes when the forbidden predicate is absent, including when no predicate is resolved at all. Such a pass is valid under the benchmark rule, but rests on weaker evidence than one where an alternative relation is established. Of the negated relation statements that pass, 19.0% pass only because no relation was resolved. This happens mostly for depth (55.5%, vs. 9.2% for directional and 2.9% for comparative), which explains why depth is the only category where negated statements score higher than positive ones at Load 1. Table 19 in Appendix E.2 reports exposure per category.

Negation-aware text encoders. NegationCLIP (Park et al., 2025) replaces SD 1.5's text encoder and raises overall accuracy from 14.5% to 14.9%. Both SD 1.5 variants score higher at 0:1 than at 1:0, and both stay far below the other models. The comparison is not clean: the two runs share a checkpoint and a 50-step schedule, but use guidance scales of 9.0 and 7.5. It is suggestive rather than a causal encoder test (Appendix E.2). Representing "not red" differently from "red" helps only



Figure 4: **Qualitative comparison of positive and negated prompts.** Each column shows one model and a different semantic category: proximity, material, comparative, directional, color, and depth. The top row contains single positive statements (1:0); the bottom row negates the corresponding statements (0:1), keeping the objects and tested predicates unchanged. Marks give the NEGT2IBENCH verdict (✓ pass, ✗ fail). In all six pairs, the model satisfies the positive prompt and fails the negated one.

if the generator can act on the distinction. The same applies to a classifier-free-guidance negative prompt, which steers sampling away from unwanted content without the model interpreting the negation (Ho & Salimans, 2021; Park et al., 2025; Cai et al., 2025c). Improving negation is therefore likely to require training or conditioning on negation-bearing supervision, not an encoder swap or a negative-prompt field. NEGT2IBENCH measures whether such methods work, both on attributes and relations.

Qualitative examples. Figure 4 illustrates the distinction NEGT2IBENCH is designed to measure: rendering a requested property versus excluding a forbidden one while preserving the named objects. Each column pairs a positive and a negated prompt from one of six categories, generated by the same model. By changing polarity at fixed statement and object counts, these examples isolate the prompt change of interest. The selected failures show why successful positive-prompt generation alone does not establish negation-following ability, motivating our separate evaluation of positive and negated requirements.

6 CONCLUSION

We introduced NEGT2IBENCH, a benchmark that measures whether T2I models can withhold what a prompt forbids while still rendering every object it names. It compares positive and negated prompts at matched complexity and object counts, guards against success by omission, and reduces accidental success. Its detector-based evaluator agrees with human labels comparably to much larger vision-language judges while using a fraction of their GPU memory, and, unlike these judges, does not depend on how the evaluation question is phrased. Across eleven T2I configurations and 211,200 generated images, nine of eleven score lower on a single negated statement than on a single positive one, with a loss that depends on the category and compounds as statements are combined. Polarity also changes which model leads and which error it makes. Generating what a prompt asks for and withholding what it forbids are thus distinct abilities, and an aggregate compositional score hides the second. NEGT2IBENCH measures it directly, providing a controlled testbed for future models and for methods that aim to improve negation.

AI USE STATEMENT

In this work, we used generative AI tools as part of the method: a quantized Qwen3-30B-A3B-Instruct model selects plausible object, attribute, and relation combinations from fixed vocabularies (Section 3); prompts are written by deterministic templates, and no language or multimodal model takes part in scoring. We have not used generative AI tools to generate research ideas or to produce, alter, or select experimental results. Additionally, we used generative AI tools to draft and edit the manuscript and to assist with software development. We have reviewed all AI-assisted work: every quantitative claim was checked against the recorded pipeline outputs, LLM-generated code was verified and tested for correctness by two authors, and every citation was checked against the official arXiv or publisher record. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

NEGT2IBENCH uses synthetic prompts about everyday objects from the COCO vocabulary, excluding "person". The prompts contain no personal data. The only human-subject component is the annotation study (Section 4): 36 annotators answered binary questions about 600 model-generated images of ordinary scenes. Annotators were members of our research group and student assistants employed by our institution, who performed the task as part of their paid work; each labeled 50 images on average (median 19 s per image, about 16 minutes in total). All annotators were informed of the study's purpose and agreed to participate. The study collected no personal data from annotators and posed minimal risk. Our per-category results measure whether models follow stated constraints; they are not a bias or fairness audit of the generators.

REPRODUCIBILITY STATEMENT

Prompt construction is described in Section 3 and Appendix A.1, generation settings for all eleven configurations in Appendix E.1, and the evaluator in Section 3.1, with rules and fitted thresholds in Appendices C.1 and D. The annotation protocol and agreement statistics are in Section 4 and Appendix B. Scoring is deterministic and uses no model judge. We release prompts with the submission. Manifests, generation and evaluation code, and per-statement results will be released upon acceptance.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H. S. Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *CVPR*, 2025.

Jessica Bader, Mateusz Pach, María A. Bravo, Serge Belongie, and Zeynep Akata. Stitch: Training-free position control in multimodal diffusion transformers. *arXiv preprint arXiv:2509.26644*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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 *ICCV*, 2023.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions, 2023. URL `https://cdn.openai.com/papers/dall-e-3.pdf`.

Black Forest Labs. FLUX.2 black-forest-labs/flux.2-klein-9b, 2026. URL `https://blackforestlabs.ai`.

María A. Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *CVPR*, 2023.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. 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: An open-source high-efficient image generative foundation model. In *ACM International Conference on Multimedia*, 2025b.

Yuliang Cai, Jesse Thomason, and Mohammad Rostami. TNG-CLIP: Training-time negation data generation for negation awareness of CLIP. *arXiv preprint arXiv:2505.18434*, 2025c.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 2023.

Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. BLIP3o-NEXT: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025a.

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

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Colin Conwell, Rupert Tawiah-Quashie, and Tomer Ullman. Relations, negations, and numbers: Looking for logic in generative text-to-image models. *arXiv preprint arXiv:2411.17066*, 2024.

Alakh Desai and Nuno Vasconcelos. Guiding diffusion models with adaptive negative sampling without external resources. In *ICCV*, 2025.

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

Haosheng Gan, Berk Tinaz, Mohammad Shahab Sepehri, Zalan Fabian, and Mahdi Soltanolkotabi. ConceptMix++: Leveling the playing field in text-to-image benchmarking via iterative prompt optimization. *arXiv preprint arXiv:2507.03275*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Wenqi Guo and Shan Du. VSF: Simple, efficient, and effective negative guidance in few-step image generation models by value sign flip. In *ICLR*, 2026.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *TPAMI*, 2025.

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

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. *International Conference on Intelligent User Interfaces*, 2022.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *EMNLP*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *TMLR*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

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

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022.

Mohammad Nadeem, Shahab Saquib Sohail, Erik Cambria, Björn W. Schuller, and Amir Hussain. Negation blindness in large language models: Unveiling the NO syndrome in image generation. *arXiv preprint arXiv:2409.00105*, 2024.

OpenAI. Addendum to GPT-4o system card: Native image generation. Technical report, OpenAI, 2025. URL https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know "no" better: A data-driven approach for enhancing negation awareness in CLIP. In *ICCV*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. RF-DETR: Neural architecture search for real-time detection transformers. In *ICLR*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Fawaz Sammani, Tzoulio Chamiti, Paul Gavrikov, and Nikos Deligiannis. When negation is a geometry problem in vision-language models. In *CVPR Workshop on Multimodal Algorithmic Reasoning*, 2026.

Mohammad Amin Shabani, Sepidehsadat Hosseini, and Yasutaka Furukawa. Housediffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising. *CVPR*, 2023.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learning the power of "no": Foundation models with negations. In *WACV*, 2025.

Jamie Stirling, Noura Al-Moubayed, Chris G. Willcocks, and Hubert P. H. Shum. Controllable image generation with composed parallel token prediction. In *CVPR Workshop on Low-Level Vision Frontiers*, 2026.

Tencent Hunyuan Team. Hunyuanimage 2.1: An efficient diffusion model for high-resolution (2k) text-to-image generation. https://github.com/Tencent-Hunyuan/HunyuanImage-2.1, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. UniGenBench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025b.

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. *arXiv preprint arXiv:2508.02324*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Rui Xiao, Sanghwan Kim, Yongqin Xian, Zeynep Akata, and Stephan Alaniz. FINER: MLLMs hallucinate under fine-grained negative queries. In *CVPR*, 2026.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. In *ICLR*, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3. technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *NeurIPS*, 2024.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. NegVQA: Can vision language models understand negation? In *Findings of ACL*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Table 4: **Representative prompts from the released corpus.**

Category	Setting	Prompt
Color	pos0_neg1	a photo of a non-brown orange
Material	pos2_neg2	a photo of a plastic bus, a plastic knife, a non-plastic sandwich, and a non-plastic toilet
Comparative	pos0_neg1	a photo of a skis and a skateboard, where the skis is not bigger than the skateboard
Depth	pos0_neg1	a photo of a broccoli and a bench, where the broccoli is not in front of the bench
Directional	pos2_neg2	a photo of a laptop, a zebra, a backpack, a surfboard, and a carrot, where the laptop is right of the zebra, is right of the backpack, is not below the surfboard, and is not right of the carrot
Proximity	pos0_neg1	a photo of a broccoli and a bench, where the broccoli is not far from the bench

A BENCHMARK CONSTRUCTION

A.1 PROMPT CONSTRUCTION AND REPRESENTATIVE EXAMPLES

Polarity settings and object-count controls. The eight settings are pos1_neg0, pos2_neg0, pos0_neg1, pos0_neg2, pos1_neg1, pos1_neg2, pos2_neg1, and pos2_neg2. A prompt contains $M = P + N$ statements, each assigning an attribute to an object or specifying a relation between two objects. At a fixed statement load, polarity varies while the number of statements remains constant: for example, pos2_neg0, pos1_neg1, and pos0_neg2 each contain two statements. Attribute prompts name one object per statement. Relation prompts use a star topology with one shared subject A and distinct targets $B_1, \ldots, B_M$, giving exactly $M + 1$ named objects. Thus, within each category, comparisons at equal statement load also preserve object count. Negating an attribute or relation never removes its objects: "a non-wooden bench" still requires a bench, and "a bicycle that is not left of a car" still requires both objects.

Taxonomy scope. The object vocabulary follows the detector-compatible COCO classes; prompts exclude person, leaving 79 eligible classes. Size is treated as a comparative relation because it is defined relative to another object, rather than as an intrinsic attribute.

Templates and candidate selection. Prompts are generated by deterministic templates (examples in Table 4). Attribute prompts name each object with its value: "a photo of a ⟨value⟩ ⟨object⟩" for a positive statement and "a photo of a non-⟨value⟩ ⟨object⟩" for a negated one. Relation prompts use the star topology: "a photo of a ⟨subject⟩ and a ⟨target$_1$⟩ ..., where the ⟨subject⟩ is [not] ⟨relation⟩ the ⟨target$_i$⟩", with one clause per statement.

The LLM selector, Qwen/Qwen3-30B-A3B-Instruct-2507, filters object/value/predicate combinations from fixed vocabularies to those judged plausible or unusual but depictable in their positive form. This ensures that the denied attribute or relation is a meaningful possibility for the selected objects. Prompt strings are generated deterministically after selection, so wording does not drift across polarity conditions. Negation is then written into the prompt itself, and the selector plays no role in the wording or in the evaluation.

For relation construction, the selector receives a shared subject, a candidate pool of target objects, and the allowed predicates for that category. It returns only allowed predicate-target pairs. Attribute construction analogously selects values from the fixed color/material vocabularies. Returned values are validated before entering the corpus, and selector outputs are cached and reused. Among valid combinations, balancing takes priority over selector preference. Relation subjects and targets have separate usage counters; candidate targets are drawn from the least-used portion of the vocabulary and under-used predicates are preferred. Attribute objects and values follow the analogous policy.

Category vocabularies and sources. Table 1 lists the complete vocabularies; every value or relation can be positive or negated. The color list is retained from GenEval (Ghosh et al., 2023). The four material values are selected from the broader OVAD vocabulary (Bravo et al., 2023). Structured

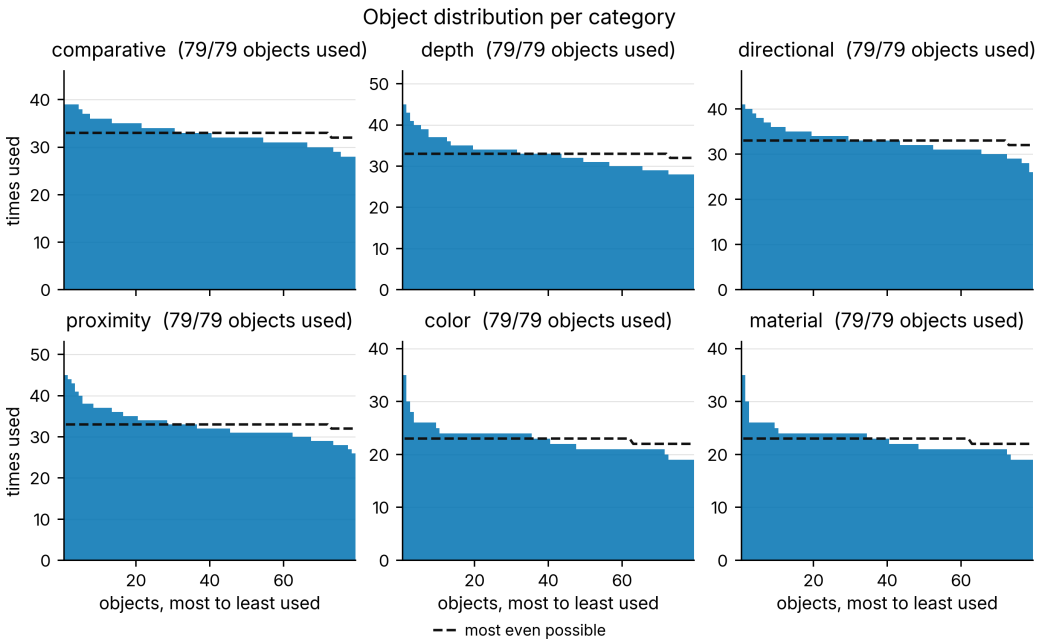


Figure 5: **Object distribution per category.** Each bar is one object's occurrence count, sorted from most to least used; the dashed line shows the most even allocation reachable for the observed number of draws. All 79 eligible COCO classes appear in every category.

prompts follow GenEval's metadata schema and PosEval's relation encoding (Bader et al., 2025). The category-specific evaluation rules are defined in Table 14 and Appendix C.1.

A.2 Dataset statistics and balancing

The final star-topology corpus contains 4,800 prompts: six categories, eight settings, and 100 prompts in every category-setting cell. Each category has 800 prompts and covers all 79 eligible COCO object classes. Averaged over the eight settings, a prompt contains 2.25 statements, half positive and half negated. Attribute prompts name one object per statement, giving 2.25 objects per prompt; relation prompts name one shared subject plus one target per statement, giving 3.25 objects per prompt.

As shown in Figures 5 and 6, all 79 eligible COCO object classes appear in every category. Pooled over the eight settings, individual object counts remain close to the most-even allocation achievable for the corresponding number of draws, and the smaller predicate/value vocabularies are distributed approximately uniformly.

B Human study and evaluator validation

B.1 Reference labels and human inter-annotator score

In this study, 36 annotators labeled 600 images, three per image, using the interface in Figure 7. The reference for an element is the majority of the three answers. The binary image label is true only when every element of the image is satisfied, and the continuous image score is the fraction of satisfied elements. We estimate the inter-annotator score (H. Inter-annotator) with a leave-one-out scheme: each annotator's labels are scored against the majority of the other two, and the score is averaged over the three held-out annotators. Table 5 reports agreement between annotators: Fleiss' κ is 72.4 over all statements, and 84.9% of statements are answered unanimously. Agreement is lower on negated statements ($\kappa = 60.2$, 70.2% unanimous) than on affirmative ones (73.0, 88.3%), so negation is harder for people as well.

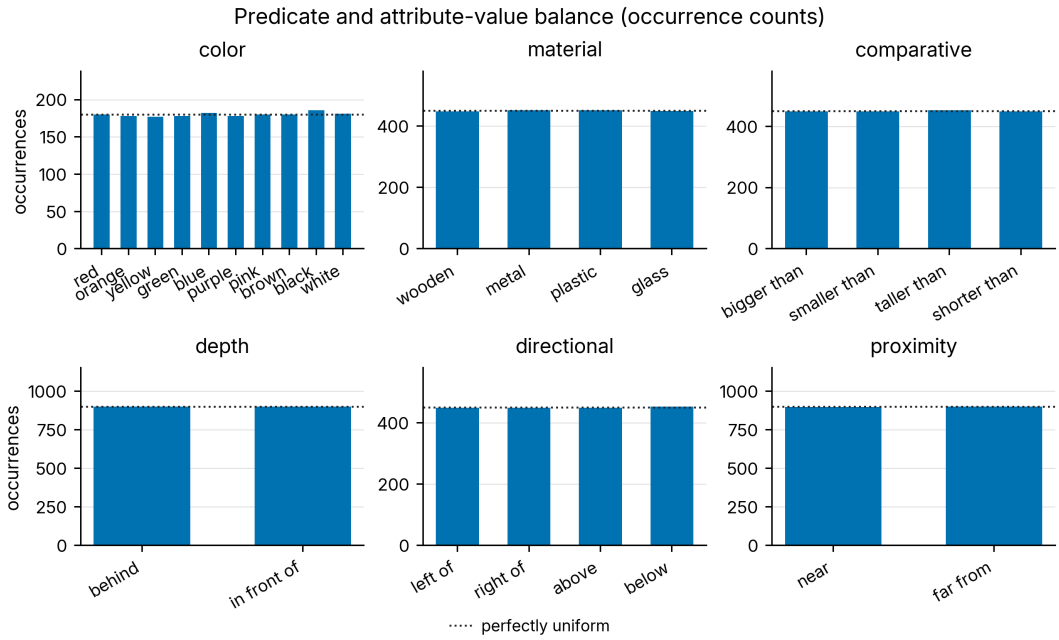


Figure 6: **Predicate and attribute-value balance.** Frequency of each relation predicate or attribute value within its category.

Table 5: **Agreement between annotators.** Fleiss' κ and the share of statements all three answered alike. Negated statements are harder for people as well, so we compare evaluators against the leave-one-out inter-annotator score rather than against perfect agreement.

Statements	n	Fleiss' κ	Unanimous	Leave-one-out κ
All statements	2772	72.4	84.9	70.9
Affirmative	2249	73.0	88.3	73.3
Negated	523	60.2	70.2	59.7

The inter-annotator score is a conservative reference, not an upper bound on achievable agreement. Each held-out annotator is compared with the majority of two annotators, a noisier target than the three-annotator majority used for every evaluator. An evaluator above $\kappa = 65.6$ therefore agrees with the three-annotator consensus more closely than a single annotator agrees with the other two; this does not make it more reliable than a human.

B.2 Evaluators' scores

- **Similarity scores** (CLIPScore (Hessel et al., 2021), BLIPScore (Li et al., 2022b), NegationCLIP). Cosine similarity between the image embedding and the embedding of the full prompt. NegationCLIP is the negation-aware CLIP of Park et al. (2025) (ViT-B/32).
- **Reward models** (ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023)). The scalar the model assigns to the image and prompt. These models predict which image a person prefers, not whether the image satisfies the prompt.
- **VQA judge, whole prompt.** The probability of "yes" to one question asking whether the image matches the whole description.
- **VQA judge, per element (pos.).** The judge answers one yes/no question per required object and one per statement, all asked in positive form, as in the annotation interface (Figure 7). The expected answer is "yes" for positive statements and "no" for negated ones. The score is the mean probability of the expected answer over the elements; elements of an absent object count as unsatisfied.

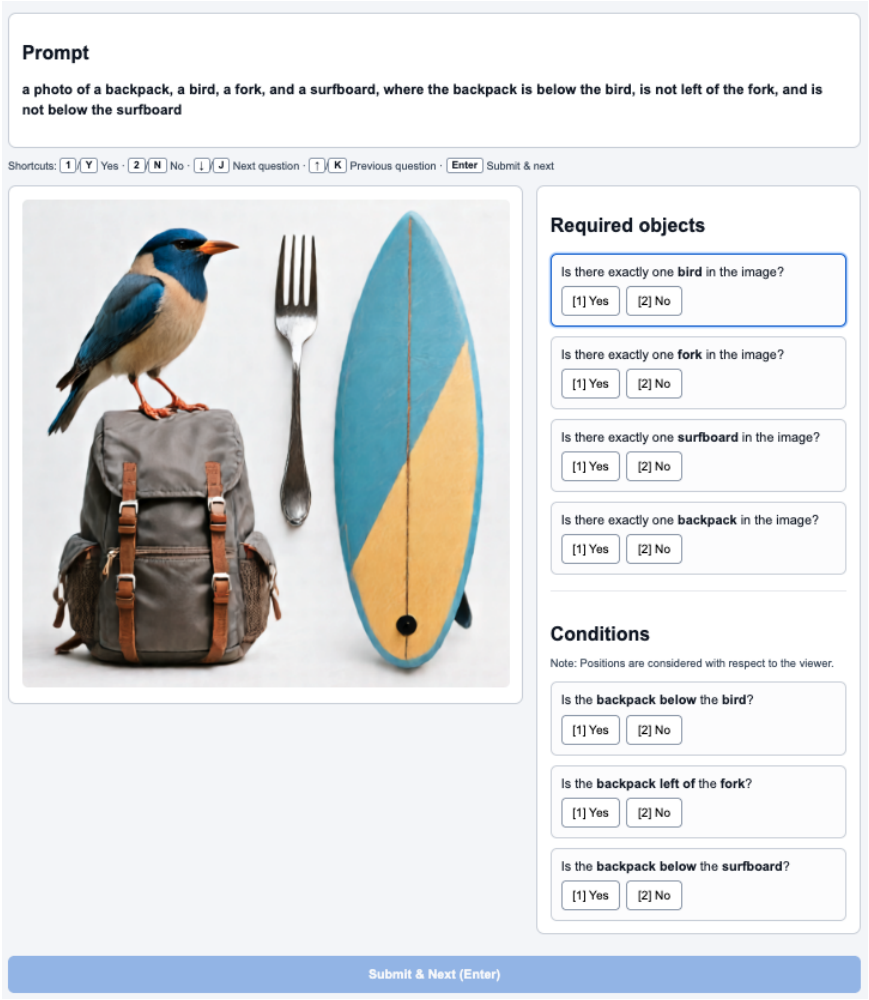


Figure 7: **The annotation interface.** The prompt is shown in full, then one question per required object and one per statement. The prompt here negates two relations, but both are asked affirmatively: the annotator answers whether the backpack is left of the fork, and a correct image gives "no". Annotators see neither the expected answers nor the T2I generator.

- **VQA judge, negated per element (neg.).** The same elements as in pos., with each negated statement asked in negated form and the expected answer flipped. Object questions and positive statements are reused from the pos. mode, so the two modes differ only in the wording of the negated statements.
- **NEGT2IBENCH.** The fraction of elements its detectors mark satisfied, on the same scale as the per-element judges and the human score.

The per-element modes give the judges the decomposition that NEGT2IBENCH produces. Without it, a judge must either answer about the whole prompt at once, which generally gives lower binary agreement (Table 6), or invent its own decomposition. NEGT2IBENCH also reports which requirement failed and why, which enables the category and polarity analyses of Section 5.

B.3 MEASURES

Binary. Accuracy and Cohen's κ against the binary human label. We exclude the similarity and reward scores, because no single threshold applies equally to them.

Table 6: **Binary agreement, all scoring modes.** Whole prompt, the benchmark's per-element decomposition, and that decomposition with the statements the prompt negates asked in negated form; all other questions are unchanged.

Evaluator	Whole prompt	Per element	Negated per element
Detector-based			
NEGT2IBENCH (ours)	–	**72.0**	**72.0**
VQA judges			
VQAScore	33.6	49.9	48.6
mPLUG	7.9	16.0	10.1
Qwen2.5-VL-7B	35.2	52.8	44.3
InternVL3-8B	25.2	59.7	41.8
Qwen3-VL-8B	48.2	69.7	54.1
Qwen3-VL-32B	48.2	74.3	66.5
InternVL3.5-8B	35.3	69.0	50.2
InternVL3.5-38B	48.4	69.5	52.6
H. Inter-annotator	–	65.6	–

Continuous. Spearman ρ, Kendall τ and pairwise ranking accuracy against the continuous human score. We report all three because they order the evaluators differently: NEGT2IBENCH leads on τ (65.1 vs. 63.3 for Qwen3-VL-32B) but trails on ρ (72.8 vs. 76.5) and pairwise accuracy (91.7 vs. 94.7). Kendall counts every discordant pair equally, while Spearman penalizes large rank shifts more, so NEGT2IBENCH makes more small ranking errors and the judges make fewer but larger ones. The three metrics treat ties and rank displacements differently, so they order the evaluators differently; we report all three.

Uncertainty. All intervals are 95% percentile intervals from a bootstrap that resamples whole images, so correlated decisions within an image stay together. We compute differences between two evaluators on the same resample, a stronger test than comparing two separate intervals.

Similarity and reward scores. Table 7 includes the six similarity and reward scores. All six rank images far below the VQA judges and NEGT2IBENCH: the best, PickScore, reaches $\tau = 27.7$, against 38.8 for the weakest VQA judge other than mPLUG. NegationCLIP, trained on negated captions, improves on CLIPScore only from 13.1 to 16.6. Neither similarity nor preference tells whether a negated requirement holds.

B.4 COST PROTOCOL

We measure every method on the same 600 images on one H200 NVL GPU, with one process, no batching, and eight CPU threads. Pinning the threads changes wall-clock time by a factor of 1.2 and CPU time by a factor of 8, but changes no verdict. Reported time excludes model loading, which Table 8 lists separately. GPU memory is `torch.cuda.max_memory_allocated` for every method, so all values share one basis; the allocator's reserved memory is larger and not comparable across frameworks. CPU memory is the peak resident set size. For NEGT2IBENCH we time the full evaluation of each image. For the judges we time each question and report the time per image as the time per question multiplied by the mean number of element questions per image, 4.66; Table 8 gives both values.

NEGT2IBENCH has four components: SigLIP 2 SO400M/14 at 378 px for color and material (1,136.0M parameters), SAM 2.1 hiera-small for mask refinement (38.5M on the image path, of 46.1M loaded; the memory-attention modules serve video tracking and are unused), RF-DETR-Seg-XL for detection, counting and masks (38.1M), and Depth-Anything-V2-Small for depth relations (24.8M). Together they total 1.24B parameters and 2.0GiB on disk.

VQAScore and Qwen2.5-VL-7B run through t2v-metrics, which computes a $K \times K$ image-text matrix and keeps only its diagonal. Their time reflects that wrapper as much as the model, so we report it but exclude both from every speed comparison in the paper.

Table 7: **Rank agreement with the human score.** Spearman ρ, Kendall τ and pairwise ranking accuracy, with 95% bootstrap intervals over images. Judges are ordered by their best ρ. Rank agreement needs no threshold, so it is the only comparison that admits the similarity and reward scores.

Evaluator	ρ	τ	Pair acc.
Qwen3-VL-32B (Bai et al., 2025a)			
whole prompt	63.2 [57.6, 68.2]	48.6 [43.9, 52.9]	87.5 [81.7, 92.4]
pos. per element	**76.5** [71.5, 81.1]	63.3 [58.7, 67.7]	**94.7** [91.6, 97.4]
neg. per element	72.4 [67.0, 77.3]	58.8 [54.2, 63.3]	94.4 [91.3, 97.1]
Qwen3-VL-8B (Bai et al., 2025a)			
whole prompt	60.4 [54.2, 66.0]	46.4 [41.2, 51.1]	86.5 [81.4, 91.2]
pos. per element	74.0 [68.5, 78.8]	61.0 [56.3, 65.6]	94.2 [91.2, 96.7]
neg. per element	68.3 [62.6, 73.4]	54.8 [50.0, 59.5]	94.3 [91.2, 96.8]
InternVL3.5-38B (Wang et al., 2025a)			
whole prompt	63.6 [57.9, 68.8]	49.2 [44.4, 53.7]	87.4 [81.7, 92.2]
pos. per element	73.9 [68.7, 78.4]	59.7 [55.0, 64.0]	94.3 [91.4, 96.7]
neg. per element	67.2 [61.5, 72.4]	52.8 [47.8, 57.5]	93.9 [91.1, 96.5]
InternVL3-8B (Zhu et al., 2025)			
whole prompt	60.1 [54.1, 65.6]	46.2 [41.2, 51.1]	86.1 [80.6, 91.0]
pos. per element	69.5 [63.5, 74.6]	56.3 [51.2, 61.0]	92.3 [88.7, 95.4]
neg. per element	64.9 [58.8, 70.4]	51.1 [45.8, 56.1]	92.7 [89.3, 95.6]
InternVL3.5-8B (Wang et al., 2025a)			
whole prompt	57.3 [51.1, 62.5]	43.2 [38.3, 47.6]	86.2 [81.5, 90.4]
pos. per element	69.2 [63.2, 74.5]	55.3 [50.1, 60.1]	90.7 [86.4, 94.2]
neg. per element	64.0 [58.0, 69.4]	49.8 [44.8, 54.6]	91.0 [87.1, 94.4]
Qwen2.5-VL-7B (Bai et al., 2025b)			
whole prompt	58.4 [52.1, 64.0]	44.6 [39.3, 49.4]	87.7 [82.5, 92.4]
pos. per element	56.3 [49.8, 62.5]	43.7 [38.3, 48.9]	90.8 [87.6, 93.6]
neg. per element	57.1 [50.6, 62.7]	44.2 [38.8, 49.0]	91.3 [88.2, 93.9]
VQAScore (Lin et al., 2024)			
whole prompt	52.3 [45.6, 58.2]	38.8 [33.6, 43.7]	83.2 [78.0, 88.1]
pos. per element	55.8 [49.0, 61.6]	42.3 [36.9, 47.1]	85.8 [80.6, 90.3]
neg. per element	56.1 [49.5, 62.1]	42.8 [37.5, 47.8]	86.8 [81.6, 91.4]
mPLUG (Li et al., 2022a)			
whole prompt	28.9 [21.0, 36.7]	21.1 [15.0, 27.1]	69.1 [62.8, 74.9]
pos. per element	-5.2 [-13.4, 3.1]	-4.3 [-10.2, 1.9]	39.1 [33.2, 45.3]
neg. per element	-6.9 [-15.2, 1.5]	-5.5 [-11.5, 0.5]	38.4 [32.7, 44.3]
Similarity and reward			
CLIPScore	18.0 [9.8, 26.2]	13.1 [7.2, 19.1]	73.5 [67.6, 78.7]
BLIPScore	17.0 [8.4, 25.2]	12.4 [6.2, 18.4]	69.8 [63.5, 75.6]
NegationCLIP	22.9 [14.6, 30.4]	16.6 [10.6, 22.1]	74.6 [69.1, 79.8]
ImageReward	25.7 [17.3, 33.7]	18.6 [12.3, 24.8]	80.2 [74.5, 85.2]
PickScore	37.5 [30.0, 44.5]	27.7 [22.0, 33.2]	82.9 [78.3, 87.0]
HPSv2	30.7 [22.7, 37.9]	22.3 [16.6, 27.7]	80.8 [75.9, 85.2]
NEGT2IBENCH (ours)	72.8 [67.9, 77.9]	**65.1** [60.4, 70.0]	91.7 [88.5, 95.5]
H. Inter-annotator	77.9	71.1	95.0

A judge answers one question per element, so its time per image grows linearly with the number of elements. NEGT2IBENCH runs its detector once per image. At the benchmark's mean of 4.66 elements per image, NEGT2IBENCH (0.78 s) is faster than the two largest judges, Qwen3-VL-32B (0.99 s) and InternVL3.5-38B (3.29 s), but slower than the two 8B judges with comparable agreement, Qwen3-VL-8B (0.49 s) and InternVL3.5-8B (0.76 s), and uses 2.7 to 11.6 times less GPU memory than all of them.

Table 8: **Cost of one evaluation.** The same 600 images on one H200 NVL, one process, no batching, eight CPU threads. Model load is reported separately and excluded from s/img. GPU memory is `torch.cuda.max_memory_allocated`. For a judge, s/img is s/Q multiplied by the 4.66 element questions an image carries on average: every judge is given the benchmark's own decomposition, so they answer identical questions and differ only in s/Q. NEGT2IBENCH asks no questions and is timed per image. †Scored through t2v-metrics, which computes a $K \times K$ image–text matrix and keeps only its diagonal, so the wall-clock reflects that wrapper rather than the model.

Evaluator	Params	s/Q	s/img	Load (s)	GPU GiB
InternVL3.5-38B	38B	0.706	3.29	28.6	73.0
Qwen3-VL-32B	32B	0.212	0.99	21.8	62.8
VQAScore†	11B	0.729	3.40	41.6	24.2
Qwen3-VL-8B	8B	0.105	0.49	12.8	16.8
InternVL3-8B	8B	0.152	0.71	26.3	15.8
InternVL3.5-8B	8B	0.164	0.76	13.8	17.1
Qwen2.5-VL-7B†	7B	1.803	8.39	30.3	15.9
mPLUG	0.6B	0.068	0.32	22.1	6.5
NEGT2IBENCH (ours)	1.24B	–	0.78	24.5	6.3

Table 9: **Error structure.** Counts against the majority label of three annotators on 600 images, for the per-element scoring mode. FP is an image the annotators reject but the evaluator passes; FN is an image the annotators accept but the evaluator rejects. Judges are ordered by Cohen's κ; NEGT2IBENCH is listed separately.

Evaluator	TP	TN	FP	FN	Prec.	Rec.	Called correct (%)
Qwen3-VL-32B, pos.	169	362	32	37	**84.1**	82.0	33.5
Qwen3-VL-8B, pos.	161	358	36	45	81.7	78.2	32.8
InternVL3.5-38B, pos.	162	356	38	44	81.0	78.6	33.3
InternVL3.5-8B, pos.	169	346	48	37	77.9	82.0	36.2
InternVL3-8B, pos.	148	344	50	58	74.7	71.8	33.0
Qwen2.5-VL-7B, pos.	117	364	30	89	79.6	56.8	24.5
VQAScore, pos.	134	332	62	72	68.4	65.0	32.7
mPLUG, pos.	36	377	17	170	67.9	17.5	8.8
NEGT2IBENCH (ours)	175	348	46	31	79.2	**85.0**	36.8
Human reference	–	–	–	–	–	–	34.3

B.5 ERROR STRUCTURE

Cohen's κ merges two kinds of error into one number. Table 9 separates them for the per-element mode, the mode behind the precision and recall of Table 2. TP: the annotators and the evaluator both accept the image; TN: both reject it; FP: the annotators reject it but the evaluator passes it; FN: the annotators accept it but the evaluator rejects it. Prec. $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and Rec. $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. Called correct (%) corresponds to the share of the 600 images the evaluator passes, $(\mathrm{TP}+\mathrm{FP})/600$. Humans accept 34.3% of the images.

NEGT2IBENCH accepts 36.8%, making 46 false passes and 31 false failures, while the strongest judge, Qwen3-VL-32B, accepts 33.5% and makes 32 and 37. The two evaluators reject about as many images that humans accept, and NEGT2IBENCH passes more images that humans reject. NEGT2IBENCH therefore rarely fails an image that humans consider correct, so the generator failures we report are not caused by a strict evaluator. The same table exposes a degenerate baseline that its aggregate score hides: mPLUG, scored per element, accepts only 8.8% of images and makes 170 false failures, close to a constant "no".

B.6 SENSITIVITY TO HOW A REQUIREMENT IS PHRASED

The pos. and neg. columns of Table 2 report κ on the benchmark's own elements and on the same elements with the negated statements asked in negated form. Every judge loses agreement when

Table 10: **Negated-question probe.** Acc. whole, pos., neg.: image accuracy against the binary human label when the judge is asked the whole prompt at once, the per-element questions in positive form, or the same questions with the negated statements asked in negated form. Flip: share of negated statements whose yes/no answer changes between the positive and negated wording. Compl. err.: mean $|p_{\text{pos}} + p_{\text{neg}} - 1|$ over those statements, where p is the probability of "yes" under each wording; it is zero when the two answers are complementary. Called corr. pos., neg.: share of the 600 images the evaluator passes in each mode; the annotators pass 34.3%. NEGT2IBENCH has one scoring mode and no question to rephrase.

Evaluator	Acc. whole	Acc. pos.	Acc. neg.	Flip (%)	Compl. err.	Called corr. pos. (%)	Called corr. neg. (%)
VQAScore	64.8	77.7	76.7	**80.0**	0.160	32.7	35.3
Qwen3-VL-32B	73.5	**88.5**	**84.7**	77.5	0.225	33.5	36.3
Qwen3-VL-8B	74.2	86.5	77.8	62.1	0.364	32.8	44.5
InternVL3.5-8B	64.7	85.8	76.5	58.9	0.361	36.2	41.2
InternVL3.5-38B	73.2	86.3	77.7	51.1	0.381	33.3	40.7
InternVL3-8B	56.8	82.0	71.5	46.7	0.431	33.0	46.8
mPLUG	46.5	68.8	67.3	46.3	0.330	8.8	6.3
Qwen2.5-VL-7B	66.2	80.2	78.3	40.7	0.589	24.5	14.7
NEGT2IBENCH (ours)	**87.2**	–	–	–	–	36.8	–
H. Inter-annotator	85.8	–	–	–	–	34.3	–

using negated prompts. The six instruction-tuned judges, from two model families, lose 7.8 to 18.8 points, and scale does not protect them: the two largest judges lose 7.8 and 16.9. NEGT2IBENCH checks a typed requirement rather than a sentence, so it is unaffected, and under negated phrasing it ranks first.

Table 10 shows the mechanism. For each element we ask both the affirmative and the negated form and measure how often the answer flips. A judge that reads the negation should flip every time. Qwen3-VL-32B flips on 77.5% of elements. Qwen2.5-VL-7B flips on only 40.7%, and its complementarity error of 0.589 shows that it mostly repeats its affirmative answer.

B.7 AGREEMENT BY POLARITY SETTING

Table 11 breaks down Kendall τ by polarity setting. Each setting holds 70 to 90 images, so a per-cell 95% interval spans about 32 points and few per-cell differences are resolvable; we read the table only for patterns shared by many evaluators. No strong evaluator collapses in any setting, and NEGT2IBENCH stays between 48.1 and 88.6. The similarity and reward scores do collapse once the prompt contains a single negation: all six fall from τ between 20.5 and 49.1 at `1:0` to between -22.2 and 20.7 at `0:1`.

B.8 AGREEMENT BY CATEGORY

Table 12 breaks down Kendall τ by category (100 images each; per-cell intervals are about 26 points wide). Judges asked about the whole prompt agree well on color (65.4 to 78.4) but fall on the four relations (25.3 to 52.8); per-element questions recover part of this gap. NEGT2IBENCH is more even across categories (54.4 to 78.5), and trails the strongest judges mainly on color and depth, the category most affected by dead zones (Section 5). Similarity and reward scores reach useful agreement only on color.

B.9 PER-ELEMENT AGREEMENT

Table 13 scores each object and statement on its own instead of one verdict per image, so a negated statement is judged even when its image fails on another element. Annotators also agree less on negated statements, so each block has its own inter-annotator score (73.3 positive, 59.7 negated). On positive statements the strongest judges agree more closely than NEGT2IBENCH (Qwen3-VL-32B 76.3 vs. 64.3). On negated statements the gap narrows: NEGT2IBENCH reaches $\kappa = 56.4$, within the interval of the best judge (60.7) and close to the inter-annotator score (59.7), and it loses less agreement from positive to negated statements than the three strongest judges (7.9 points, against 14.3 to 18.5). Asked in negated form, every judge falls further, to between 2.3 and 43.5.

Table 11: Kendall τ against the continuous human score, by polarity setting (positive:negated statements). Judges are shown in their three scoring modes. The inter-annotator score (H. inter-annotator) is recomputed within each setting. Settings with no negated statement (1:0, 2:0) have nothing to rephrase, so their negated row is omitted. Each cell holds 70 to 90 images; per-cell 95% bootstrap intervals have a median width of 32 points. *All* is computed on all 600 images.

	Load 1		Load 2			Load 3		Load 4	
Evaluator	1:0	0:1	2:0	1:1	0:2	2:1	1:2	2:2	All
Qwen3-VL-32B (Bai et al., 2025a)									
whole prompt	46.4	57.0	43.8	50.3	38.8	42.9	49.0	30.3	48.6
pos. per element	**66.8**	66.8	62.3	65.4	**67.4**	60.2	51.7	57.7	63.3
neg. per element	–	68.5	–	55.5	53.8	57.9	51.5	49.6	58.8
Qwen3-VL-8B (Bai et al., 2025a)									
whole prompt	54.3	55.2	39.6	32.5	40.1	38.1	47.3	39.2	46.4
pos. per element	61.7	72.8	53.3	63.7	58.0	**64.6**	**58.8**	53.9	61.0
neg. per element	–	57.2	–	47.6	49.3	62.0	55.8	48.3	54.8
InternVL3.5-38B (Wang et al., 2025a)									
whole prompt	52.0	57.7	39.6	38.5	44.0	49.8	43.0	32.6	49.2
pos. per element	62.7	69.7	55.6	63.0	63.4	50.4	44.1	59.3	59.7
neg. per element	–	62.6	–	49.4	20.2	46.6	42.0	49.8	52.8
InternVL3-8B (Zhu et al., 2025)									
whole prompt	47.5	55.4	42.4	43.4	22.9	47.6	36.7	38.5	46.2
pos. per element	65.1	74.1	40.0	57.6	54.7	51.5	50.6	52.6	56.3
neg. per element	–	74.4	–	44.4	31.5	39.9	42.1	36.6	51.1
InternVL3.5-8B (Wang et al., 2025a)									
whole prompt	58.3	51.8	49.0	30.9	37.1	35.1	40.7	36.1	43.2
pos. per element	57.9	63.0	63.7	55.2	57.2	45.8	36.4	45.3	55.3
neg. per element	–	63.4	–	36.7	34.7	41.3	27.0	33.3	49.8
Qwen2.5-VL-7B (Bai et al., 2025b)									
whole prompt	46.9	58.2	42.4	49.0	19.8	46.3	45.9	35.5	44.6
pos. per element	51.7	58.9	45.2	55.6	40.2	34.1	38.9	43.4	43.7
neg. per element	–	66.5	–	55.4	23.5	33.7	39.5	38.6	44.2
VQAScore (Lin et al., 2024)									
whole prompt	36.5	50.2	26.8	10.3	23.2	48.2	48.6	24.8	38.8
pos. per element	54.3	57.2	42.5	34.2	10.9	34.5	19.7	41.9	42.3
neg. per element	–	50.0	–	33.8	7.8	36.5	33.3	42.6	42.8
mPLUG (Li et al., 2022a)									
whole prompt	37.1	18.7	22.7	-0.5	1.4	11.8	15.1	4.6	21.1
pos. per element	-4.8	12.3	-7.3	-29.0	-35.8	0.7	-15.3	-9.8	-4.3
neg. per element	–	7.3	–	-28.1	-36.0	-0.2	-8.3	-9.6	-5.5
Similarity and reward									
CLIPScore	38.5	-1.3	22.7	10.2	9.1	17.1	29.8	29.7	13.1
BLIPScore	23.0	-22.2	12.2	23.5	2.7	17.3	23.4	42.8	12.4
NegationCLIP	29.4	2.9	30.7	22.5	0.7	14.7	17.6	36.3	16.6
ImageReward	20.5	-11.1	33.3	29.9	17.1	30.4	30.4	31.8	18.6
PickScore	49.1	-1.9	32.3	41.7	6.6	23.9	23.1	35.7	27.7
HPSv2	38.3	20.7	27.1	33.5	-0.6	23.1	34.6	22.2	22.3
NEGT2IBENCH (ours)	62.3	**88.6**	**76.3**	**77.4**	66.8	48.1	50.1	**63.2**	**65.1**
H. Inter-annotator	79.1	88.1	73.6	66.7	76.4	65.2	69.2	73.9	71.1
n	90	90	70	70	70	70	70	70	600

Table 12: Kendall τ against the continuous human score, by semantic category. Judges are shown in their three scoring modes. The inter-annotator score (H. inter-annotator) is recomputed within each category. Each cell holds 100 images; per-cell 95% bootstrap intervals have a median width of 26 points. *All* is computed on all 600 images.

	Attributes		Relations				
Evaluator	Color	Material	Comparative	Depth	Directional	Proximity	All
Qwen3-VL-32B (Bai et al., 2025a)							
whole prompt	69.7	58.9	47.2	39.9	47.9	38.2	48.6
pos. per element	83.2	69.7	57.3	**63.7**	61.7	53.2	63.3
neg. per element	83.5	67.8	41.9	59.7	53.2	50.1	58.8
Qwen3-VL-8B (Bai et al., 2025a)							
whole prompt	69.8	56.6	52.8	35.3	42.1	36.6	46.4
pos. per element	**83.8**	62.7	55.1	60.4	56.4	**61.2**	61.0
neg. per element	75.5	60.0	42.7	50.7	47.6	58.1	54.8
InternVL3.5-38B (Wang et al., 2025a)							
whole prompt	74.6	53.5	47.2	37.3	48.5	39.0	49.2
pos. per element	82.0	57.3	61.7	60.5	59.0	45.1	59.7
neg. per element	83.4	59.5	40.0	44.4	49.0	50.1	52.8
InternVL3-8B (Zhu et al., 2025)							
whole prompt	75.2	48.2	43.0	36.2	47.2	36.1	46.2
pos. per element	82.3	50.6	48.3	59.3	59.7	47.0	56.3
neg. per element	82.0	52.4	47.1	42.2	36.0	51.1	51.1
InternVL3.5-8B (Wang et al., 2025a)							
whole prompt	71.3	45.1	46.0	35.3	39.1	36.8	43.2
pos. per element	78.8	58.7	48.7	44.9	58.4	40.9	55.3
neg. per element	70.6	64.3	27.8	36.3	42.3	49.5	49.8
Qwen2.5-VL-7B (Bai et al., 2025b)							
whole prompt	78.4	60.3	37.7	33.6	42.2	39.3	44.6
pos. per element	66.3	51.0	31.7	36.8	41.3	34.5	43.7
neg. per element	74.1	54.5	36.3	32.2	39.7	36.3	44.2
VQAScore (Lin et al., 2024)							
whole prompt	65.4	46.5	25.3	44.6	29.1	37.3	38.8
pos. per element	82.4	48.8	30.0	28.3	31.9	36.0	42.3
neg. per element	82.3	44.3	30.3	36.1	32.2	42.3	42.8
mPLUG (Li et al., 2022a)							
whole prompt	64.4	-12.4	11.3	13.2	5.3	26.9	21.1
pos. per element	15.9	9.2	-20.3	-5.4	-23.1	-3.4	-4.3
neg. per element	15.5	3.6	-20.5	-4.4	-20.1	-6.7	-5.5
Similarity and reward							
CLIPScore	45.9	-4.0	9.0	17.2	21.1	18.8	13.1
BLIPScore	20.0	-2.4	15.6	13.9	20.2	11.7	12.4
NegationCLIP	46.1	-1.8	15.3	15.1	17.3	14.2	16.6
ImageReward	38.1	-8.6	23.2	28.8	23.0	15.0	18.6
PickScore	56.6	11.2	34.8	23.3	25.5	10.6	27.7
HPSv2	44.2	25.3	27.7	13.1	9.1	18.3	22.3
NEGT2IBENCH (ours)	75.2	**78.5**	**72.3**	54.4	**65.3**	60.0	**65.1**
H. Inter-annotator	84.2	81.4	78.8	66.5	65.8	64.3	71.1

C EVALUATION RULES AND METRICS

C.1 EVALUATION RULES

All reported benchmark results use RF-DETR-XL with the current evaluator configuration. The formal rules below are deterministic. Unless otherwise noted, only relation measurement operates on a SAM-2-refined instance mask.

Table 13: **Element-level agreement, by polarity.** One decision per object and per statement instead of one verdict per image. Negated statements are harder for the annotators too, so each block has its own inter-annotator score. The last three columns give the share of satisfied and of unsatisfied statements each method gets right, and their mean. Intervals on κ are 95% bootstrap intervals resampling images.

Evaluator	n	Acc.	κ	Satisfied	Unsatisfied	Balanced
Affirmative statements						
NEGT2IBENCH (ours)	2249	90.5	64.3 [59.6, 68.9]	95.6	65.5	80.5
VQAScore, pos.	2242	88.8	53.4 [48.1, 58.4]	96.9	48.9	72.9
mPLUG, pos.	2242	42.5	-1.0 [-3.9, 2.0]	39.2	58.8	49.0
Qwen2.5-VL-7B, pos.	2242	79.0	48.3 [44.4, 52.1]	75.9	94.4	85.1
InternVL3-8B, pos.	2242	90.9	63.8 [59.2, 68.6]	97.2	60.1	78.6
Qwen3-VL-8B, pos.	2242	92.6	72.9 [69.0, 76.9]	95.9	75.9	85.9
Qwen3-VL-32B, pos.	2242	93.3	**76.3** [72.8, 79.8]	95.7	81.4	88.6
InternVL3.5-8B, pos.	2242	91.2	64.9 [60.1, 69.5]	97.4	60.7	79.1
InternVL3.5-38B, pos.	2242	91.7	68.1 [63.8, 72.3]	96.6	67.1	81.9
H. inter-annotator	2249	91.6	73.3	–	–	–
Negated statements						
NEGT2IBENCH (ours)	523	78.2	56.4 [48.8, 63.5]	76.4	80.2	78.3
VQAScore, pos.	519	61.3	23.7 [15.6, 31.1]	41.3	82.8	62.0
VQAScore, neg.	519	58.8	18.3 [9.9, 26.6]	46.8	71.6	59.2
mPLUG, pos.	519	59.3	20.0 [11.9, 27.4]	37.5	82.8	60.2
mPLUG, neg.	519	50.7	2.3 [-6.2, 10.5]	37.2	65.2	51.2
Qwen2.5-VL-7B, pos.	519	72.4	44.0 [37.2, 51.0]	92.2	51.2	71.7
Qwen2.5-VL-7B, neg.	519	59.7	21.4 [15.4, 27.9]	26.4	95.6	61.0
InternVL3-8B, pos.	519	73.0	46.6 [39.1, 54.0]	57.6	89.6	73.6
InternVL3-8B, neg.	519	54.7	8.5 [0.7, 17.2]	67.7	40.8	54.2
Qwen3-VL-8B, pos.	520	77.1	54.4 [46.8, 62.4]	70.7	84.0	77.4
Qwen3-VL-8B, neg.	519	64.9	28.4 [21.3, 36.9]	91.4	36.4	63.9
Qwen3-VL-32B, pos.	520	80.4	**60.7** [53.8, 67.4]	80.7	80.0	80.4
Qwen3-VL-32B, neg.	519	72.1	43.5 [36.3, 51.1]	85.5	57.6	71.6
InternVL3.5-8B, pos.	520	75.0	50.4 [43.1, 57.9]	64.1	86.8	75.4
InternVL3.5-8B, neg.	519	58.4	16.7 [7.9, 26.1]	59.1	57.6	58.4
InternVL3.5-38B, pos.	520	76.7	53.8 [45.9, 61.7]	66.3	88.0	77.1
InternVL3.5-38B, neg.	519	59.9	18.9 [10.1, 28.2]	75.5	43.2	59.3
H. inter-annotator	523	80.0	59.7	–	–	–

Detection and instance counting. RF-DETR-XL (`RFDETRSegXLarge`) is called with an internal prediction floor of 0.005. A requested class is considered present when at least one detection reaches

$$\tau_{\text{presence}} = 0.03.$$

Instance counting uses the stricter threshold.

$$\tau_{\text{count}} = 0.5.$$

If no detection reaches τ_{count}, the most confident detection for that requested class is retained so that a faint but detected object can still be judged. If more counted instances of a requested class are present than requested, the prompt is marked as an over-count. `skis` is exempt from this rule because one semantic referent is often returned as multiple detections. At most 16 detections per class are retained, and overlap suppression is disabled (IoU $= 1.0$).

SAM 2 mask refinement. We refine detections with `facebook/sam2.1-hiera-small`. Detector boxes are passed as SAM box prompts; the candidate mask is selected by overlap with the detector's own mask rather than by SAM confidence. The refined mask is post-processed with morphological closing of radius 5 pixels. Connected components smaller than 5% of the largest component are removed, while larger disconnected pieces are retained to preserve partially occluded objects. SAM-refined masks are used for only the four relation categories. SAM refinement is restricted to detections corresponding to objects named in the prompt. Each relation prompt names at most five objects.

Table 14: **NEGT2IBENCH evaluation rules.** Boolean tests on masks by category: positive statements pass when condition holds, negated ones when it does not. Categories span diverse negation forms.

Category (#values)	What we check	Condition for the attribute or relation to hold
Color (10) / Material (4)	The object carries the requested color / material	$s_v \geq s_{\max} - \beta$, ($\beta = 0.5$ for color, 0 for material) s_v is SigLIP 2 score for color / material v on object crop and $s_{\max}$ the highest; for color, objects count as having every value within 0.5 of $s_{\max}$, so when two values score closely, a statement denying either fails; material uses the top value only
Directional (4)	The subject lies on the requested side of the target	left / right: $d_x < -0.45$, $d_x > 0.45$ above / below: $d_y < -0.45$, $d_y > 0.45$ (d_x, d_y) is offset between mask centers, shrunk by $0.075\times$ summed mask extents and normalized by original center distance
Proximity (2)	Two objects are near each other or far apart	near: $g < 0.75R$; far from: $g \geq 0.75R$ g is minimum mask-to-mask gap and R is mean equivalent radius. Every measurable pair is either near or far, so proximity has no designed ambiguous middle
Comparative (4)	One object is bigger or taller than another	bigger / smaller: $a_A > 1.05a_B$; $a_B > 1.05a_A$ taller / shorter: $h_A > 1.05h_B$; $h_B > 1.05h_A$ a is mask area and h mask height; pairs within the factor are called neither way
Depth (2)	The subject is in front of the target, or behind it	Let $\bar{\delta}_A$ and $\bar{\delta}_B$ be the mean disparities over the two objects' facing surfaces, defined within the mask-to-mask gap plus 16 pixels. A depth relation is emitted only if $\frac{\lvert\bar{\delta}_A - \bar{\delta}_B\rvert}{\delta_{\max} - \delta_{\min}} \geq 0.05 + 0.10\frac{g}{R}.$ If the condition holds, the object with larger disparity is in front; otherwise no depth predicate is emitted

Attribute decision rule. Color and material are evaluated with `ViT-SO400M-14-SigLIP2-378` SigLIP 2 features (`pretrained=webli`). Pixels outside the refined instance mask are filled with gray (`#999`) before the instance crop is classified.

Let V be the attribute vocabulary and s_v the SigLIP 2 logit for value $v \in V$. The evaluator forms the multi-label confusion set.

$$C(I, o) = \left\{ v \in V : s_v \geq \max_{u \in V} s_u - \beta \right\}, \qquad \beta = 0.5 \text{ for color and } \beta = 0 \text{ for material (plain top-1).}$$

For a requested value $v^\star$, an asserted attribute passes if and only if $v^\star \in C(I, o)$, while a negated attribute passes if and only if $v^\star \notin C(I, o)$. Thus, a near-tied denied value is not treated as confidently absent.

The color vocabulary is

$$\{\text{red, orange, yellow, green, blue, purple, pink, brown, black, white}\},$$

and the material vocabulary is

$$\{\text{wooden, metal, plastic, glass}\}.$$

Directional relation rule. We adapt GenEval's position rule (Ghosh et al., 2023), using refined-mask centroids and extents in place of bounding-box centers and sizes, with the thresholds reported below.

Let M_A, M_B be the refined masks, c_A, c_B their centroids, and $(w_A, h_A), (w_B, h_B)$ their mask extents. With centroid offset $o = c_A - c_B$, define the extent-discounted offset

$$o_e = \max(|o| - \tau_{\text{dir}}(w_A + w_B,\ h_A + h_B), 0) \odot \text{sign}(o),$$

with $\tau_{\text{dir}} = 0.075$. The evaluator then normalizes the discounted offset using the original centroid-displacement magnitude, $(d_x, d_y) = \frac{o_e}{\|o\|_2}$. A corresponding signed directional predicate fires when the magnitude of its normalized component exceeds $\alpha_{\text{dir}} = 0.45$. If neither axis clears the threshold, no directional predicate is emitted for the pair.

Proximity relation rule.

The minimum edge-to-edge distance between the two refined masks is

$$g = \min_{p \in M_A,\, q \in M_B} \|p - q\|_2.$$

Define the mean equivalent-circle radius

$$R = \frac{1}{2}\left(\sqrt{\frac{|M_A|}{\pi}} + \sqrt{\frac{|M_B|}{\pi}}\right).$$

The decision boundary is

$$\text{near} \iff g < \gamma_{\text{prox}} R, \qquad \text{far from} \iff g \geq \gamma_{\text{prox}} R,$$

with $\gamma_{\text{prox}} = 0.75$. Near and far are therefore exact complements, so proximity has no geometric dead zone.

Comparative relation rule. Let $a_X = |M_X|$ denote refined-mask area and let h_X denote its vertical extent. With $\rho_{\text{size}} = 1.05$, the area predicates are

$$A \text{ bigger than } B \iff a_A > \rho_{\text{size}} a_B,$$
$$A \text{ smaller than } B \iff a_B > \rho_{\text{size}} a_A,$$

and the height predicates are

$$A \text{ taller than } B \iff h_A > \rho_{\text{size}} h_B,$$
$$A \text{ shorter than } B \iff h_B > \rho_{\text{size}} h_A.$$

If neither direction exceeds the ratio margin, the evaluator emits no comparative predicate. These relations measure visible in-image size rather than real-world physical size.

Depth relation rule. Depth is estimated with `depth-anything/Depth-Anything-V2-Small-hf`, whose disparity output δ increases for surfaces closer to the camera.

Let M_A, M_B be the refined masks. Define the mask gap

$$g = \min_{p \in M_A,\, q \in M_B} \|p - q\|_2$$

and the pair's mean equivalent-circle radius

$$R = \frac{1}{2}\left(\sqrt{\frac{|M_A|}{\pi}} + \sqrt{\frac{|M_B|}{\pi}}\right).$$

For a fixed facing-window width $w = 16$ px, the facing surfaces are

$$F_A = \{p \in M_A : d(p, M_B) \leq g + w\},$$
$$F_B = \{q \in M_B : d(q, M_A) \leq g + w\},$$

where $d(\cdot, M)$ is Euclidean distance to the nearest pixel of mask M. Let $\bar{\delta}_A$ and $\bar{\delta}_B$ be the mean disparity over F_A and F_B, and let Δ_{img} be the disparity range of the image. The evaluator commits to a depth predicate only if

$$\frac{|\bar{\delta}_A - \bar{\delta}_B|}{\Delta_{\text{img}}} \geq \beta_0 + \beta_1 \frac{g}{R},$$

where $\beta_0 = 0.05$, $\beta_1 = 0.10$. If the condition holds, $\bar{\delta}_A > \bar{\delta}_B$ yields A *in front of* B, while $\bar{\delta}_A < \bar{\delta}_B$ yields A *behind* B. Otherwise no depth predicate is emitted.

This coupling makes the required depth evidence increase with spatial separation: touching objects require only the base margin, while distant pairs must differ more strongly in disparity before the evaluator commits.

Table 15: **Current evaluator configuration used for generator benchmark scoring.** Attribute buffers are measured in $100\times$ cosine-similarity units.

Stage	Setting	Current value
Detection	Backbone	RF-DETR-XL (`RFDETRSegXLarge`)
Detection	Presence threshold	0.03
Detection	Counting threshold	0.5
Detection	Per-class cap	16
Detection	Overlap suppression	disabled (IoU $= 1.0$)
Masks	Refiner	SAM 2.1 Hiera Small
Masks	Closing radius	5 px
Masks	Minimum component	$0.05\times$ largest component
Masks	Refined-mask scope	prompt instances in all four relation categories
Masks	Box-prompt chunk size	50
Attributes	Backbone / weights	ViT-SO400M-14-SigLIP2-378 / webli
Attributes	Crop mask	detector mask; no SAM
Attributes	Confusion buffer β	0.5 (color), 0 (material)
Attributes	Additional color answer	gray (not requested in prompts)
Directional	Extent margin τ_{dir}	0.075
Directional	Axis threshold α_{dir}	0.45
Proximity	Gap factor γ_{prox}	0.75
Comparative	Size ratio ρ_{size}	1.05
Depth	Rule	facing surfaces with coupled confidence
Depth	Facing window w	16 px
Depth	Base margin β_0	0.05
Depth	Separation slope β_1	0.10

Positive, negated, and dead-zone semantics. For each relation category the evaluator returns the set of predicates that fire for a given object pair. An asserted statement passes if and only if its requested predicate belongs to this set; a negated statement passes if and only if its denied predicate does not.

Directional, comparative, and depth contain a deliberate dead zone in which no predicate fires because the measured geometry is not sufficiently decisive. A positive statement fails in such a region, whereas a negated statement passes because the denied predicate is absent. Proximity has no dead zone by construction because its two predicates are exact complements. Attribute confusion sets are also never empty because they always include the top-scoring value.

C.2 Evaluator configuration

Table 15 summarizes the evaluator used for all generator benchmark scores. RF-DETR-XL provides detections and instance masks; SAM 2 refines masks for relation measurements only. SigLIP2 scores detector-masked attribute crops without SAM, while Depth-Anything V2 provides depth estimates. No benchmark verdict uses a generative language-model judge. The decision rules are detailed in Appendix C.1.

The ablation varies the detector, SAM refinement, and relation-rule configuration. Table 16 lists the previous relation constants; this axis leaves the SigLIP2 attribute configuration unchanged. Detector-specific thresholds and SAM usage are controlled by their respective ablation axes.

C.3 Metrics and failure definitions

Metrics. Let $\mathcal{I}$ be the set of generated images and $\mathrm{correct}(I) \in \{0, 1\}$ denote complete-prompt correctness. **Image accuracy** is

$$\mathrm{Acc}_{\mathrm{img}} = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \mathrm{correct}(I).$$

Table 16: **Previous relation-rule configuration used in the evaluator ablation.** These settings replace the corresponding current relation rules; attribute scoring is unchanged.

Stage	Setting	Previous value
Directional	Extent margin τ_{dir}	0.10
Directional	Axis threshold α_{dir}	0.50
Proximity	Gap factor γ_{prox}	0.50
Comparative	Size ratio ρ_{size}	1.15
Depth	Rule	contact-band rule
Depth	Contact-band width	12 px
Depth	Minimum rim/contact pixels	20

Table 17: **Full evaluator ablation against human annotations.**

Config.	Detector	SAM	Constants	n	Agreement (%)	κ
C6	RF-DETR-XL	✓	Current	140	88.5	0.732
C8	RF-DETR-XL	–	Current	140	87.1	0.698
C2	Mask R-CNN	✓	Current	140	85.6	0.654
C4	Mask R-CNN	–	Current	140	84.9	0.640
C7	RF-DETR-XL	–	Previous	140	82.0	0.578
C1	Mask R-CNN	✓	Previous	140	81.3	0.557
C3	Mask R-CNN	–	Previous	140	81.3	0.557
C5	RF-DETR-XL	✓	Previous	140	81.3	0.564

Let $\mathcal{P}$ denote the prompt set and $\mathcal{I}(p)$ the four images generated for prompt p. **Prompt accuracy** is

$$\text{Acc}_{\text{prompt}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}\left[\exists I \in \mathcal{I}(p) : \text{correct}(I) = 1\right].$$

Failure decomposition. For diagnosis, the evaluator also records why an incorrect sample failed. Statement-level failures are assigned to *undetected object*, *over-count*, *positive violated*, *negated violated*, or *dead-zone failure*; row-level failures distinguish object/count errors, positive violations, negated violations, or both polarities. These diagnostics do not define a separate benchmark score.

D EVALUATOR ABLATIONS

Compared configurations. We ablate three binary design choices: the object detector (Mask R-CNN (He et al., 2017) vs. RF-DETR-XL (Robinson et al., 2026)), SAM refinement (on vs. off), and the relation-rule constants (previous vs. current). This yields: C1 = Mask R-CNN + SAM + previous constants; C2 = Mask R-CNN + SAM + current constants; C3 = Mask R-CNN + no SAM + previous constants; C4 = Mask R-CNN + no SAM + current constants; C5 = RF-DETR-XL + SAM + previous constants; C6 = RF-DETR-XL + SAM + current constants; C7 = RF-DETR-XL + no SAM + previous constants; and C8 = RF-DETR-XL + no SAM + current constants. Table 17 reports agreement with human annotations on the 140 examples with a clean human-majority verdict. We use C6 throughout NEGT2IBENCH.

Threshold selection. All configurations are scored with the evaluator from §3.1: RF-DETR-XL at presence/counting thresholds $0.03/0.5$, SAM 2-refined masks for all relation evaluations, SigLIP 2 for color/material, and Depth-Anything V2 for depth. The geometric cut points and attribute confusion buffer were fitted against 140 human-annotated images and then fixed. The evaluator is deterministic and contains no language-model judge.

The ablation was sampled from 140 of 600 annotated images distributed across model, setting, and category; they have a clear human-majority verdict and enter the headline table. The point estimates should therefore be read as a compact diagnostic of evaluator design rather than as a precise ranking of configurations. SAM refinement is retained in C6 to tighten object masks for geometric relation measurements, including area, centroid, spacing, and depth computations.

Table 18: **Generation settings used in the benchmark sweep.** Resolutions are pipeline defaults. Janus-Pro-7B is autoregressive, so a diffusion step count is not applicable.

Model	Steps	Resolution
FLUX.2-klein-9B	4	1024^2
HunyuanImage-2.1 (distilled)	8	1024^2
Z-Image-Turbo	9	1024^2
FLUX.2-dev (4-bit)	28	1024^2
SD 3.5 Large	28	1024^2
BLIP3-o NEXT SFT 3B	30	1024^2
HiDream-I1 (4-bit)	50	1024^2
Qwen-Image	50	1024^2
SD 1.5	50	512^2
SD 1.5 + NegationCLIP	50	512^2
Janus-Pro-7B	–	384^2

E GENERATOR EXPERIMENTS

E.1 EXPERIMENTAL SETUP

Models and sample counts. We evaluate eleven model configurations on the full NEGT2IBENCH corpus: BLIP3-o NEXT SFT 3B (Chen et al., 2025a), FLUX.2-klein-9B and FLUX.2-dev (4-bit) (Black Forest Labs, 2026), HiDream-I1 (4-bit) (Cai et al., 2025b), HunyuanImage-2.1 (distilled) (Tencent Hunyuan Team, 2025), Janus-Pro-7B (Chen et al., 2025b), Qwen-Image (Wu et al., 2025), SD 3.5 Large (Esser et al., 2024), Z-Image-Turbo (Cai et al., 2025a), and SD 1.5 (Rombach et al., 2022) with and without the negation-aware text encoder of Park et al. (2025). Each configuration generates four samples for the same 4,800 prompts, yielding 19,200 evaluated images per configuration and 211,200 images in total.

Generation settings. All model configurations use the same NEGT2IBENCH prompt files and four samples per prompt with seeds $42, \ldots, 45$. We intentionally use each pipeline's native generation configuration rather than forcing an equal inference-compute budget. Consequently, absolute between-model comparisons also reflect differences in inference schedule and output resolution; within-model comparisons across benchmark settings do not.

No model-specific prompt engineering is applied, and the CFG negative-prompt field is unused. The SD 1.5 and SD 1.5+NegationCLIP entries share the same base checkpoint and 50-step schedule, but their registered guidance scales are 9.0 and 7.5 respectively; they should therefore not be interpreted as a perfectly controlled encoder intervention.

E.2 ADDITIONAL RESULTS

Failure analysis. The main paper reports failure composition by benchmark setting and semantic category. Here we provide the corresponding statement-count calculation and additional image- and model-level diagnostics. Across the full benchmark there are 475,200 statement evaluations. For one prompt from each of the eight settings, the total statement load is

$$1 + 1 + 2 + 2 + 2 + 3 + 3 + 4 = 18.$$

With six categories, 100 prompts per category–setting cell, four generated samples per prompt, and eleven model configurations, this gives

$$6 \times 100 \times 4 \times 11 \times 18 = 475{,}200$$

statement evaluations. Of these, 280,067 are failed statements; the statement-level percentages in Section 5 are computed over these failed statements, or over the corresponding failed statements within each setting or category.

At the image level, 158,421 of 211,200 generations are incorrect. Among these, 40.1% contain only negated-statement violations, 25.9% only positive-statement violations, and 13.9% violations of both polarities. Over-counting accounts for a further 10.5% and undetected objects for 9.6%.

Table 19: **Dead-zone exposure among successful negated relations.** A dead-zone pass occurs when a negated statement succeeds while the relation evaluator returns no predicate for that object pair.

Category	Evaluable	Passed	Dead-zone passes	Share of passes
Depth	30,335	17,947	9,963	**55.5%**
Directional	29,777	17,117	1,567	9.2%
Comparative	29,883	12,962	374	2.9%
Proximity	30,620	14,764	3	0.0%
Total	120,615	62,790	11,907	19.0%

Images that violate both positive and negated statements are retained as a separate category rather than assigned to either polarity.

Dead-zone analysis. Directional, comparative, and depth relations contain a deliberate *dead zone*: when the measured geometry is insufficient to support any predicate in the category, the evaluator returns an empty predicate set rather than forcing a decision. This can be the correct outcome; for example, two objects at approximately the same depth may be neither clearly *in front of* nor *behind* one another.

Let r denote the relation predicate a statement names (e.g., *left of*) and $R(A, B)$ the set of predicates the evaluator returns for the object pair (A, B). A positive statement passes if and only if $r \in R(A, B)$, whereas a negated statement passes if and only if $r \notin R(A, B)$. Consequently, when $R(A, B) = \emptyset$, a positive statement fails while a negated statement passes. We call the latter a *dead-zone pass*. Such a pass is valid under the benchmark rule, but differs from a pass in which an alternative predicate is positively established.

Across the full experiment, 158,400 negated relation statements are evaluated across the 211,200 generated images. Of these, shown in Table 19, 37,785 are unevaluable because an endpoint is missing or over-counted. Among the remaining evaluable statements, 62,790 pass, and 11,907 of those passes (19.0%) arise from the dead zone. Equivalently, dead-zone passes account for 7.5% of all negated relation statements.

The effect is concentrated in depth: 55.5% of successful negated depth statements occur when no depth predicate is resolved. This does not imply that those images are incorrectly scored; an object pair may genuinely lack a clear front–behind ordering. It does mean, however, that such cases provide weaker evidence of negation following than cases in which the evaluator positively establishes the alternative relation.

Proximity provides a useful contrast. Its *near/far* predicates are exact complements under a single decision boundary, so it has no designed dead zone. The three observed empty-set cases are degenerate geometry rather than a consequence of the decision rule.

Negation-aware text encoder. Replacing SD 1.5's text encoder with NegationCLIP changes overall image accuracy only from 14.5% to 14.9%. Both SD 1.5 configurations are unusual in scoring higher on `pos0_neg1` than on `pos1_neg0`, but the pair remains well below the other configurations overall. This comparison is not fully controlled: the checkpoint and 50-step schedule are shared, but the registered guidance scales differ (9.0 vs. 7.5). We therefore treat the pair as suggestive rather than as an isolated causal test of the text encoder.

E.3 Additional qualitative examples.

Figure 8 presents selected generations for six NegT2IBench prompts, one from each semantic category, across all eleven evaluated T2I models. Showing each model's response to the same prompt illustrates differences in rendering requested properties and excluding forbidden ones while preserving the named objects. Each image is also scored by our evaluator, with pass/fail marks indicating whether it satisfies the complete prompt. These selected examples complement the quantitative results by making the benchmark's requirements and the models' successes and failures concrete.

F LIMITATIONS

NEGT2IBENCH is limited to what can be verified automatically in both polarities: color, material, and four spatial and size relations over 79 COCO classes. Contact, interactions such as holding or riding, and the negated presence of whole objects are not covered, and are natural extensions as detectors improve. Comparative relations also measure apparent image size rather than real-world size. Negation is expressed in one English form ("non-*value*", "is not *relation*"), and relation prompts link one subject to every target. This keeps polarity comparisons controlled, but robustness to paraphrases, other languages, and richer relation structures is not tested; these can be added on the same evaluation pipeline.

NEGT2IBENCH measures whether negated content is absent from the image, not whether the model understood the negation. Because negated statements are easier to satisfy by chance than positive ones, a model that ignores a negation can still pass some negated statements. A control prompt that omits the tested attribute or relation would separate the two, and fits the current design. Scores depend on off-the-shelf perception models (RF-DETR, SAM 2, SigLIP 2, and Depth-Anything V2), so benchmark accuracy is bounded by their quality; the modular design lets each component be replaced as stronger models appear. Finally, we evaluate open-weight models under their native inference settings rather than a matched compute budget; closed systems can be scored with the same protocol.

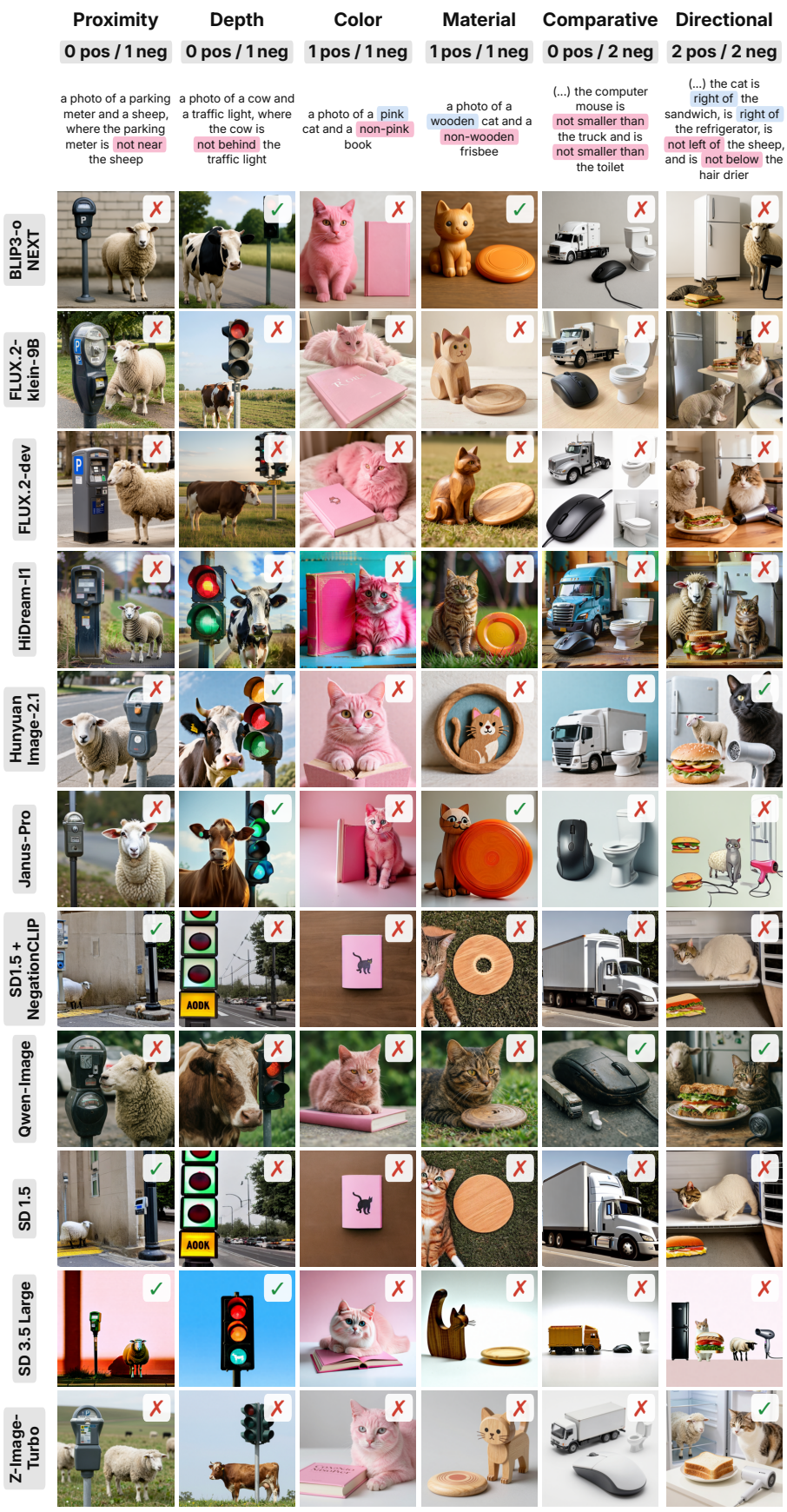


Figure 8: **Qualitative examples.** Six selected prompts (one per category) generated by all eleven T2I models.