SUBTEXT: CAN IMAGE GENERATORS INFER WHAT CAPTIONS LEAVE IMPLICIT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Can image generators infer an outcome that a caption implies but does not state? Captions routinely omit direct descriptions of outcomes and rely on the reader to infer them from context. For example, a caption describing a chocolate bar left on a dashboard on a hot afternoon implies that the bar has melted. We introduce SUBTEXT, a benchmark of 140 minimal image pairs that differ in one physical or socio-emotional outcome within an otherwise matched scene. Each image has an implicit caption that specifies only the antecedent context and an explicit caption that is identical except for an inserted outcome term such as "melted". Following work on diffusion classifiers, we use seven frozen image generators as zero-shot classifiers that match captions to images and images to captions. We compare them with three contrastive models, Claude Fable 5.1, and human participants. With implicit captions, the generators select the correct image at near-chance rates and fall below chance when both captions of a pair must be matched. Human participants and Claude Fable 5.1 select the correct image in at least 93% of cases. Explicit captions significantly improve accuracy for six of the seven generators, which indicates that inference adds difficulty beyond visual discrimination. Unlike prior results on compositional benchmarks, the generators show no advantage over contrastive models, even when the two share a text encoder. These results suggest that learning to generate realistic images does not by itself teach a model to reliably link context to visible outcomes. SUBTEXT therefore provides a controlled test of whether future generators close the gap between what a caption states and what it implies.

1 INTRODUCTION

Language comprehension routinely relies on inferences that go beyond the literal content of a text (Graesser et al., 1994; McKoon & Ratcliff, 1992). For example, a bicycle chained outside for several winters implies rust, and an unwanted gift implies disappointment. Writers tend to leave unstated what a reader can infer, and image captions show the same bias (Gordon & Van Durme, 2013; Misra et al., 2016). Aligning captions with images often requires inferring visual content from context rather than from direct description.

Frozen image generators can serve as zero-shot classifiers that score how well a caption fits an image by how much the caption reduces their denoising error. Contrastive models such as CLIP ViT-L/14 instead score this fit by embedding similarity and are trained explicitly to align images with text (Radford et al., 2021). Prior evaluations of these generative classifiers focused on object recognition and compositional structure such as word order and attribute binding (Krojer et al., 2023; Li et al., 2023; Clark & Jaini, 2023). On compositional benchmarks, these classifiers matched or exceeded contrastive models, although later work found the advantage to be task dependent (Jeong et al., 2025). In generation, Niu et al. (2026) found that nearly all tested generators produced substantially better images when indirect prompts were rewritten as explicit ones. Realistic generation is sometimes taken as evidence of learned world knowledge (OpenAI, 2024). We extend the evaluation of generative classifiers from stated to implied content, which tests whether generative training links contextual cues in a caption to their visual consequences.

Existing benchmarks probe inference beyond visible content in vision-language models through question answering and abductive reasoning (Zellers et al., 2019; Hessel et al., 2022). Related

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.



Image A

Implicit The chocolate bar had been left on the dashboard during a hot afternoon.

Explicit The **melted** chocolate bar had been left on the dashboard during a hot afternoon.



Image B

Implicit The chocolate bar had been left on the dashboard during a chilly morning.

Explicit The **solid** chocolate bar had been left on the dashboard during a chilly morning.

Figure 1: The two images form one example pair from the benchmark. Both images show the same chocolate bar on a car dashboard and differ only in whether the bar has melted. Each image has an implicit caption that gives only the context and an explicit caption that adds the outcome word, shown in bold.

benchmarks assess such inference in generators by evaluating their outputs with a judge (Fu et al., 2024; Niu et al., 2026). To our knowledge, generative classifiers have not been evaluated on minimal image pairs that isolate a single implied outcome.

We introduce SUBTEXT, a benchmark of 140 minimal image pairs in which a single physical or socio-emotional outcome varies while the scene is held constant. Each image is paired with an implicit caption that conveys the outcome only through context and an explicit caption that states it (Figure 1). We evaluate seven frozen image generators as zero-shot classifiers, with three contrastive models, Claude Fable 5.1, and human participants as baselines.

We find that generative classifiers largely fail to recover implied outcomes. With implicit captions, image-choice accuracy ranges from 48.6% to 55.7% across generators against a chance level of 50% (Table 3). The benchmark itself is solvable, as human participants attain 93.0% and Claude Fable 5.1 (Anthropic, 2026) reaches 99.3%. Explicit captions significantly raise accuracy for six of seven generators, so inferring an outcome is harder for them than matching a stated one. Even so, canceling a fixed preference for one image reveals only a weak signal from implicit captions (Section 5.1).

On object-state and material pairs, generators favor the image in which nothing has changed, which suggests a default toward typical states when no outcome is stated (Figure 3). Contrastive models score modestly higher (56.4% to 60.0%), including the two whose text encoders a generator shares. For applications that use generators as zero-shot classifiers, captions that leave an outcome implicit may provide little usable signal.

2 Related work

Inference beyond what is visible. A substantial body of work evaluates inference beyond the visible content of images and videos. VCR (Zellers et al., 2019) and Sherlock (Hessel et al., 2022) show that models trail humans in inferring mental states and the implications of visual clues, and VisualCOMET (Park et al., 2020) targets events before and after the depicted moment. Social-IQ (Zadeh et al., 2019), Social Genome (Mathur et al., 2025), and EMOTIC (Kosti et al., 2019) assess social and emotional inference from context. On Spot The Ball (Balamurugan et al., 2026), humans localize a removed ball from gaze and body pose, and vision-language models rely on spatial heuristics. Physion (Bear et al., 2021) likewise finds that vision models lag far behind humans in

predicting how physical scenes unfold. These benchmarks pose inference as a question about a given image or video, whereas SUBTEXT places the context in the caption and the outcome in the image. Matching in either direction therefore requires inference.

Minimal-pair image and caption benchmarks. Minimal-pair benchmarks vary a single detail of an image or caption so that it alone determines the correct match. Winoground (Thrush et al., 2022) pairs two images with two captions that contain the same words in a different order. EqBen (Wang et al., 2023) and ARO (Yuksekgonul et al., 2023) find that vision-language models perform poorly on such changes, and VisMin (Awal et al., 2024) reports higher accuracy in selecting captions than in selecting images. SugarCrepe (Hsieh et al., 2023) shows that text-only models outperform vision-language models on earlier compositional benchmarks. Additionally, NaturalBench (Li et al., 2024) finds that models often give the same answer for both images of a pair, a bias that parallels the fixed image preference we observe. In these benchmarks, the text directly describes the distinguishing content, whereas our implicit captions leave it unstated.

Generators used as classifiers. DiffusionITM (Krojer et al., 2023) and Diffusion Classifier (Li et al., 2023) surpass CLIP on compositional tasks, and Clark & Jaini (2023) find such classifiers competitive with CLIP in zero-shot classification and stronger on some attribute-binding tasks. Jaini et al. (2024) report near-human accuracy on out-of-distribution images and argue for generative pre-training as an alternative to contrastive training. In contrast, Li et al. (2023) find their classifier weaker than discriminative models at zero-shot recognition, and Jeong et al. (2025) find that such classifiers excel at spatial position but trail CLIP on object recognition and counting. These evaluations target content that a caption states, and SUBTEXT extends them to content that a caption implies.

Implied content in generation. Commonsense-T2I (Fu et al., 2024), WISE (Niu et al., 2026), and T2I-ReasonBench (Sun et al., 2026) use a judge to assess whether generated images depict what their prompts imply. Yang et al. (2024) argue that text-to-image benchmarks neglect implicit prompts and find that generators can accurately depict the symbols such prompts hint at. These benchmarks evaluate what a generator draws. In contrast, SUBTEXT scores fixed images without a judge, which allows an explicit caption on the same images to serve as a control.

3 BENCHMARK

3.1 DESIGN

Each pair comprises two images of the same scene that differ in a single outcome, such as an object's physical state or a person's emotional state. Each image is the target of one implicit and one explicit caption (Figure 1). Each explicit caption is its implicit caption with an outcome term inserted. In socio-emotional pairs, this term is a single adjective that names an emotional state or interpersonal manner. In physical pairs, it is a phrase of one to three words that names the resulting state of an object or scene.

3.2 CATEGORIES

Physical pairs span object state (41 pairs), material outcomes (17 pairs), and event-level scenes (12 pairs). Object state pairs depict one object in two states, such as a soggy and a dry newspaper. Material outcome pairs depict objects of different materials with and without an event that affects them differently, such as a fall that breaks a ceramic mug but not a plastic cup. Event-level scene pairs depict a scene with and without traces of an event, such as a ransacked and a tidy apartment. In physical pairs, image A always shows the changed state. Socio-emotional pairs span emotion (38 pairs) and interpersonal manner (32 pairs). Emotion pairs depict a person whose emotional state follows from the stated circumstances, such as an overjoyed and a disheartened boy. Interpersonal manner pairs show how one person treats another, such as welcoming or threatening.[1] Overall, the 140 pairs comprise 280 images and 560 captions.

[1]Two interpersonal manner pairs contrast emotional reactions rather than manners toward another person.

Table 1: The table lists all evaluated models. Size counts the denoiser of each generator and the image encoder of each contrastive model.

Model	Type	Size	Text encoder	Scored at
SD 1.5	Diffusion (UNet)	0.86B	CLIP ViT-L/14	512 px
SD 2.1 Base	Diffusion (UNet)	0.87B	OpenCLIP ViT-H/14	512 px
SDXL	Diffusion (UNet)	2.57B	CLIP-L, OpenCLIP-bigG	1024 px
SD3 Medium	Flow (MMDiT)	2.03B	CLIP-L, OpenCLIP-bigG, T5-XXL	1024 px
SD3.5 Large	Flow (MMDiT)	8.06B	CLIP-L, OpenCLIP-bigG, T5-XXL	1024 px
FLUX.2 dev	Flow (transformer)	32.2B	Mistral Small 3.2 (24B)	1024 px
Qwen-Image 2.1	Flow (DiT)	7.12B	Qwen3-VL (8B)	1024 px
CLIP ViT-L/14	Contrastive	0.30B	Own text tower	224 px
OpenCLIP ViT-H/14	Contrastive	0.63B	Own text tower	224 px
SigLIP 2	Contrastive	0.43B	Own text tower	384 px
Claude Fable 5.1	Hosted multimodal	not disclosed	n/a	Native

3.3 Construction

All 280 images were generated with gpt-image-2 (OpenAI, 2026), and many pairs were constructed by editing one image into the other so that the two differ only in the target outcome. The 560 captions were written by the authors, and both images and captions were refined over several iterations. Images were regenerated or edited until they appeared photorealistic and depicted the intended contrast unambiguously. Implicit captions were revised until the described context supported the intended inference without stating the outcome. A subset of images was generated at 1024×1024 pixels and upsampled to 1254×1254 with Lanczos resampling, in some cases following Real-ESRGAN super-resolution (Wang et al., 2021). This procedure introduced a resolution mismatch within 21 pairs, and excluding these pairs leaves the main results unchanged, with implicit image choice at 49.2% to 55.0% for the generators (Appendix A). The dataset is distributed under the CC BY 4.0 license.

4 Evaluation

4.1 Models and scoring

We evaluate the seven frozen image generators in Table 1 (Rombach et al., 2022; Podell et al., 2024; Stability AI, 2023; Esser et al., 2024; Stability AI, 2024a;b; Black Forest Labs, 2025; Qwen, 2026). Following DiffusionITM (Krojer et al., 2023), each generator scores a caption c against an image x by how much c reduces its denoising error, giving the score $s(c, x) = e(\emptyset, x) - e(c, x)$.

Here $e(c, x)$ is the mean ℓ_2 error of the generator's noise prediction for x conditioned on c, and $\emptyset$ is a blank caption. The error is averaged over 245 fixed noise levels with one Gaussian draw per level, which every image and caption share (Appendix A). For the four flow transformers, we compute the same error from the noise prediction implied by their velocity output. The three contrastive models (Radford et al., 2021; Cherti et al., 2023; Tschannen et al., 2025) instead score a caption against an image by the cosine similarity of their embeddings. Every model then selects the image or caption with the higher score. To validate the generator scoring, we reproduce DiffusionITM on 200 sampled Flickr30K queries (Young et al., 2014), where SD 2.1 Base reaches 65.0% retrieval accuracy. The published value of 61.5% lies inside our 95% Wilson interval (Appendix B).

4.2 Claude Fable 5.1 and human participants

Claude Fable 5.1 (Anthropic, 2026) answered every question in both display orders at the high effort setting (Appendix A). For the human baseline, 122 participants recruited through Prolific (Palan & Schitter, 2018) answered the questions of 139 of the 140 pairs (Appendix C). Most questions received three responses, and every question received two to four.

4.3 MEASURES

Every measure is computed separately for implicit and explicit captions, and c_A and c_B denote the target captions of images x_A and x_B.

Pair scores. Following Winoground (Thrush et al., 2022), three pair scores evaluate each pair as a whole. The image score requires the target image for both captions, the text score requires the target caption for both images, and the group score requires both.

Swap score. Because a model with a fixed preference for one image fails the image score on that pair, the swap score instead counts a pair as correct when $s(c_A, x_A) + s(c_B, x_B) > s(c_A, x_B) + s(c_B, x_A)$, which cancels any such preference.

Single-answer measures. Image choice is the fraction of captions for which the target image is selected, and caption choice is the fraction of images for which the target caption is selected.

Chance and baselines. For the generators and contrastive models, chance is 25.0% for the image and text scores and 16.7% for the group score. It is 50.0% for the swap score, image choice, and caption choice. Claude Fable 5.1 receives pair-score credit for a question only when it answers both display orders correctly, and its single-answer measures average the two orders. Human pair scores use a majority vote per question with ties counted as half, and human single-answer measures average the fraction of correct answers per question. These rules change the pair-score chance levels reported in the notes of Table 2. The swap score requires continuous scores and is therefore undefined for Claude Fable 5.1 and human participants.

4.4 STATISTICS

Unless stated otherwise, 95% confidence intervals are percentile bootstrap intervals over pairs with 10,000 resamples (Efron, 1979). Sign tests over pairs compare explicit with implicit captions on each measure and compare each implicit single-answer measure with chance. We use binomial tests to compare the pair scores and the swap score with chance, and permutation tests compare the gain from explicit captions between physical and socio-emotional pairs. Holm correction (Holm, 1979) treats the pair-score tests and the single-answer tests of each model as two separate families, and all other tests are uncorrected. Human scores receive intervals but no tests.

5 RESULTS

5.1 PAIR SCORES AND SWAP SCORE

With implicit captions, all seven generators score below chance on the three pair scores (Table 2). The low implicit scores reflect fixed preferences, as each generator selects the same image for both implicit captions in 76.4% to 90.7% of pairs and the same caption for both images in 82.1% to 87.1%. Human participants and Claude Fable 5.1 select the same image for both implicit captions in only 13.6% (95% CI 10.0 to 17.5) and 1.4% (95% CI 0.0 to 3.2) of pairs. Once the swap score cancels these preferences, four generators exceed its 50.0% chance level with implicit captions. These four reach 62.1% to 74.3%, and uncorrected exact binomial tests give $p \leq 0.0052$ for each. SD3 Medium, FLUX.2 dev, and Qwen-Image 2.1 remain indistinguishable from chance at 49.3% to 56.4% ($p \geq 0.15$).

Explicit captions numerically raise every pair score for every generator. After Holm correction, exact sign tests confirm this gain in the image score for six generators and in the text and group scores for four. With explicit captions, the swap score rises to 60.7% to 89.3% and exceeds its implicit value for all seven generators and all three contrastive models (uncorrected exact sign tests, $p \leq 0.030$). This gain holds for every generator under every blank-corrected scoring variant (Appendix A).

Table 2: Pair scores and swap scores in percent of pairs (Section 4.3).

	Image score		Text score		Group score		Swap score	
Model	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit
Chance	*25.0*	*25.0*	*25.0*	*25.0*	*16.7*	*16.7*	*50.0*	*50.0*
SD 1.5	13.6	28.6	11.4	25.7	2.9	10.7	67.9	86.4
SD 2.1 Base	12.9	27.9	10.7	22.1	1.4	9.3	67.1	89.3
SDXL	12.9	31.4	10.7	17.9	3.6	10.7	74.3	86.4
SD3 Medium	5.0	14.3	9.3	17.1	2.1	5.7	56.4	73.6
SD3.5 Large	12.9	26.4	13.6	27.9	5.0	14.3	62.1	85.7
FLUX.2 dev	13.6	17.9	8.6	10.7	2.9	4.3	49.3	60.7
Qwen-Image 2.1	6.4	25.7	8.6	24.3	0.7	11.4	53.6	77.9
CLIP ViT-L/14	24.3	36.4	21.4	37.9	10.7	19.3	77.9	85.0
OpenCLIP ViT-H/14	16.4	32.1	22.9	39.3	5.7	20.0	74.3	91.4
SigLIP 2	17.9	40.0	17.9	40.0	4.3	20.7	72.1	92.9
Claude Fable 5.1[†]	97.9	100.0	85.7	94.3	85.0	94.3	n/a	n/a
Human participants[‡]	92.1	97.8	93.2	94.6	85.3	92.4	n/a	n/a

The swap score is undefined for Claude Fable 5.1 and human participants. [†]Both display orders must be correct, and chance is 6.3 (image, text) and 0.4 (group). [‡]A majority vote decides each question over 139 pairs, and chance is 25.0 (image, text) and 6.3 (group).

Table 3: The table gives image choice and caption choice in percent of questions, with chance at 50.0 (Section 4.3).

	Image choice		Caption choice	
Model	Implicit	Explicit	Implicit	Explicit
Chance	*50.0*	*50.0*	*50.0*	*50.0*
SD 1.5	55.0	63.9	53.9	62.5
SD 2.1 Base	54.3	63.2	53.9	60.7
SDXL	55.7	65.4	53.6	58.2
SD3 Medium	50.4	55.7	52.1	57.5
SD3.5 Large	54.6	62.1	54.6	62.5
FLUX.2 dev	51.8	56.1	52.1	52.1
Qwen-Image 2.1	48.6	61.1	50.7	59.3
CLIP ViT-L/14	60.0	66.8	60.0	68.6
OpenCLIP ViT-H/14	56.4	65.7	60.7	69.6
SigLIP 2	56.8	70.0	57.9	70.0
Claude Fable 5.1[†]	99.3	100.0	94.6	97.9
Human participants[‡]	93.0	97.2	90.9	93.7

[†]Values average both display orders. [‡]Values cover 139 pairs.

5.2 Single-answer measures and baselines

With implicit captions, generator image choice ranges from 48.6% to 55.7% against a chance level of 50.0% (Table 3). Exact sign tests with Holm correction place it above chance only for SD 1.5, SDXL, and SD3.5 Large ($p \leq 0.043$). The same tests detect a gain in image choice from explicit captions for six of the seven generators, ranging from 5.3 to 12.5 points ($p \leq 0.0079$). FLUX.2 dev is the only exception, with a gain of 4.3 points (95% CI -0.7 to 9.3, $p = 0.76$).

Human participants and Claude Fable 5.1 far exceed the generators on both single-answer measures. In implicit image choice, human participants reach 93.0% (95% CI 90.9 to 94.8) and Claude Fable 5.1 reaches 99.3% (95% CI 98.4 to 100.0). In implicit caption choice, they reach 90.9% (95% CI 88.7 to 93.0) and 94.6% (95% CI 92.1 to 96.8), respectively.

The contrastive models also exceed every generator in implicit image choice, at 56.4% to 60.0%. SD 1.5 and SD 2.1 Base share text-encoder weights with CLIP ViT-L/14 and OpenCLIP ViT-H/14,

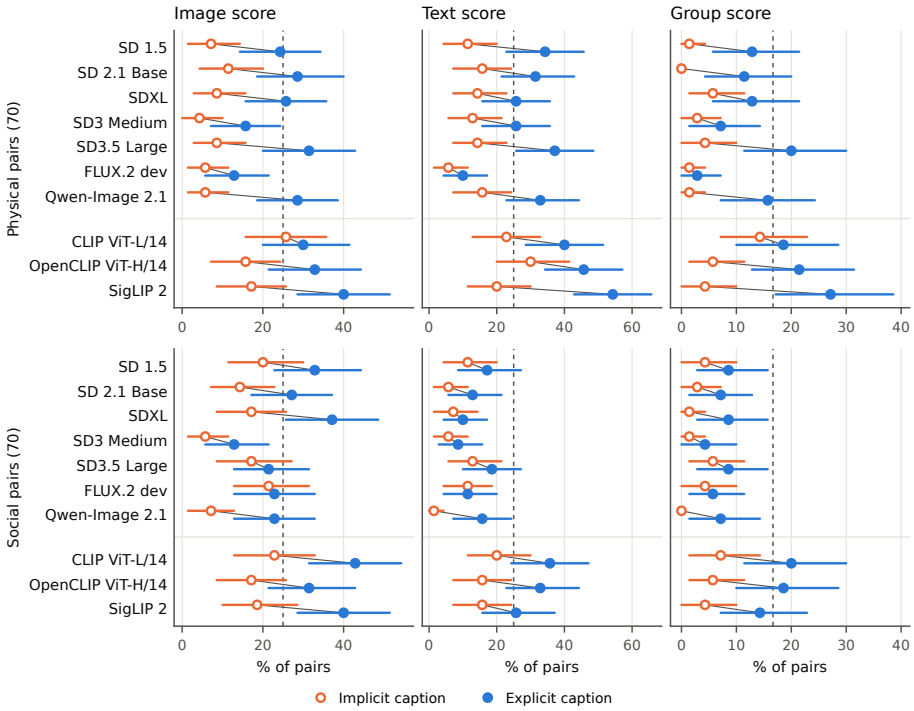


Figure 2: Each panel shows one pair score for the physical pairs (top) and the socio-emotional pairs (bottom), with implicit and explicit captions for each model. Error bars show 95% bootstrap confidence intervals, and dashed lines mark chance.

respectively, which permits a direct comparison (Appendix B). In implicit image choice, the contrastive model leads by 5.0 points for SD 1.5 (95% CI -0.4 to 10.4, $p = 0.11$) and by 2.1 points for SD 2.1 Base (95% CI -2.5 to 6.8, $p = 0.54$). Neither difference reaches significance in uncorrected exact sign tests. On all three pair scores, both generators score below their matched contrastive models with implicit and explicit captions alike (Table 2). Uncorrected exact sign tests detect this gap in five of six comparisons for SD 1.5 and three of six for SD 2.1 Base ($p < 0.05$). For instance, the implicit group score of SD 1.5 falls 7.9 points below that of CLIP ViT-L/14 (95% CI 2.1 to 13.6, $p = 0.013$).

5.3 Physical versus socio-emotional pairs

The gain from explicit captions does not differ detectably between the 70 physical and the 70 socio-emotional pairs for any generator (Figure 2). Permutation tests with Holm correction give $p \geq 0.073$ for each generator on the pair scores and single-answer measures. Among the contrastive models, these tests detect such a difference only for SigLIP 2, whose gain is 12.9 points smaller on socio-emotional pairs in caption choice ($p = 0.013$) and 24.3 points smaller in the text score ($p = 0.044$).

6 Discussion

6.1 Separating inference from visual discrimination

Explicit captions keep both images fixed and add only the outcome term, so their gain for most generators (Section 5.2) locates an added difficulty in inference rather than in faint depictions of the outcome. Image choice with explicit captions nonetheless remains modest (Table 3), so matching a stated outcome to the images is also difficult for the generators.

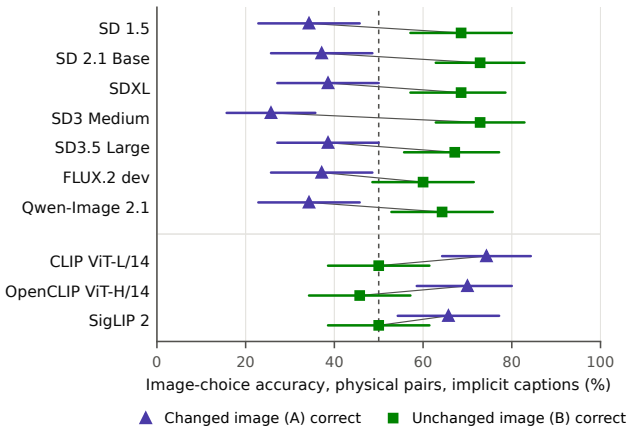


Figure 3: Each row shows one model's implicit image choice on the 70 physical pairs, split by whether the changed image A or the unchanged image B is correct. Error bars show 95% bootstrap confidence intervals, and the dashed line marks chance.

6.2 Opposite leans on physical pairs

On physical pairs with implicit captions, every generator is more accurate when the unchanged image B is correct (Figure 3). For each generator, the 95% confidence interval of this accuracy difference excludes zero. Every generator shows this lean on object state and material outcome pairs, but five of the seven reverse it on event-level scene pairs. On the physical pairs as a whole, contrastive models are instead more accurate when the changed image A is correct. We interpret this lean as a preference for the more typical state of an object when no outcome is stated.

6.3 No advantage over contrastive models

The generators do not outperform the contrastive models (Table 3), which contrasts with earlier reports that generators used as classifiers match or outperform CLIP on some compositional tests (Krojer et al., 2023; Li et al., 2023; Clark & Jaini, 2023). Because the Stable Diffusion model evaluated on Winoground shares its text encoder with OpenCLIP ViT-H/14, its higher text score there has been attributed to better binding of concepts to images (Li et al., 2023). SD 1.5 and SD 2.1 Base instead trail their matched contrastive models on our implicit text score by 10.0 points (95% CI 2.1 to 17.9) and 12.1 points (95% CI 5.0 to 19.3), respectively. Both generators condition on the per-token hidden states of the shared encoder through cross-attention, whereas the contrastive models use one projected embedding of the full caption. The comparison therefore does not isolate generative training (Appendix B).

6.4 Realistic generation and contextual inference

For video generators, prior work cautions that visual realism does not demonstrate adherence to physical principles (Motamed et al., 2026) and finds that models trained on simulated 2D mechanics fail out of distribution (Kang et al., 2025). Our results extend this caution to image generators and socio-emotional outcomes but do not address world models.

6.5 Limitations

Several factors across scope, design, and measurement limit our study. In scope, we evaluate the generators only as classifiers, so the results do not establish whether images generated from implicit captions would show the implied outcome. The benchmark images are also synthetic, and a replication with photographs would show whether the results transfer to natural images. The captions are in English, and both the captions and the images largely reflect Western cultural contexts. Because socio-emotional inferences can differ across languages and cultures, the socio-emotional

results should be read as specific to this setting. In design, the human study is between-subjects, so its pair scores combine the majority votes of different participants rather than the consistency of any one person. The benchmark also contains only 140 pairs, which leaves the confidence intervals within each category wide. Each image has one caption of each type, so the sensitivity of the results to caption wording remains untested. In measurement, the blank-corrected readout is validated against a published result only for SD 2.1 Base, and FLUX.2 dev and Qwen-Image 2.1 score higher under an alternative readout (Appendix A).

7 Conclusion

Realistic image generation is often assumed to include the ability to infer how a situation changes what a scene looks like. SUBTEXT tests this assumption directly, and the seven generators we evaluated show little of this ability. Their difficulty lies in the inference itself, because naming the outcome on the same images improves their accuracy. Evaluations that state outcomes explicitly would therefore miss this gap, and benchmarks for generative models should include captions that leave outcomes implicit.

Future work can test whether the same gap appears when generators produce images from implicit captions, and whether it arises in their text encoders or in the mapping from text to images. Training on implied outcomes would then show whether the gap can be closed directly. Until then, a generator's realism should not be read as evidence that it infers what a caption leaves implicit.

Ethics statement

Human judgments for different parts of the benchmark were collected through separate studies on Prolific. All participants were adults, provided informed consent before participation, and could withdraw at any time by exiting the study. The task consisted of two-alternative forced-choice judgments over everyday scenes and posed no more than minimal risk. All visual stimuli were synthetic images generated by text-to-image models and did not depict real identifiable individuals. Across studies, participants were compensated at effective hourly rates ranging from $13.11 to $19.30 based on observed completion times. The study protocol was approved by the authors' institutional review board (IRB).

We collected only the participants' task data and the Prolific IDs required for payment. The task data were responses, confidence ratings, and response times. Prolific IDs were removed before data release, and released human data were de-identified. No individual responses are released. For each question, the release contains only the number of answers and the number of correct answers.

Reproducibility statement

The supplementary material contains the evaluation code, the benchmark captions with a checksum for every image, and the stored outputs of every model. These outputs include the error of each image generator in removing noise at all 245 noise levels, the similarity scores of each contrastive model, and every reply of Claude Fable 5.1. The anonymous repository at `https://anonymous.4open.science/r/implicit-inference-benchmark-paper-C6AF` contains the same files plus the 280 images and their dataset card. The images and captions are released under the CC BY 4.0 license. Two analysis scripts regenerate every table and results figure from the stored model outputs and the released human answer counts, and the regenerated tables match the released files byte for byte. The scripts run in about 10 seconds and need no GPU, no images, and no network access. Rerunning the scoring requires the images, and the code loads the same version of each generator that we scored. Scoring the seven generators took about 88 GPU-hours on NVIDIA H100 GPUs. The appendix gives the exact scoring recipe for each kind of model (Appendix A), the validity checks (Appendix B), and the details of the human study (Appendix C).

AI use statement

In this work, we used generative AI tools to generate synthetic datasets, to support data analysis, and to interpret results. All 280 benchmark images were generated with gpt-image-2 (OpenAI, 2026), and some were upscaled with Real-ESRGAN (Wang et al., 2021). Theoretical frameworks,

mathematical claims and proofs, and translation are not applicable to this work. Additionally, we used generative AI tools to aid and polish our writing. We have reviewed all AI-assisted work. The authors have inspected every image for realism and for the intended difference, and every number in the paper was checked against the released result files. Every reference was verified against its published source, and the analysis scripts regenerate every results table and figure from the released outputs. We take responsibility for the final content of this work and for all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. Claude Fable 5.1. https://platform.claude.com/docs/en/models/fable-5-1/overview, 2026. Accessed September 25, 2026.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual Minimal-Change Understanding. In *Advances in Neural Information Processing Systems 37*, pp. 107795–107829. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3423. URL http://dx.doi.org/10.52202/079017-3423.

Neha Balamurugan, Sarah Wu, Cristobal Eyzaguirre, and Tobias Gerstenberg. Spot The Ball: A Benchmark for Visual Social Inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30919–30928, 2026.

Daniel Bear, Elias Wang, Damian Mrowca, Felix Jedidja Binder, Hsiao-Yu Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel LK Yamins, and Judith E Fan. Physion: Evaluating Physical Prediction from Vision in Humans and Machines. In *NeurIPS 2021 Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=CXyZrKPz4CU.

Black Forest Labs. black-forest-labs/FLUX.2-dev. https://huggingface.co/black-forest-labs/FLUX.2-dev, 2025. Accessed September 25, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/cvpr52729.2023.00276.

Kevin Clark and Priyank Jaini. Text-to-Image Diffusion Models are Zero Shot Classifiers. In *Advances in Neural Information Processing Systems 36*, pp. 58921–58937. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2571. URL http://dx.doi.org/10.52202/075280-2571.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1979. doi: 10.1214/aos/1176344552. URL http://dx.doi.org/10.1214/aos/1176344552.

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, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-T2I Challenge: Can Text-to-Image Generation Models Understand Commonsense? In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MI52iXSSNy.

Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pp. 25–30. ACM, 2013. doi: 10.1145/2509558.2509563. URL http://dx.doi.org/10.1145/2509558.2509563.

Arthur C. Graesser, Murray Singer, and Tom Trabasso. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371–395, 1994. doi: 10.1037/0033-295x.101.3.371. URL http://dx.doi.org/10.1037/0033-295X.101.3.371.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. *The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning*, pp. 558–575. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-20059-5_32. URL http://dx.doi.org/10.1007/978-3-031-20059-5_32.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing Hackable Benchmarks for Vision-Language Compositionality. In *Advances in Neural Information Processing Systems 36*, pp. 31096–31116. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1355. URL http://dx.doi.org/10.52202/075280-1355.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing Properties of Generative Classifiers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rmg0qMKYRQ.

Yujin Jeong, Arnas Uselis, Seong Joon Oh, and Anna Rohrbach. Diffusion Classifiers Understand Compositionality, but Conditions Apply. In *Advances in Neural Information Processing Systems 38*, pp. 137728–137777. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-4141. URL http://dx.doi.org/10.52202/085713-4141.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How Far Is Video Generation from World Model: A Physical Law Perspective. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DLlVjZQ7vD.

Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context Based Emotion Recognition using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019. doi: 10.1109/tpami.2019.2916866. URL http://dx.doi.org/10.1109/tpami.2019.2916866.

Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Chris Pal, and Siva Reddy. Are Diffusion Models Vision-And-Language Reasoners? In *Advances in Neural Information Processing Systems 36*, pp. 8385–8405. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-0367. URL http://dx.doi.org/10.52202/075280-0367.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your Diffusion Model is Secretly a Zero-Shot Classifier. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217. IEEE, 2023. doi: 10.1109/iccv51070.2023.00210. URL http://dx.doi.org/10.1109/iccv51070.2023.00210.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean De Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. In *Advances in Neural Information Processing Systems 37*, pp. 17044–17068. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-0542. URL http://dx.doi.org/10.52202/079017-0542.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social Genome: Grounded Social Reasoning Abilities of Multimodal Models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24868–24891. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1264. URL https://aclanthology.org/2025.emnlp-main.1264/.

Gail McKoon and Roger Ratcliff. Inference during reading. *Psychological Review*, 99(3):440–466, 1992. doi: 10.1037/0033-295x.99.3.440. URL http://dx.doi.org/10.1037/0033-295X.99.3.440.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2939. IEEE, 2016. doi: 10.1109/cvpr.2016.320. URL http://dx.doi.org/10.1109/CVPR.2016.320.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026. doi: 10.1109/wacv61042.2026.00099. URL http://dx.doi.org/10.1109/wacv61042.2026.00099.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World Knowledge-Informed Semantic Evaluation for Text-to-Image Generation. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=f1q8dFnxtY.

OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. Accessed September 25, 2026.

OpenAI. GPT-Image-2 Model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026. Accessed September 25, 2026.

Stefan Palan and Christian Schitter. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018. doi: 10.1016/j.jbef.2017.12.004. URL http://dx.doi.org/10.1016/j.jbef.2017.12.004.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. *VisualCOMET: Reasoning About the Dynamic Context of a Still Image*, pp. 508–524. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58558-7_30. URL http://dx.doi.org/10.1007/978-3-030-58558-7_30.

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 *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Qwen. Qwen's most powerful open-source image generation model · GitHub. https://github.com/QwenLM/Qwen-Image-2.1, 2026. Accessed September 25, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01042. URL http://dx.doi.org/10.1109/cvpr52688.2022.01042.

Stability AI. Announcing SDXL 1.0. https://stability.ai/news-updates/stable-diffusion-sdxl-1-announcement, 2023. Accessed September 25, 2026.

Stability AI. stabilityai/stable-diffusion-3.5-large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024a. Accessed September 25, 2026.

Stability AI. Stable Diffusion 3 Medium. https://stability.ai/news-updates/stable-diffusion-3-medium, 2024b. Accessed September 25, 2026.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, Aoxue Li, and Xihui Liu. T2I-ReasonBench: Benchmarking Reasoning-Informed Text-to-Image Generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8919–8944. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.433. URL http://dx.doi.org/10.18653/v1/2026.findings-acl.433.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022. doi: 10.1109/cvpr52688.2022.00517. URL http://dx.doi.org/10.1109/cvpr52688.2022.00517.

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, 2025. URL https://arxiv.org/abs/2502.14786.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant Similarity for Vision-Language Foundation Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11964–11974. IEEE, 2023. doi: 10.1109/iccv51070.2023.01102. URL http://dx.doi.org/10.1109/iccv51070.2023.01102.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021. doi: 10.1109/iccvw54120.2021.00217. URL http://dx.doi.org/10.1109/ICCVW54120.2021.00217.

Yue Yang, Yuqi Lin, Hong Liu, Wenqi Shao, Runjian Chen, Hailong Shang, Yu Wang, Yu Qiao, Kaipeng Zhang, and Ping Luo. Position: Towards Implicit Prompt For Text-To-Image Models. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=H9fNj8ivTy.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8799–8809. IEEE, 2019. doi: 10.1109/cvpr.2019.00901. URL http://dx.doi.org/10.1109/cvpr.2019.00901.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From Recognition to Cognition: Visual Commonsense Reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6713–6724. IEEE, 2019. doi: 10.1109/cvpr.2019.00688. URL http://dx.doi.org/10.1109/cvpr.2019.00688.

A Scoring details and robustness

An image generator scores a caption against an image by how accurately it predicts the noise added to that image when conditioned on the caption. We resize each image with bicubic resampling to the resolution in Table 1 and use the posterior mean of the generator's autoencoder as the clean latent. Noise is added at the 245 levels $t = 10, 14, \ldots, 986$ used in the released DiffusionITM code. At

each level, a single draw from a standard normal distribution serves as the noise for every image and caption. The error at a level is the ℓ_2 norm of the difference between the predicted and the added noise, and the error of a caption is its mean over the 245 levels.

Krojer et al. (2023) observe that this raw error favors a visually familiar image regardless of the caption. We adopt their correction and define the score of a caption on an image as the error with a blank caption minus the error with that caption. We write x_t for the latent of image x noised with the draw ϵ_t at level t and $\hat{\epsilon}_\theta(x_t, t, c)$ for the generator's noise prediction given caption c. With T denoting the set of 245 levels and $\emptyset$ denoting the blank caption, the error and the score defined above take the form

$$e(c, x) = \frac{1}{|T|} \sum_{t \in T} \left\| \epsilon_t - \hat{\epsilon}_\theta(x_t, t, c) \right\|_2 \quad \text{and} \quad s(c, x) = e(\emptyset, x) - e(c, x). \tag{1}$$

For the four flow transformers, x_t is the interpolated latent x_σ with $\sigma = t/1000$ defined below, and $\hat{\epsilon}_\theta$ is the noise prediction implied by the predicted velocity.

The generator then chooses the image or caption with the higher score. A contrastive model maps the image and the caption to one projected embedding each and scores the caption by the cosine similarity of the two embeddings. As with the generators, the contrastive model chooses the image or caption with the higher score. Table 4 lists the prediction target and the blank caption of each generator.

Table 4: The table lists the prediction target and the blank caption that each image generator uses in scoring. Velocity errors are converted to the noise scale before they are averaged.

Generator	Target	Blank caption
SD 1.5	Noise	Empty string through the CLIP ViT-L/14 text encoder
SD 2.1 Base	Noise	Empty string through the OpenCLIP ViT-H/14 text encoder
SDXL	Noise	All-zero caption and pooled embeddings, which the pipeline uses as its unconditional input
SD3 Medium	Velocity	Empty string through all three text encoders
SD3.5 Large	Velocity	Empty string through all three text encoders
FLUX.2 dev	Velocity	Empty string through the Mistral Small 3.2 text encoder, padded to 512 tokens
Qwen-Image 2.1	Velocity	A single space through the Qwen3-VL-8B text encoder, because the pipeline replaces an empty caption with a space

The four flow transformers in Section 4 interpolate linearly between the clean latent x_0 and the noise ϵ (Liu et al., 2023; Esser et al., 2024). With $\sigma = t/1000$, the noisy latent is $x_\sigma = (1-\sigma)\, x_0 + \sigma\, \epsilon$, and the generator predicts the velocity $v = \epsilon - x_0$. A velocity prediction $\hat{v}$ implies the noise prediction $x_\sigma + (1 - \sigma)\, \hat{v}$, whose error is $1 - \sigma$ times the velocity error. We therefore multiply the velocity error by $1 - \sigma$ so that all seven generators are scored by their error in predicting the noise. FLUX.2 dev additionally takes a guidance scale as input, which we fix at 4.0.

Claude Fable 5.1 answered each question once in each display order through the Anthropic Message Batches API, with images uploaded at their native size. We set its effort parameter to high and capped its output at 16,000 tokens. Adaptive thinking is always enabled for this model, and it produced thinking tokens in 275 of the 2,240 answers. The system prompt instructed the model to reply with exactly 1 or 2 to select the better-matching option. The image-choice and caption-choice questions were "Which photograph better matches the situation described by the sentence?" and "Which sentence better matches the situation shown in the photograph?" All 2,240 replies ended in 1 or 2, although 43 of them contained text before this final digit. Excluding these 43 replies leaves implicit image choice unchanged at 99.3% and lowers implicit caption choice from 94.6% to 94.3%.

We rescored every generator under alternative definitions of the error, and under each of these variants the explicit caption raises the image choice and swap score of every generator (Table 5). The squared variant averages the mean squared error over levels instead of the ℓ_2 norm, and the native variants leave the velocity error of the flow transformers unconverted. Across the 22 combinations of generator and variant, implicit image choice stays between 46.8% and 58.9% against a chance level of 50.0%. In every combination, the implicit image and text scores reach at most 21.4% against a chance level of 25.0%, and the implicit group score reaches at most 5.7% against a chance level

of 16.7%. We define the gain in image choice as explicit minus implicit image choice, and its 95% interval excludes zero in 19 of the 22 combinations. Without the blank correction, image choice falls to between 49.6% and 50.7% because the raw error then prefers the same image for both captions in 138 to 140 of the 140 pairs.

Table 5: The table reports image choice and the swap score of each image generator under up to four ways of computing the error. Headline is the scoring used in the main text. Squared averages the mean squared error over noise levels instead of the square root of the summed squared error. Native and native squared keep the velocity error of a flow transformer without converting it to the noise scale, so they apply only to the four flow transformers. Image choice is a percentage of 280 questions, and the swap score is a percentage of 140 pairs. The gain is explicit minus implicit image choice in points, with its 95% interval in brackets.

		Image choice			Swap score	
Generator	Variant	Implicit	Explicit	Gain	Implicit	Explicit
Chance		50.0	50.0		50.0	50.0
SD 1.5	Headline	55.0	63.9	8.9 [5.4, 12.9]	67.9	86.4
	Squared	58.9	65.0	6.1 [2.1, 10.0]	68.6	85.7
SD 2.1 Base	Headline	54.3	63.2	8.9 [5.4, 12.5]	67.1	89.3
	Squared	52.5	61.4	8.9 [5.4, 12.9]	66.4	88.6
SDXL	Headline	55.7	65.4	9.6 [5.7, 13.6]	74.3	86.4
	Squared	55.4	63.2	7.9 [3.9, 11.8]	65.0	84.3
SD3 Medium	Headline	50.4	55.7	5.4 [2.5, 8.6]	56.4	73.6
	Squared	50.4	52.1	1.8 [−1.8, 5.4]	53.6	62.1
	Native	54.6	58.6	3.9 [1.1, 7.1]	55.0	77.9
	Native squared	53.9	58.9	5.0 [1.8, 8.6]	59.3	77.9
SD3.5 Large	Headline	54.6	62.1	7.5 [3.6, 11.8]	62.1	85.7
	Squared	51.1	58.6	7.5 [2.9, 12.1]	54.3	76.4
	Native	54.6	65.0	10.4 [6.4, 14.6]	63.6	86.4
	Native squared	55.0	65.0	10.0 [6.1, 13.9]	63.6	85.0
FLUX.2 dev	Headline	51.8	56.1	4.3 [−0.7, 9.3]	49.3	60.7
	Squared	50.7	52.1	1.4 [−3.9, 7.1]	46.4	55.0
	Native	55.4	64.3	8.9 [3.9, 13.9]	59.3	72.9
	Native squared	56.4	63.9	7.5 [2.9, 12.1]	57.9	72.9
Qwen-Image 2.1	Headline	48.6	61.1	12.5 [7.9, 17.5]	53.6	77.9
	Squared	46.8	56.1	9.3 [5.0, 13.6]	45.7	60.7
	Native	51.4	68.2	16.8 [12.1, 21.8]	57.1	80.7
	Native squared	51.4	67.5	16.1 [11.4, 20.7]	57.1	78.6

Resampling the 245 noise levels with replacement 1,000 times flips 4.4% to 8.2% of a generator's decisions on average (Table 6). Decisions computed from the even levels and from the odd levels agree in 83.8% to 92.1% of cases. Across the resamples, the central 95% range of the gain in image choice stays above zero for every generator. This range reflects only variation across noise levels because the pairs are held fixed.

Table 6 also reports image choice without the 21 pairs whose images differ in size. In each of these pairs, one image measures 1254 pixels per side and the other 1280, so resizing rescales the two images by different factors. On the remaining 119 pairs, implicit image choice is 49.2% to 55.0% for the generators and 55.9% to 58.8% for the contrastive models. For every generator, the implicit image and text scores reach at most 13.4% against a chance level of 25.0%, and the implicit group score reaches at most 5.0% against a chance level of 16.7%. The explicit image score exceeds the implicit image score for every generator, and each generator's 95% interval for the gain in image choice lies above zero. These intervals and the resampled ranges above do not replace the sign test over all 140 pairs in Section 5, and that test does not detect the gain for FLUX.2 dev.

Apart from the pipeline check in Appendix B, we did not validate the scores of FLUX.2 dev and Qwen-Image 2.1, and the retrieval reproduction there covers only SD 2.1 Base. FLUX.2 dev is guidance-distilled, so its blank-corrected score is not a calibrated likelihood. Qwen-Image 2.1 was

Table 6: The table reports how a generator's decisions change when the noise levels are resampled and how image choice changes when the 21 pairs whose images differ in size are dropped. Flipped is the mean share of a pair's eight image-choice and caption-choice decisions that change across 1,000 resamples of the 245 noise levels. Halves is the share of decisions on which the even and odd levels agree. The first Gain column gives the central 95% range of the gain in image choice across these resamples. The last three columns use the 119 remaining pairs, and their gain comes with its 95% interval from resampling pairs.

	Resampled noise levels			Without the 21 pairs		
Generator	Flipped	Halves	Gain	Implicit	Explicit	Gain
SD 1.5	8.2	83.9	[3.9, 10.4]	54.6	63.9	9.2 [5.0, 13.4]
SD 2.1 Base	7.6	83.8	[5.0, 10.7]	52.5	62.2	9.7 [5.9, 13.4]
SDXL	7.0	87.0	[5.4, 11.4]	55.0	65.1	10.1 [5.9, 14.3]
SD3 Medium	5.2	89.8	[2.1, 7.1]	49.6	54.6	5.0 [2.1, 8.4]
SD3.5 Large	6.6	88.8	[5.0, 11.1]	55.0	61.8	6.7 [2.5, 10.9]
FLUX.2 dev	4.4	92.1	[0.4, 6.1]	50.0	56.3	6.3 [0.8, 11.8]
Qwen-Image 2.1	5.8	88.8	[7.9, 13.9]	49.2	61.3	12.2 [7.1, 17.6]

scored at 1024 pixels per side, below its native resolution of about 2K. The conversion to the noise scale also gives little weight to the highest noise levels, where the factor $1 - \sigma$ approaches zero. Relative to the headline scoring, the native variant removes this down-weighting and raises the implicit swap score from 49.3% to 59.3% for FLUX.2 dev and from 53.6% to 57.1% for Qwen-Image 2.1 (Table 5). It also raises their explicit image choice from 56.1% to 64.3% and from 61.1% to 68.2%, respectively. Implicit image choice for both generators stays between 51.4% and 56.4% under the native variants.

B VALIDITY CHECKS

Our scoring code reproduces the denoiser output of each generator's official pipeline in the diffusers library. Before scoring, we ran each official pipeline on a caption of the pair in Figure 1 and captured the input to its denoiser at one step. Our code then passed the same noisy latent and noise level to the denoiser along with caption embeddings computed independently of the pipeline. For all seven generators, the maximum absolute difference between the two outputs was 0.0, and our caption embeddings equaled those of the pipeline exactly. The blank caption embedding of SDXL, SD3 Medium, and SD3.5 Large also equaled the pipeline's own unconditional input. For FLUX.2 dev, we checked only that the blank caption embedding is finite. The blank captions of SD 1.5, SD 2.1 Base, and Qwen-Image 2.1 pass through the same text-encoder code as every other caption and were not compared separately. The checks of the six generators other than FLUX.2 dev ran on a code version that predates a one-line change to the FLUX.2 dev check, and this change does not alter their code.

We scored SD 1.5 and SD 2.1 Base from community copies, and the original SD 2.1 Base repository is no longer available. The SHA-256 hashes of the denoiser, text encoder, and autoencoder files we scored match those of four independent copies of SD 1.5 created in 2023. For SD 2.1 Base, the hashes match those of three independent copies, two of which were created in 2023.

Section 5 compares SD 1.5 and SD 2.1 Base with the contrastive models whose text-encoder weights they share, but the generators use the encoder output differently. The SD 1.5 text encoder has the same weights as the CLIP ViT-L/14 text model in all 197 tensors, and CLIP ViT-L/14 adds only a final projection. The SD 2.1 Base text encoder has the same weights as the first 23 of the 24 blocks of the OpenCLIP ViT-H/14 text model in all 372 tensors we compared. Both generators condition on the encoder's per-token hidden states through cross-attention, whereas the contrastive models use a single projected embedding of the full caption. The comparison is therefore not a causal test of the text encoder.

Our implementation closely reproduces the published image-retrieval accuracy of Diffusion-ITM (Krojer et al., 2023) on Flickr30K (Young et al., 2014). In this task, a model chooses among 10 candidate images for each caption. With a fixed seed, we drew 200 of the 1,014 queries used by

the DiffusionITM loader. SD 2.1 Base ranks the correct image first for 65.0% of these queries (95% CI 58.2 to 71.3, Wilson interval) against a chance level of 10.0%. The published accuracy of 61.5% lies inside this interval. Without the blank correction, SD 2.1 Base ranks the correct image first for 14.0% of the queries, against the published 11.1%.

C HUMAN STUDY

We recruited participants on Prolific (Palan & Schitter, 2018) in three separate studies, and they completed the task in a desktop browser through JATOS. Eligibility was restricted to fluent English speakers aged 18 to 70 with a Prolific approval rating of at least 95%. The first screen described the purpose, expected duration, and pay of the study. It also listed the data to be recorded and stated that participants had to be adults. All participants were adults and gave informed consent before the task began, and they could withdraw at any time by closing the study. The authors' institutional review board approved the protocol, and its identifiers are withheld for anonymous review. Pay did not depend on accuracy, and across the three studies it ranged from $2.00 to $3.00 for an advertised length of 8 to 12 minutes. Observed completion times put the effective pay at $13.11 to $19.30 per hour.

Each trial presented either one caption with the pair's two images or one image with the pair's two implicit or two explicit captions. Each study began with instructions and a worked example of each trial type in which the correct button was highlighted. The two trial types then ran as separate parts in counterbalanced order, and the screen position of the correct option varied across presentations. After each choice, participants rated their confidence on a four-point scale that ranged from guessing at 1 to complete certainty at 4, and response times were also recorded. No analysis in the paper uses the confidence ratings or the response times.

After excluding preview runs and three participants whose question sets were collected again, we retained 122 participants. Different participants answered the questions of a pair, and each question was answered by two to four participants and most often by three (964 of the 1,112 questions). In total, the 1,112 questions received 3,406 answers. For each question, the supplement reports only the number of answers and the number of correct answers. The three human studies together cover 139 of the 140 pairs because one pair was added to the benchmark after data collection. Model scores use all 140 pairs, and restricting them to the 139 rated pairs changes no value in Tables 2 and 3 by more than 1 point. Three pairs were shown to participants in an earlier version with a different image or slightly different caption wording. Their answers are included in the human scores, and excluding these three pairs changes each human score by less than 1 point.

An alternative to the majority vote of Section 4.3 treats the questions of a pair as answered independently and multiplies their fractions of correct answers. This alternative yields implicit and explicit human image scores of 86.2% and 94.4%. The corresponding text scores are 82.8% and 87.7%, and the group scores are 72.0% and 82.7%.