WHEN CHEKHOV'S GUN FIRES ANYWAY: OVER-REALIZATION IN TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In text-to-image generation, a target that the prompt mentions often appears in the generated image in a way the prompt does not call for. We call this failure *over-realization*. Mental-spaces theory points to the situations where it can happen: under a negation, inside what a character believes or imagines, as the vehicle of a figurative comparison, or behind a viewpoint that hides the target. Prior work has studied only negation in depth, and the conventional alignment and preference metrics do not penalize the failure. We build OverReal to measure it: OverReal-Gen, 698 prompts covering the four situations, on which we measure each generator's over-realization rate; and OverReal-Det, 2,841 generated images human-labeled for whether and how the target is over-realized, on which we verify the automatic judge that produces this rate. Our evaluation of eight text-to-image systems reveals that all of them over-realize, and that conventional metrics such as CLIPScore reward the failure rather than penalize it. Prompt expansion only partly mitigates the failure. Inside the generator, the diffusion backbone attends to a mentioned target almost as much as to a requested one. The cue words that should hold the target back receive less attention than ordinary words.



(a) *"Tom is reading a letter behind the desk. The letter is from his mom, which says: I miss you."*



(b) *"The coach is as fierce as a tiger."*

Figure 1: Two over-realizations, both generated by Gemini, with the prompt under each image. In (a) the model turns the page to face the camera: the words cannot reach the reader, so the scene contradicts itself. In (b) the model puts a tiger on the team's emblem and banner, which is not what the comparison meant, but the image still holds together.

1 INTRODUCTION

If in the first act you have hung a pistol on the wall, then in the following one it should be fired.

—Anton Chekhov

Chekhov's principle says that a planted gun is meant to fire, and to fire at the right moment. However, generative models have only learned to fire every gun, not to fire it properly. We study this failure in text-to-image generation, and Figure 1 shows two examples. In Figure 1(a) the prompt mentions a letter that says "I miss you": to render the words legibly, the model turns the page toward the camera instead of its reader, and the scene contradicts itself. In Figure 1(b) the prompt calls the coach as

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.

fierce as a tiger: the model renders the tiger onto the team's emblem and banner, which is not what the comparison meant. In both cases the model fired a gun it should have fired differently or left hanging. We call this failure *over-realization*: a mentioned target appears in the generated image in a way the prompt does not call for.

Mental-spaces theory (Fauconnier, 1994) suggests where over-realization can happen: some expressions in a prompt open a space alongside the scene to be drawn, and a target placed in that space is not part of the scene itself. Such a space can be opened by a *negation*, by what a character *believes or imagines*, by a *figurative comparison*, or by a *viewpoint* that hides the target. A model can then over-realize in two ways: the target disrupts the image, as in Figure 1(a), or settles silently into the scene in a way the prompt did not call for, as in Figure 1(b).

Of the four situations above, only negation has been studied in depth (Alhamoud et al., 2025; Park et al., 2025; Cai et al., 2025); the figurative and the perspectival case have each been studied within a narrow scope, as idioms (Sun et al., 2025) and as occlusion (Wang et al., 2026). Each is treated as a separate defect, and no prior work gives a systematic account of when a model may over-realize. Nor does current evaluation target the failure: it rests on two requirements, "a mentioned target should appear in the image" (Huang et al., 2023; Lin et al., 2024) and "an unmentioned target should not appear" (Lim et al., 2025), and neither covers the case where "a mentioned target does not necessarily appear", so an over-realized image is scored as correct.

We hypothesize that these situations share one failure mode: the model attends to the mentioned target almost as it would to a requested one, and renders it. To measure over-realization across the four situations and to examine this hypothesis, we build OverReal. **OverReal-Gen** is a prompt set covering the four situations, in which the target is prone to over-realization; a system's over-realization rate on it is our measurement. **OverReal-Det** is a set of generated images human-labeled for whether and how the target is over-realized; it verifies the automatic judge that produces this rate.

All eight systems we evaluate over-realize, on 41% to 73% of the images of OverReal-Gen. The proprietary systems have largely solved the templated negation prompts, but every system over-realizes at least 32% of the images in the other three situations, and the form of the failure follows the situation. The conventional alignment and preference metrics do not penalize the failure and in fact reward it. Prompt expansion helps mostly because the rewriter drops the target; whenever the generator sees the target word, it tends to put the target in the image.

Our contributions are (1) a taxonomy of over-realization, motivated by mental-spaces theory, covering four situations and two forms of the failure; (2) OverReal, a prompt set for measuring over-realization and a human-labeled image set for verifying automatic judges; and (3) an evaluation of eight text-to-image systems, with an analysis of how conventional metrics, prompt expansion and the diffusion backbone relate to over-realization. We release OverReal, the images and the labels.

2 Related Work

Text-to-image generation and its evaluation Current text-to-image systems consist of a text encoder and a diffusion transformer trained by flow matching (Esser et al., 2024; Black Forest Labs, 2024; Qwen Team, 2025; Wu et al., 2025), and the deployed proprietary systems (OpenAI, 2025b; Google DeepMind, 2025; Ideogram, 2025) expand the user's prompt before generation (Betker et al., 2023; Hao et al., 2023). Faithfulness to the prompt is the central criterion of text-to-image evaluation, and it has been understood as getting mentioned content into the image: T2I-CompBench (Huang et al., 2023), TIFA (Hu et al., 2023), Davidsonian Scene Graph (Cho et al., 2024) and VQAScore (Lin et al., 2024) decompose the prompt into elements or questions and reward an image for each one it contains, and methods that improve faithfulness pursue the same goal on the generation side (Chefer et al., 2023). Every mention is an obligation in these protocols, so an image that shows something the prompt mentions but does not call for scores as correct, or better. Hallucination benchmarks (Qin et al., 2024; Lim et al., 2025; Kasaei & Rohban, 2025) measure content the prompt never mentions. However, the opposite failure, content that is mentioned and should stay out, has not been well-studied.

Negation and semantic leakage Among the cases in which a mentioned element should stay out of the image, negation is the one that has been studied most. Vision-language models largely ignore it

(Alhamoud et al., 2025; Conwell et al., 2024), and negation-aware training data improves CLIP (Park et al., 2025; Cai et al., 2025); we build our existence-canceling items on such prompts. Semantic leakage has been studied in generation within a narrower scope: a property of one requested entity appears on another (Ventura et al., 2026). Reasoning benchmarks touch two further cases in passing, idioms rendered literally (Sun et al., 2025) and objects that should be occluded (Wang et al., 2026). No prior work gives a systematic account and analysis of where and how over-realization can happen.

3 PRELIMINARIES

A generative model *over-realizes* when a *target* word or concept that a *prompt* mentions appears in the generated *image* in a way the prompt does not call for. Other ways an image can fail to match its prompt, unrelated to the target, fall outside the definition of over-realization. We now describe where over-realization can happen, and how a model over-realizes in these situations; Appendix A.1 gives an example of each combination in Table 2.

3.1 WHERE OVER-REALIZATION CAN HAPPEN

To locate where over-realization occurs, we use mental-spaces theory from cognitive linguistics (Fauconnier, 1994). In its terms, a prompt describes not one but several spaces: the scene to be generated (the *base space*), and alongside it the user's own assumptions, beliefs, memories and comparisons (the *non-base spaces*). A text-to-image model should render only the base space, but it sometimes fails to do so. **Over-realization is content from a non-base space leaking into the base space in a way the prompt does not call for.** We group these situations into the following families:

Family 1: existence-canceling The target is mentioned under an operator that cancels its existence in the described scene. Typical explicit cues include *no*, *without* and *never*. It is well studied (Alhamoud et al., 2025; Park et al., 2025; Cai et al., 2025) and included for completeness; the target should not appear in the image.

Family 2: mental-state The target is mentioned inside the mental state of a person in the scene, so that it exists only in what that person believes, remembers or imagines. Typical explicit cues include *believes*, *thinks*, *remembers* and *dreams*. Unlike family 1, the target may legitimately be depicted, for instance inside a thought bubble or a hazy vignette, but it must not be shown as though it were physically present in the scene.

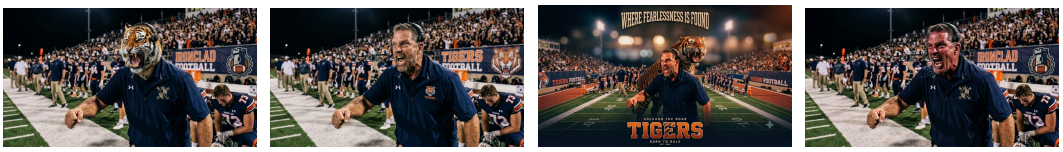
Family 3: figurative The target is mentioned as the vehicle of a figurative comparison: it characterizes something else in the scene rather than belonging to the scene itself. Typical cues include *like*, *as* ... *as*, *resembles* and the suffix *-like*. The target should not appear in a photorealistic image, but where the image itself is stylized it may legitimately appear, for instance as a faded shape in the background of a poster.

Family 4: perspectival The target is mentioned as something that may not be visible from the viewpoint of the image: it lies behind a barrier, or on a surface turned away from the camera. Typical explicit cues include *hidden*, *inside*, *closed* and *facing away*. The target may be brought into view by a camera angle that does not contradict the scene, introduced through a quote box, or shown only in part or not at all.

3.2 HOW MODELS OVER-REALIZE

In these situations a model may over-realize the target or handle it correctly. We divide over-realization into two kinds according to whether the picture still holds together as a rendering of the requested scene:

Disruptive over-realization The target appears, and its appearance stops the picture from making sense: the scene contradicts itself or the prompt, or no longer reads as the requested scene from an ordinary view. In Figure 2(a) the coach is given the tiger's head: the vehicle of the comparison has displaced the person it described.



(a) Disruptive over-realization ✗ (b) Silent over-realization ✗ (c) Integrated realization ✓ (d) Withheld realization ✓

Figure 2: Four possible outcomes for the prompt *"The coach is **as fierce as a tiger**."* The target word *tiger* is mentioned only as the vehicle of a figurative comparison, so it is not strictly requested in the scene. (b) is the outcome the model actually produced. (a) and (b) are over-realization; (c) and (d) are correct.

Silent over-realization The target slips in while the picture holds together, typically in the background or in a detail. In Figure 2(b) the tiger appears on the team's emblem and banner. Nothing in the picture is wrong on its own, but the emblem shows a tiger because the prompt mentioned one, not because the scene called for one. This is the warning level: the model has put the target in the scene in a way the prompt did not call for, and the scene happened to offer a harmless place for it. None of the evaluation methods we study detects this defect.

A generation is correct when it does not over-realize the target and the picture matches the scene the prompt describes. We divide correct generations by whether the target appears.

Integrated realization The target appears, depicted coherently and consistently with the prompt. Figure 2(c) is a poster: the tiger looms behind the coach in a composition no one reads as a photograph, so it is taken as an attribute of the man. This is correct behavior wherever the target may legitimately be shown. However, in most of our cases the target must not appear in the scene as a physical entity (§3.1), which makes integrated realization hard to produce.

Withheld realization The target is left out and the picture is coherent without it. A mention does not oblige the model to depict, so this is as correct as an integrated depiction. In Figure 2(d) no tiger appears and nothing is lost: the coach's face and stance carry the fierceness. Prior work treats a missing mention as a defect (Chefer et al., 2023; Hu et al., 2023; Huang et al., 2023); here it is correct behavior wherever the target should not be shown.

Note that our definition covers only targets that appear because they were mentioned. A target can also belong to the scene, as a plane belongs to an airport, or come in by coincidence, as the tiger in Figure 2(b) would if the team were named after tigers. Such appearances are not over-realization; §5.3 measures how often the target appears when it is not mentioned, to discount them.

4 OverReal Benchmark

4.1 OverReal-Gen: the generation benchmark

Data format Every entry of OverReal-Gen pairs a *prompt* with a *target*: the prompt describes a scene in which the target is prone to over-realization. Following Section 3.1, we classify these situations into four main prompt families: existence-canceling, mental-state, figurative, and perspectival. Table 2 in Appendix A.1 provides several examples.

Prompt collection Existence-canceling is well studied and its construction is direct, so we synthesize these prompts from templates following prior negation benchmarks (Alhamoud et al., 2025; Conwell et al., 2024). The mental-state, figurative, and perspectival prompts are all written by hand. Annotators write in natural language about everyday scenes, so that the prompts resemble those of a user; each annotator writes a prompt and names its target, Gemini generates three images, and the pair is kept if the annotator judges at least one image to over-realize the target and a second review confirms it. In total, 698 pairs enter OverReal-Gen (Table 3, Appendix B). OverReal-Gen is therefore a challenge set: the three hand-written families (not the templated existence-canceling one) are enriched for prompts on which a generator fails, so its rates measure how the systems handle

prompts that are prone to over-realization. As a comparison free of this selection, §6.1 also evaluates an in-the-wild set drawn from DiffusionDB (Wang et al., 2023) and Pick-a-Pic (Kirstain et al., 2023) by wording alone.

Prompt expansion Expanding the prompt with an LLM before generation is now a common part of the text-to-image pipeline. To control this stage, we collect prompts expanded from OverReal-Gen with the Qwen-Image rewriter (Qwen Team, 2025) and Ideogram Magic Prompt (Ideogram, 2024). These 1,396 derived prompts form the expanded split (Table 3), which we use to analyze the effect of expansion on over-realization (§6.2).

4.2 OverReal-Det: the detection benchmark

Data format Every entry of OverReal-Det is a triple of a prompt, its target, and one generated image, and carries two independent annotations, each a set of labels from {disruptive, silent, integrated, withheld, other}, as defined in §3.2.

Image collection The images come from two sources: (1) the pilot images produced during prompt collection (§4.1), and (2) images generated from the OverReal-Gen prompts with the models listed in §5.

Annotation and splits Each image is labeled independently by two annotators following the guide in Appendix E.2, who use two labels only for borderline images. An image whose two label sets intersect enters the *gold* split with the intersection as its label (a third annotator decides the three images left with two labels); an image whose sets are disjoint enters the *ambiguous* split, released with both label sets; images that both annotators mark *other* are discarded. This yields 2,253 gold and 588 ambiguous images, with an overall κ agreement of 0.68 (Table 3). Appendix B.2 analyzes the ambiguous split.

5 Experimental Setup

We run eight generators, three proprietary and five open-weight (§5.1), on every prompt of OverReal-Gen (§4.1), and evaluate the generated images both with conventional metrics (§5.2) and from the perspective of over-realization (§5.3).

5.1 Text-to-Image Generators

We evaluate eight text-to-image generators. The proprietary generators are GPT-Image-1.5 (OpenAI, 2025b), Ideogram v3 (Ideogram, 2025), and Nano Banana (gemini-2.5-flash-image) (Google DeepMind, 2025), while the open-weight models are the diffusion models FLUX.1-dev (Black Forest Labs, 2024), Qwen-Image (Qwen Team, 2025), SD3.5-Medium and SD3.5-Large (Esser et al., 2024), and the unified understanding-and-generation model OmniGen2 (Wu et al., 2025). An open-weight model receives the exact prompt; a proprietary generator rewrites it internally before generation, so the text that conditions its image is not under our control. Ideogram returns its rewritten prompt, which we record and reuse as a rewriter in §6.2; GPT-Image and Nano Banana do not expose theirs. All images are generated at 1024×1024, the default resolution of most generators, and every open-weight model produces two images per prompt under fixed seeds.

5.2 Conventional evaluation

For each generator, family and prompt condition we score a fixed random sample of up to 100 images (7,050 in total) with five conventional text-to-image metrics. For text–image alignment we use CLIPScore (Hessel et al., 2021) and the VQA-based VQAScore (Lin et al., 2024); for learned human preference, PickScore (Kirstain et al., 2023) and HPSv2 (Wu et al., 2023); and as an LLM judge, a VIEScore-style semantic-consistency rating (Ku et al., 2024) produced by GPT-5.2 (OpenAI, 2025a) and Claude Opus 5 (Anthropic, 2026); we report the Opus rating in the main text, since Opus shares no vendor with the evaluated generators, and the GPT-5.2 rating in Appendix D.1 for reference. The scores of every generator are given in Table 6.

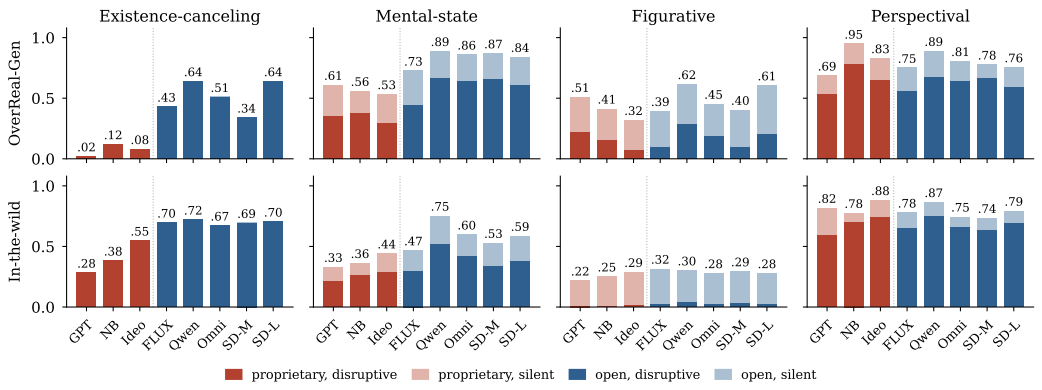


Figure 3: Main results: over-realization rate of each generator by prompt family, on the raw prompts of OverReal-Gen (top) and on the in-the-wild prompts (bottom). Bar height is the over-realization rate; the dark part is disruptive (DO) and the light part silent (SO) over-realization; red bars are the proprietary generators, blue bars the open models. The full outcome shares, including integrated (I) and withheld (W) realization and the unprompted appearance rate u, are in Table 7 (Appendix D).

5.3 Over-realization evaluation

Automatic labeling Evaluating 698 prompts on eight generators, under up to three prompt conditions and two seeds, yields tens of thousands of images, so exhaustive manual annotation is infeasible. We therefore label the images with a third-party model, Claude Opus 5 (Anthropic, 2026). The model substantially agrees with human annotators ($\kappa = 0.70$) on whether a generator over-realizes or not over the test split of OverReal-Det. Details are given in Appendix C. All rates below are over the images that received one of the four labels; the 15 images (0.5%) labeled *other* are excluded.

Unprompted appearance rate A target can appear in a scene for reasons unrelated to its mention. To discount such coincidences from the over-realization rates, we also report the unprompted appearance rate u of the target: we replace the target in the prompt with a blank ("___"), generate images from the edited prompt with the same settings, and count how often the target still appears. Ideally u is close to zero.

6 Results and Analysis

6.1 Main results

Figure 3 shows the over-realization rate of three proprietary and five open generators (§5.1) on the raw prompts of OverReal-Gen and on a set of in-the-wild prompts; the full outcome shares are in Table 7 (Appendix D). Conclusions are given as follows.

All generators over-realize, and the rate depends on the prompt family. We first read the over-realization rate, the sum of the DO and SO shares; its prompt-level bootstrap 95% interval is within ± 0.13 in every cell (Appendix D.2). Over all prompts, weighting the four families by their prompt counts, each generator over-realizes 41% to 73% of the images. The proprietary generators rarely over-realize the templated existence-canceling prompts (at most 12% of the images), while the open models still do so on 34% to 64%. Every model fails on mental-state and perspectival prompts, over-realizing about half or more of the images; on perspectival prompts the proprietary generators are no better than the open ones, and Nano Banana over-realizes 95% of the images. Figurative prompts sit in between, at 32% to 62%.

The form of the failure depends on the family. The two forms of over-realization are not distributed alike. On perspectival prompts the failures are mostly disruptive: the hidden object or content is exposed, contradictory with the prompt or common sense. On figurative prompts, by

	Existence-canceling					Mental-state					Figurative					Perspectival				
	CS↑	VQ↑	PS↑	HP↑	LJ↑	CS↑	VQ↑	PS↑	HP↑	LJ↑	CS↑	VQ↑	PS↑	HP↑	LJ↑	CS↑	VQ↑	PS↑	HP↑	LJ↑
DO	**.72**	.32	**22.9**	**.30**	3.0	.73	.71	**21.8**	**.32**	6.4	**.69**	.74	**22.0**	**.30**	4.8	.74	.74	**22.5**	**.30**	3.6
SO	—	—	—	—	—	.73	.69	**21.8**	.30	6.7	.63	.74	21.5	.29	6.0	.74	.72	22.3	.29	3.5
I	—	—	—	—	—	**.74**	**.81**	21.7	.31	**8.3**	**.69**	**.86**	21.9	.29	**6.9**	**.76**	**.85**	22.2	**.30**	4.3
W	.66	**.85**	**22.9**	**.30**	**8.8**	.67	.59	21.3	.29	6.6	.60	.75	21.2	.27	6.8	.67	.59	21.9	.28	**4.5**

Table 1: Mean score of the conventional metrics over the images with each label, pooled over generators; cells with fewer than 10 images are left blank.

contrast, silent over-realization accounts for 25% to 40% of the images and exceeds the disruptive form for all eight generators: the vehicle of the comparison is drawn into the scene as a harmless detail.

Unprompted appearance does not explain the rates. The unprompted appearance rate u (Table 7) stays at or below 9% in every cell, and below 5% for the open models, far below the over-realization rates. The rates of Figure 3 are therefore not explained by the target belonging to the scene, and coincidental appearances inflate them only marginally.

Silent over-realization is a precursor of disruptive over-realization, except in the figurative family. Silent over-realization does not disrupt the image, but it is a warning sign of disruptive over-realization rather than a lucky form of correct realization: an image labeled silent is over-realized again on the second seed 90% of the time (disruptive 33%, silent 57%) and correct only 10% of the time (Appendix D.3). The pattern holds for every open model and is strongest on perspectival and mental-state prompts; on figurative prompts a silent image is followed by another silent one 76% of the time and rarely turns disruptive, since the scene can usually hold the vehicle of the comparison.

Non-LLM conventional metrics reward over-realization and penalize withholding. Table 1 reports the mean score of each metric over the images with each label. A metric sensitive to over-realization would score the two correct outcomes (I, W) above the two over-realizations (DO, SO). The alignment (CS, VQ) and preference (PS, HP) metrics credit integrated images but give withheld images the lowest score in most cases: they reward the presence of the target, not its correct handling. Only the LLM judge scores withheld images above over-realized ones, in three of the four families. Within matched pairs of images from the same prompt and generator whose two seeds differ in outcome, CLIPScore, PickScore and HPSv2 score the over-realized image higher in 64% to 66% of the pairs (Appendix D.4): the conventional metrics reward over-realization even with prompt and generator held fixed.

The conclusions hold on in-the-wild prompts. For comparison with the curated OverReal-Gen (§4.1), we take 397 prompts of the four families from DiffusionDB (Wang et al., 2023) and Pick-a-Pic (Kirstain et al., 2023), real user prompts selected by their wording alone (Appendix D.6). The bottom row of Figure 3 shows that every generator still over-realizes in every family: pooled over generators, the rate is 60%, 51%, 28% and 80% on existence-canceling, mental-state, figurative and perspectival prompts, with the same family patterns as on the curated prompts, which run 0.1 to 0.3 higher on mental-state and figurative prompts and lower on existence-canceling ones.

6.2 The effect of prompt expansion

The proprietary generators expand the prompt to include more details before generating, so their advantage could come from either the prompt expansion or the generator. This section asks **whether prompt expansion reduces over-realization**. We compare five open models, with their own expansion disabled, on the raw prompts and on the prompts rewritten by the Qwen-Image official rewriter and by Ideogram Magic Prompt (Table 13, Appendix D.5).

Expansion lowers the rate but does not remove the failure. On existence-canceling, mental-state and figurative prompts, a rewrite lowers the over-realization rate in 29 of the 30 generator–rewriter pairs; the largest gains are on existence-canceling prompts, where the Ideogram rewrite brings the rate to 11% or below. The Ideogram rewrite is the more reliable of the two: it improves every cell of the

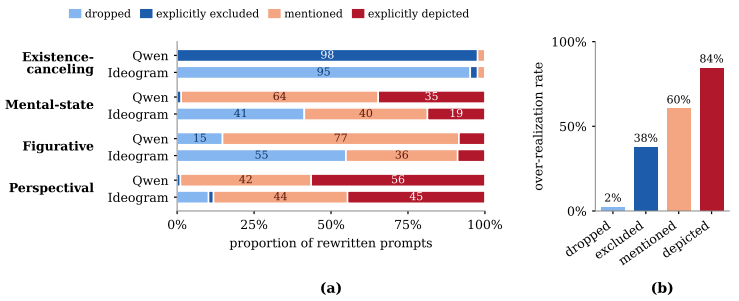


Figure 4: What the rewriters do with the target. (a) Proportion of rewritten prompts in which the target is dropped, explicitly excluded, mentioned, or explicitly to be depicted. (b) Over-realization rate of the images generated from those prompts, by the four categories, pooled over rewriters and open generators. Full numbers in Tables 14 and 15.

three families and roughly halves the rates in the mental-state and figurative families, eliminating 58% of the over-realizations against 21% for the Qwen rewrite, which on existence-canceling prompts even raises the rate of its own generator, Qwen-Image, from 64% to 81%. Perspectival prompts are the exception: the Qwen rewrite raises the rate for all five generators, to 84% to 91%, and the Ideogram rewrite leaves it within a few points of the raw prompt. And even with the Ideogram rewrite, the open generators still over-realize about 48% of the mental-state prompts, 21% of the figurative ones and 79% of the perspectival ones.

The gain comes from dropping the target. We classified every rewritten prompt by its treatment of the target: *dropped* (no longer mentioned), *explicitly excluded* (stated to be absent), *mentioned* (named, but nothing calls for it to be shown), or *explicitly depicted* (stated to be shown); Figure 4 summarizes the distribution and Appendix D.7 gives it in full. When a target is not meant to appear, Ideogram drops it whereas Qwen explicitly excludes it, and Figure 4(b) shows that only dropping substantially suppresses over-realization; Ideogram also recognizes more often that the target should not appear, dropping it in 41% of the mental-state and 55% of the figurative prompts, against 0% and 15% for Qwen. Neither rewriter reasons correctly about perspectival prompts: both keep the target in nearly every rewrite and describe it as visible in about half. The three treatments other than dropping still yield over-realization in 38% to 84% of the images, so the generator over-realizes on its own, and Section 6.3 turns to its side.

6.3 Attention to the target and the cue words

An open text-to-image model has two stages, a text encoder and a diffusion backbone that reads the encoder's output while denoising the image. Here we read the attention inside the backbones of FLUX.1-dev, SD3.5-Large and Qwen-Image under the setting of §5; the encoder is examined in Appendix D.8. To compare a target the prompt does not call for with a requested one, we write a *plain-mention control* for each prompt by hand, replacing the cue words of each family (*no*, *imagines*, *like*, *hidden*) by words that put the target into the scene, and generate it under the same seed: *"His vest is as green as a cucumber"* becomes *"His vest is green, and he holds a cucumber"* (more in Table 18). Table 19 in Appendix D.9 reports the share of attention on the target words, the cue words and the other words, per family and averaged. Three observations follow.

A mentioned target receives almost as much attention as a requested one. Under the original prompt, one token of the target words receives 2.8% to 4.4% of the image-to-text attention (a_t), while under the control the same words receive 3.0% to 5.0% (a'_t). The ratio ρ_t, 0.83 to 0.90, means that the target receives almost the same attention under the original prompt as under the control, although the control requests the target and the original only mentions it. Therefore, the backbone's attention barely distinguishes a mentioned target from a requested one. Figure 10 (top row, first two columns) shows the map for the cucumber of Table 2: under both prompts the attention sits where the cucumber is drawn, on the man's head under the original prompt and in his hand under the control.

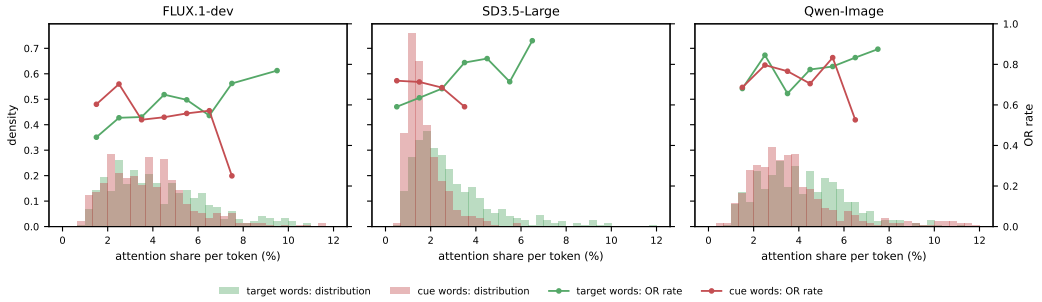


Figure 5: Histograms: the distribution over the attention of the target words (green) and of the cue words (red). Lines: the over-realization rate (OR rate) within each interval of the share; intervals with fewer than 12 images are not drawn.

The cue words receive less attention than ordinary words. The cue words (*no*, *imagines*, *like*, *hidden*) receive 1.7% to 3.8% per token (a_c), less than an ordinary word (γ = 0.60 to 0.76), and also less than the replacing words (*with*, *holding*, *next to*, *open*; ρ_c = 0.84 to 0.97). The backbone's attention, then, does not reflect what the cue words say: that the target need not appear in the image, and if it does, only in a form the prompt allows, such as a thought bubble, a stylized shape, a reflection or a viewpoint that brings it into sight.

Over-realization rate increases with more attention on the target words and less attention on the cue words. Figure 5 relates the attention share of target words and cue words to the over-realization. From the lowest to the highest interval of the figure, the over-realization rate rises with the share of the target words ($0.44 \rightarrow 0.77$, $0.59 \rightarrow 0.92$ and $0.68 \rightarrow 0.88$ for FLUX.1-dev, SD3.5-Large and Qwen-Image) and falls with the share of the cue words ($0.60 \rightarrow 0.25$, $0.72 \rightarrow 0.59$ and $0.69 \rightarrow 0.53$), the fall being steepest where the cue words receive an unusually large share. The over-realization rate thus tracks both attention shares, rising with the share of the target words and falling with the share of the cue words on all three backbones.

7 Conclusion

We introduced over-realization, the appearance of a mentioned target in a generated image in a way the prompt does not call for, and drew on mental-spaces theory to organize the four situations where it can happen. OverReal measures it with a prompt set prone to over-realization and a human-labeled image set. All eight text-to-image systems over-realize outside negation, on the benchmark and on real user prompts; the form of the failure follows the family, and conventional metrics reward over-realization. Prompt expansion helps mainly when the rewriter drops the target; when the target word reaches the generator, its backbone attends to it almost as to a requested one and to the cue words less than to ordinary words. Faithful generation therefore requires a model that can decide on its own which mentioned target to leave out (Appendix A.3), and evaluation that credits it for withholding.

Limitations

Modality. Over-realization is not specific to text-to-image generation: a language model can realize a mentioned target in text, and a video generator in its own medium. This paper covers images, where the failure produces a visible contradiction and generation and evaluation are affordable; extending the taxonomy and the benchmark to other modalities is left for future work.

Subjectivity of the outcome labels. The four outcome labels are defined by the effect that the target's appearance has on the image, and this effect is a matter of degree, in particular between disruptive, silent and integrated. The annotators did not always agree; Appendix A.2 gives borderline examples and Appendix B.2 counts the disagreements.

Ethics Statement

The prompts are written by the authors or synthesized from templates, and the images are generated by the evaluated models; no personal data is involved. The work studies a faithfulness failure of text-to-image generation and provides no means of attack.

Reproducibility Statement

OverReal-Gen and OverReal-Det are released with the annotations, the plain-mention controls and the expanded prompts. All generator settings are provided in §5.1. The judges and the automatic annotator are identified by model version in §5.2 and §5.3, and the exact question text of the annotation protocol and the human annotation guide are given in Appendix E. All code for generation, annotation, probing and attention recording is released with the data.

AI Use Statement

Large language models are part of the evaluated pipeline and of the evaluation itself: Claude Opus 5 and GPT-5.2 serve as the LLM judge (§5.2) and Claude Opus 5 as the automatic annotator (§5.3); the Qwen-Image rewriter and Ideogram Magic Prompt produce the expanded prompts (§6.2). The authors used Claude Code as a coding and writing assistant; all analyses, claims and text were checked by the authors, who take full responsibility for the content.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Anthropic. Claude Opus 5. `https://docs.anthropic.com/en/docs/about-claude/models`, 2026. Model version `claude-opus-5`, accessed September 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023.

Black Forest Labs. Flux.1. `https://github.com/black-forest-labs/flux`, 2024.

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

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*, 42(4), 2023. doi: 10.1145/3592116.

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 *International Conference on Learning Representations (ICLR)*, 2024.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Gilles Fauconnier. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press, 1994.

Google DeepMind. Gemini 2.5 flash image. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

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

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (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 *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Ideogram. Magic prompt. https://docs.ideogram.ai/using-ideogram/generation-settings/magic-prompt, 2024.

Ideogram. Ideogram 3.0. https://about.ideogram.ai/3.0, 2025.

Seyed Amir Kasaei and Mohammad Hossein Rohban. Hallucination as an upper bound: A new perspective on text-to-image evaluation. *arXiv preprint arXiv:2509.21257*, 2025. NeurIPS 2025 Workshop on Generative and Protective AI for Content Creation.

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

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of ACL*, 2024.

Youngsun Lim, Hojun Choi, and Hyunjung Shim. Evaluating image hallucination in text-to-image generation with question-answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

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 *European Conference on Computer Vision*, 2024.

OpenAI. GPT-5.2. https://platform.openai.com/docs/models/gpt-5.2, 2025a. Model version gpt-5.2-2025-12-11.

OpenAI. GPT image in the API. https://platform.openai.com/docs/models/gpt-image-1.5, 2025b.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2825–2835, 2025.

Ziyuan Qin, Dongjie Cheng, Haoyu Wang, Huahui Yi, Yuting Shao, Zhiyuan Fan, Kang Li, and Qicheng Lao. Evaluating hallucination in text-to-image diffusion models with scene-graph based question-answering agent. *arXiv preprint arXiv:2412.05722*, 2024.

Qwen Team. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2I-ReasonBench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025.

Mor Ventura, Michael Toker, Or Patashnik, Yonatan Belinkov, and Roi Reichart. DeLeaker: Dynamic inference-time reweighting for semantic leakage mitigation in text-to-image models. In *International Conference on Learning Representations (ICLR)*, 2026.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. In *International Conference on Learning Representations (ICLR)*, 2026.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, et al. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 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.

A Example of Generated Images

A.1 Taxonomy

Table 2 shows more examples of over-realization with different labels in each category. All images are produced by the generators of §5.1 with prompts from OverReal-Gen and labels annotated by two human experts. Existence-canceling prompts have no silent or integrated example, since a target whose existence is cancelled should not appear in any status.

A.2 Borderline cases

The outcome labels are defined by the effect that the target's appearance has on the image, and for some images this effect is a matter of degree. Figure 6 shows three pilot images on which the two annotators disagreed, one for each pair of labels that is hard to separate.

Disruptive or silent. The two labels differ on whether the target's appearance fits the scene. In Figure 6(a), the sponge, the vehicle of the comparison, is drawn into the office: one annotator read it as an object an office may plausibly hold, the other as something that does not belong there.

Disruptive or integrated. The two labels differ on whether an exaggerated rendering is read as a stylized expression or as a scene that breaks common sense. In Figure 6(b) a bowling ball rolls through the department: one annotator took it as a stylization of the comparison, the other as an object that a real office cannot contain.

Integrated or silent. The two labels differ on whether a depiction of the target inside the scene is a form the prompt allows or the target slipping in. In Figure 6(c) the cat appears as a framed photograph on the counter: one annotator read the photograph as a fitting way to show an absent cat, the other as a cat that was not supposed to appear at all.

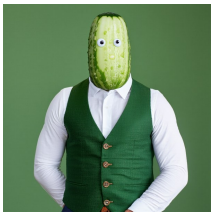
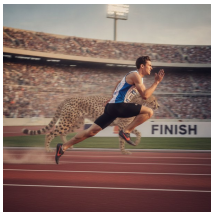
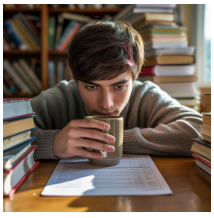
Family	Disruptive over-realization	Silent over-realization	Integrated realization	Withheld realization
1. Existence-canceling	 prompt: a bath room with sinks and mirrors, no horse anywhere in the scene target: horse model: OmniGen2	—	—	 prompt: boston Terrier sitting on a couch with a shoe, and not a single sandwich target: sandwich model: GPT-Image-1.5
2. Mental-state	 prompt: Maya at the ridge, believing that there is a fierce tiger at the ridge and being super afraid. target: tiger model: SD3.5-Large	 prompt: Sitting in math class, the thirsty girl stares at the whiteboard, wondering if she can buy a bottle of apple juice as soon as class ends. target: apple juice model: SD3.5-Large	 prompt: Sitting on the sofa one warm morning, the kid thought that there were bears outside at night, although he had never seen them. target: bears model: GPT-Image-1.5	 prompt: A woman waits at a train station, imagining her deceased friend arriving on the next train. target: friend model: OmniGen2
3. Figurative	 prompt: His vest is as green as a cucumber. target: cucumber model: SD3.5-Large	 prompt: The unexplained expense was a red flag during an otherwise routine audit meeting. target: red flag model: Nano Banana	 prompt: The runner is as fast as a cheetah as he races toward the finish line. target: cheetah model: Nano Banana	 prompt: His wrinkles were as layered as tree rings. target: tree rings model: GPT-Image-1.5
4. Perspectival	 prompt: A man holds a whiteboard against his chest with “Good luck!” written on the hidden side target: Good luck model: SD3.5-Large	 prompt: A closed batman cookie jar contains a gingerbread men target: gingerbread men model: SD3.5-Large	 prompt: Tom is reading a letter. The letter is from his mom, which says: I miss you. target: I miss you model: GPT-Image-1.5	 prompt: The student covered the grade on his exam sheet with a mug, pretending it wasn’t a bad score. target: score model: Ideogram v3

Table 2: Taxonomy examples.



prompt: *The new employee absorbed every instruction like a sponge soaking up water.*
target: **sponge**
labels: disruptive / silent



prompt: *The tiny disagreement rolled through the department like a bowling ball knocking down everything in its path.*
target: **bowling ball**
labels: disruptive / integrated



prompt: *The woman is making her breakfast, wondering what her cat is doing, since it is living with her parents right now.*
target: **cat**
labels: silent / integrated

Figure 6: Borderline cases: (a) disruptive or silent, (b) disruptive or integrated, (c) integrated or silent. The two labels under each image are the two annotators' labels.

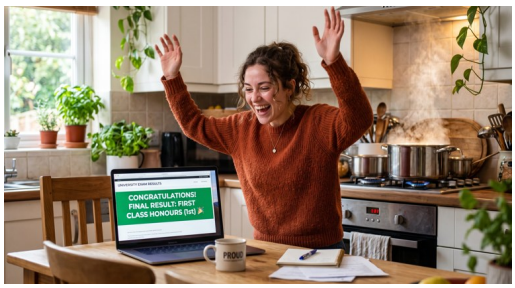
A.3 Unexpected over-realized targets

An annotator writes a prompt with one target in mind, but the generator may over-realize another mention, and even another family. Figure 7 shows two example images which over-realize both the vehicle of the comparison and the viewpoint. In (a) the leaves drift past the window (silent), and the student, who should face the lecturer, is turned toward the camera so that his face is visible (disruptive). In (b) a pot bubbles on the stove (silent), and the screen that carries the results faces the camera rather than the woman who is reading it (disruptive).

In practical use, a generator is not told which mention to withhold, and a single prompt can over-realize several of them. Perspectival over-realization in particular can arise in any prompt that places a person or a text-bearing object in a scene, without a cue word. Methods that suppress a given word, such as a negative prompt or a reweighting of the attention on that word, must first be told which word to suppress, and a user does not say.



prompt: *The student's attention drifted like a leaf in the wind during the long afternoon lecture.*
targets: **leaf** (figurative), **the student** (perspectival)
labels: silent / disruptive



prompt: *Her excitement bubbled over like a pot on the stove when the results came in.*
targets: **pot** (figurative), **the results** (perspectival)
labels: silent / disruptive

Figure 7: Two pilot images (Gemini) that over-realize two targets each: the vehicle of the comparison, silently, and the viewpoint of the person, disruptively. (a) The student faces the camera instead of the lecturer. (b) The screen faces the camera instead of the woman reading it.

B Dataset Details

B.1 Statistics

Table 3 gives the size of the raw and expanded splits of OverReal-Gen, the size of the gold and ambiguous splits of OverReal-Det, the label distribution of the gold split, and the agreement between the two annotators.

Family	OverReal-Gen				OverReal-Det						
	Raw		Expanded		Gold size	Label distribution (%)				Amb. size	Overall agreement (κ)
	Size	Avg. tokens	Size	Avg. tokens		DO	SO	I	W		
Existence-canceling	200	16.9	400	112.1	204	32.4	0.5	0.0	67.2	23	0.73
Mental-state	75	23.5	150	124.0	239	53.6	18.0	11.3	17.2	112	0.51
Figurative	237	16.4	474	116.4	932	36.6	28.1	2.9	32.4	299	0.65
Perspectival	186	23.1	372	118.7	878	70.5	10.3	7.3	12.0	154	0.72
Overall	698	19.1	1,396	117.8	2,253	51.2	17.6	5.2	26.0	588	0.68

Table 3: OverReal statistics. Label distribution: share of the gold split labeled disruptive over-realization (DO), silent over-realization (SO), integrated realization (I) and withheld realization (W).

B.2 The ambiguous split

An image enters the ambiguous split when the label sets of its two annotators are disjoint (§4.2). Table 4 sorts the 588 such images, 20.7% of the 2,841 annotated images, by what the two annotators disagree on. In 245 of them, the annotators agree on whether the target is over-realized; in 266 of them, one annotator read the image as over-realized and the other as correct; the remaining 77 involve the label *other*.

	Existence-canceling	Mental-state	Figurative	Perspectival	All
Form of the outcome only	6	52	125	62	245
Over-realized against correct	13	43	141	69	266
One label is *other*	4	17	33	23	77

Table 4: The ambiguous split of OverReal-Det by the kind of disagreement.

C Automatic Annotator

C.1 Cascade design

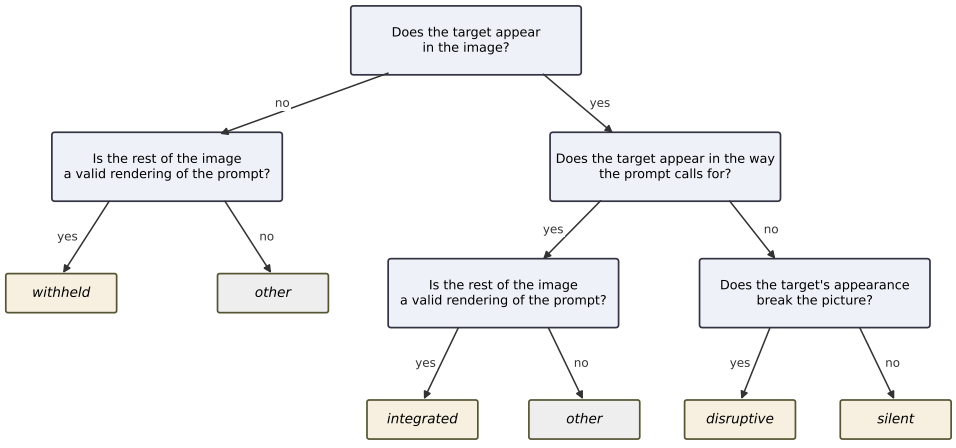


Figure 8: The annotation cascade as a decision tree. Each decision node (blue) is one yes/no question put to the annotator in a separate call, with the image and the prompt; the leaves (italic) are the labels. The family-specific wording of the questions is given in Appendix E.1.

Rather than asking for one of four labels at once, the annotator follows the decision tree of Figure 8, one yes/no question per call. The first question is whether the target appears in the image. If it

does not, the image is *withheld* when the rest of the image is a valid rendering of the prompt, and *other* when it is not. If it does, the next question is whether the target appears in the way the prompt calls for. If so, the same validity question decides between *integrated* and *other*; if not, the image is *disruptive* when the target's appearance breaks the picture, and *silent* when it does not. The questions that instantiate these decisions differ by family for higher precision. The full question set per family is given in Appendix E.1.

C.2 Self-evolution of the AI annotator

The AI annotator is adapted to the task without updating any model weights, by a form of self-evolution: the model annotates a development set with its current questions, revises the questions from the mistakes it made, and repeats until its agreement with the human labels stops improving. We build it on the human labels of OverReal-Det, split evenly, stratified by family and generator, into a development split and a test split; the development split alone drives the self-evolution, and the frozen annotator is then evaluated once on the test set. In each round the model scores the development split, receives the images on which its label disagrees with the human one, traces each disagreement to the question that produced it, and proposes a rewrite of that question; we accept the rewrite and score the revised questions again. The best wording is selected per family.

C.3 Agreement with the human labels

The annotator uses Claude Opus 5 (Anthropic, 2026) as the base model. Table 5 reports the AI-human agreement and human-human agreement in Cohen's κ. Conclusions are given as follows.

Self-evolution improves the cascade annotator. Asking the model for one of the four labels outright is unreliable: a multiple-choice prompt over the four labels reached only $\kappa = 0.46$ against the human labels. Self-evolution of the cascade questions then raised the agreement with the human labels on the development split from $\kappa = 0.61$ to 0.68 over the four labels, and from 0.65 to 0.76 on the split between over-realization and correct output.

The annotator agrees with the human labels nearly as well as the humans agree with each other. Over the four labels, κ is 0.58 overall against 0.70 between the humans; on the split between over-realization and correct output, which the rates of Table 7 rest on, it is 0.70 against 0.75, a substantial agreement. The automatic labels are therefore reliable enough for the rates we report, if not for a fine reading of every image.

Part of the gap to the humans is the ambiguity of the task. The κ values in Table 5 are not uniformly high, but they come close to the agreement between the humans on this task. Deciding whether the target appears is easy for both the annotator and the humans ($\kappa = 0.83$ against 0.86). Deciding what its appearance amounts to is harder for both, because the outcome labels are defined by the effect that the target's appearance has on the image, and this effect is a matter of degree, as described in the limitations; on mental-state prompts the annotator is close to the humans (0.42 against 0.47). On perspectival prompts it falls short of them (0.41 against 0.75), so there the gap is the model's.

Decision		Existence-canceling	Mental-state	Figurative	Perspectival	Overall
Four labels	AI-human	1.00	0.42	0.60	0.41	0.58
	humans	0.83	0.47	0.67	0.75	0.70
Target appears	AI-human	1.00	0.74	0.82	0.71	0.83
	humans	0.88	0.84	0.83	0.87	0.86
Over-realized vs. correct	AI-human	1.00	0.53	0.69	0.60	0.70
	humans	0.88	0.49	0.74	0.79	0.75

Table 5: Cohen's κ of the automatic annotator (Claude Opus 5) against the human gold labels on the held-out test split of OverReal-Det (1,118 images), and of the two human annotators against each other on the 2,688 doubly annotated images with one label each, for the four-outcome label, for whether the target appears, and for the collapse to over-realized against correct.

	Existence-canceling					Mental-state					Figurative					Perspectival				
	CS↑	VQ↑	PS↑	HP↑	LJ↑	CS↑	VQ↑	PS↑	HP↑	LJ↑	CS↑	VQ↑	PS↑	HP↑	LJ↑	CS↑	VQ↑	PS↑	HP↑	LJ↑
GPT-Image	.67	**.87**	**23.3**	**.31**	**9.4**	**.75**	.78	**22.1**	**.33**	**8.2**	**.69**	**.84**	**22.1**	**.31**	**7.5**	.74	.66	22.4	.29	**5.6**
Nano Banana	.64	.82	22.9	.28	8.7	.71	**.78**	21.8	.30	7.8	.63	.83	21.6	.28	**7.5**	.71	.76	22.3	.28	4.5
Ideogram	.61	.82	22.7	.29	8.3	.66	.62	21.2	.29	6.9	.57	.73	21.1	.27	6.9	.70	.70	22.4	.30	3.6
FLUX.1-dev	.66	.59	22.9	.30	5.8	.73	.66	21.7	.30	6.2	.62	.70	21.4	.29	5.5	.73	.69	22.6	.30	3.2
Qwen-Image	**.73**	.51	23.2	.31	5.2	.75	.75	22.0	.32	6.7	.66	.79	21.7	.30	6.3	**.77**	**.77**	**22.8**	**.31**	3.3
OmniGen2	.70	.58	22.9	.30	5.6	.71	.64	21.5	.32	5.8	.63	.71	21.2	.28	5.3	.75	.72	22.3	.30	3.0
SD3.5-Medium	.69	.69	22.5	.28	6.5	.72	.65	21.4	.29	6.0	.62	.69	21.1	.26	5.5	.76	.69	21.9	.28	3.2
SD3.5-Large	.71	.50	22.9	.30	5.0	.73	.67	21.8	.31	6.3	.66	.71	21.6	.29	5.4	.75	.71	22.5	.30	3.6

Table 6: Conventional evaluation of the eight systems on the raw prompts of OverReal-Gen: mean CLIPScore (CS), VQAScore (VQ), PickScore (PS), HPSv2 (HP) and LLM judge (LJ) per family.

D Results

D.1 Conventional Evaluation: LLM Judge Details

	Existence-canceling					Mental-state					Figurative					Perspectival				
Generator	DO↓	SO↓	I↑	W↑	u↓	DO↓	SO↓	I↑	W↑	u↓	DO↓	SO↓	I↑	W↑	u↓	DO↓	SO↓	I↑	W↑	u↓
(a) OverReal-Gen																				
GPT-Image	**.02**	.00	.00	**.98**	.02	.35	.26	.28	.11	.08	.22	.29	**.16**	.33	.05	**.54**	.15	**.03**	**.28**	.05
Nano Banana	.12	.00	.00	.88	.03	.37	**.19**	**.37**	.07	.09	.15	.26	.15	.43	.09	.78	.17	.02	.03	.07
Ideogram	.08	.00	.00	.92	.02	**.29**	.24	.03	**.44**	**.02**	**.07**	**.25**	.03	**.65**	.04	.65	.18	.00	.17	**.02**
FLUX.1-dev	.43	.00	.00	.57	.02	.44	.28	.03	.24	**.02**	.09	.30	.14	.46	**.02**	.56	.19	.01	.24	.03
Qwen-Image	.64	.00	.00	.36	.01	.67	.22	.01	.10	.03	.28	.33	.07	.31	**.02**	.67	.22	.01	.10	.04
OmniGen2	.51	.00	.00	.49	.01	.64	.22	.03	.11	**.02**	.19	.27	.08	.46	.03	.64	.16	.01	.18	.03
SD3.5-Medium	.34	.00	.00	.66	**.00**	.66	.21	.01	.12	**.02**	.09	.31	.06	.54	**.02**	.67	**.11**	.00	.22	**.02**
SD3.5-Large	.64	.00	.00	.36	.01	.61	.23	.03	.13	**.02**	.20	.40	.11	.28	**.02**	.59	.16	.02	.22	.03
(b) In-the-wild prompts																				
GPT-Image	.28	.00	.00	.72	—	.21	.12	.64	.03	—	.01	.21	.47	.31	—	.59	.23	.00	.18	—
Nano Banana	.38	.00	.00	.62	—	.26	.10	.58	.06	—	.01	.24	.35	.40	—	.70	.07	.01	.21	—
Ideogram	.55	.00	.00	.45	—	.28	.16	.14	.41	—	.02	.27	.26	.45	—	.74	.15	.00	.12	—
FLUX.1-dev	.70	.00	.00	.30	—	.30	.17	.04	.49	—	.02	.29	.14	.54	—	.65	.13	.00	.22	—
Qwen-Image	.72	.00	.00	.28	—	.51	.24	.06	.19	—	.04	.26	.36	.34	—	.75	.12	.00	.13	—
OmniGen2	.67	.00	.00	.33	—	.42	.18	.04	.36	—	.02	.26	.26	.46	—	.66	.09	.00	.25	—
SD3.5-Medium	.69	.00	.00	.31	—	.33	.20	.03	.44	—	.03	.26	.23	.47	—	.63	.10	.00	.26	—
SD3.5-Large	.70	.00	.00	.30	—	.38	.21	.03	.38	—	.02	.26	.29	.43	—	.69	.10	.01	.19	—

Table 7: Main results: proportion of images under each outcome label (definition in §3.2), namely disruptive over-realization (DO), silent over-realization (SO), integrated realization (I) and withheld realization (W), and the mean unprompted appearance rate (u, definition in §5.3). (a) On the curated prompts of OverReal-Gen. (b) On the 397 in-the-wild prompts of §6.1, one image per prompt and system; u was not measured there.

Both judges score the same images with the verbatim VIEScore text-to-image semantic-consistency template: Claude Opus 5 (`claude-opus-5`, accessed 2026-09) and GPT-5.2 (`gpt-5.2-2025-12-11`). Table 6 reports the Opus rating; Table 8 gives the GPT-5.2 rating for the same cells and Table 9 the agreement of the two judges.

D.2 Intervals for the over-realization rates

Table 10 gives the over-realization rate of Table 7, the sum of the DO and SO columns, with half the width of a 95% bootstrap interval.

D.3 Outcomes across the two seeds

For the raw prompts of OverReal-Gen, each model generates two images per prompt under different seeds. Table 11 conditions on the label of one image and gives the distribution of the label of the other, over the labeled sample of §5.3 (172 silent, 428 correct and 588 disruptive conditioning images). Existence-canceling prompts have no silent images.

	Existence-canceling	Mental-state	Figurative	Perspectival
GPT-Image	9.5	8.8	9.2	6.3
Nano Banana	9.1	8.8	9.0	6.2
Ideogram	9.0	7.8	8.2	5.2
FLUX.1-dev	6.2	7.7	7.9	4.9
Qwen-Image	5.0	8.3	8.7	5.3
OmniGen2	6.0	7.6	8.0	5.0
SD3.5-Medium	7.1	7.4	7.4	5.0
SD3.5-Large	4.9	7.6	8.1	5.0

Table 8: LLM judge (LJ) ratings by GPT-5.2, the same cells as the LJ column of Table 6.

Family	Opus mean	GPT mean	Pearson r	Mean $\lvert\Delta\rvert$	n
Existence-canceling	6.86 ± 2.83	7.37 ± 3.27	0.90	1.02	1700
Mental-state	6.78 ± 1.52	7.97 ± 1.30	0.62	1.32	1750
Figurative	6.71 ± 1.95	8.29 ± 1.39	0.57	1.69	1800
Perspectival	3.63 ± 1.81	5.27 ± 2.38	0.56	1.98	1800
Overall	5.98 ± 2.49	7.22 ± 2.52	0.78	1.51	7050

Table 9: Between-judge statistics: mean rating $\pm$ standard deviation per judge, correlation, and mean absolute rating difference.

D.4 Conventional metrics on matched pairs

For the raw prompts of OverReal-Gen, each model generates two images per prompt under different seeds. Over the labeled sample of §5.3, we take the pairs whose two images differ in outcome, one over-realized and one correct, and ask which of the two each conventional metric scores higher. Table 12 gives this winning rate of the over-realized image, by kind of over-realization and per family.

The conventional metrics reward over-realization or are indifferent to it. CLIPScore, PickScore and HPSv2 score the over-realized image higher in 64% to 66% of the pairs. VQAScore is split: on existence-canceling prompts it scores the correct image higher in 95% of the pairs, but on mental-state and figurative prompts it scores the over-realized image higher in 67% and 75%, and on perspectival prompts it scores either image higher equally often (50%). The LLM judge scores the correct image higher in 88%. The preference of the three metrics is strongest on the disruptive images (69% to 72%); on the pairs whose over-realized image is silent they score either image higher about equally often (44% to 50%), so they do not penalize the silent target either. A metric that credited correct realization would score the correct image higher in both cases; the conventional metrics reward the disruptive over-realization and are indifferent to the silent one.

D.5 Prompt expansion: full results

Table 13 gives the outcome distribution of the five open generators on the raw prompts and on the prompts rewritten by the Qwen-Image rewriter and by Ideogram Magic Prompt, per family, with the change from the raw prompt in parentheses. §6.2 discusses the over-realization rates; Appendix D.7 classifies what each rewrite does with the target.

D.6 Experiment on the in-the-wild set

The curated prompts of OverReal-Gen are written or synthesized for the benchmark, and the hand-written families are admitted only after a pilot model over-realized them (§4.1). We also evaluate the systems on an in-the-wild set of prompts that users wrote for their own purposes, selected on the wording of the prompt alone.

Sources and sampling We pool the prompts of DiffusionDB (Wang et al., 2023) and Pick-a-Pic (Kirstain et al., 2023), collapse whitespace, drop duplicates, drop DiffusionDB prompts whose text or

	Existence-canceling	Mental-state	Figurative	Perspectival	All families
GPT-Image	.02 ± .03	.61 ± .12	.51 ± .11	.69 ± .10	.43 ± .05
Nano Banana	.12 ± .09	.56 ± .11	.41 ± .10	.95 ± .04	.49 ± .05
Ideogram	.08 ± .07	.53 ± .11	.32 ± .09	.83 ± .08	.41 ± .04
FLUX.1-dev	.43 ± .13	.73 ± .11	.39 ± .10	.75 ± .10	.53 ± .06
Qwen-Image	.64 ± .13	.89 ± .07	.62 ± .10	.89 ± .06	.73 ± .05
OmniGen2	.51 ± .11	.86 ± .08	.45 ± .11	.81 ± .08	.61 ± .05
SD3.5-Medium	.34 ± .11	.87 ± .07	.40 ± .10	.78 ± .09	.54 ± .05
SD3.5-Large	.64 ± .12	.84 ± .08	.61 ± .10	.76 ± .10	.68 ± .06

Table 10: Over-realization rate of each system per family and over all families, weighting the families by their prompt counts in OverReal-Gen, ± half the width of a 95% bootstrap interval over prompts.

	Existence-canceling	Mental-state	Figurative	Perspectival	All
$P(\mathrm{DO}_1 \mid \mathrm{DO}_0)$	.84	.77	.65	.71	.79
$P(\mathrm{SO}_1 \mid \mathrm{DO}_0)$	.00	.17	.10	.20	.10
$P(\mathrm{C}_1 \mid \mathrm{DO}_0)$	.16	.06	.26	.09	.12
$P(\mathrm{DO}_1 \mid \mathrm{SO}_0)$	—	.44	.05	.64	.33
$P(\mathrm{SO}_1 \mid \mathrm{SO}_0)$	—	.50	.76	.32	.57
$P(\mathrm{C}_1 \mid \mathrm{SO}_0)$	—	.06	.19	.04	.10
$P(\mathrm{DO}_1 \mid \mathrm{C}_0)$	.16	.23	.09	.19	.16
$P(\mathrm{SO}_1 \mid \mathrm{C}_0)$	.00	.09	.13	.03	.04
$P(\mathrm{C}_1 \mid \mathrm{C}_0)$	.84	.68	.78	.78	.80

Table 11: Label of the other seed's image given the label of one image of the same prompt and open model, on the raw prompts of OverReal-Gen. $P(y_1 \mid y_0)$ is the share of images with label y_0 whose other-seed image has label y_1; DO disruptive, SO silent, C correct (integrated or withheld). Existence-canceling prompts have no silent images.

image carries an NSFW score of 0.5 or more, keep prompts of 3 to 120 words that are at least 90% ASCII, and draw 7,000 DiffusionDB and 3,000 Pick-a-Pic prompts at random.

Classification Each of the 10,000 prompts is classified into one of the four families of §3.1 or none, based on the cue words of each family. Of the 10,000 prompts, 202 (2.0%) qualify: 62 existence-canceling, 3 mental-state, 129 figurative and 8 perspectival. To obtain enough prompts of the two rare families, we add another 135 mental-state and 60 perspectival prompts from the rest of the datasets, also based on cue words.

Experiments The 397 prompts are generated under the settings of §5, one image per prompt and system, and the images are labeled by the automatic annotator of §5.3. Table 7(b) gives the results, and the conclusions of §6.1 carry over. First, every system over-realizes in every family: pooled over all systems, the over-realization rate is 0.60 on existence-canceling prompts, 0.28 on figurative, 0.51 on mental-state and 0.80 on perspectival. Second, the form of the failure follows the family: perspectival targets break the scene (mostly DO), figurative targets slip in silently (almost entirely SO), and mental-state targets sit between; the proprietary systems handle mental-state prompts largely by integrated depiction (I of 0.58 to 0.64), which the open models rarely manage. Third, the curated prompts run 0.1 to 0.3 higher than the in-the-wild prompts in the mental-state and figurative families. Existence-canceling goes the other way: negations in the wild mostly cancel an attribute of the subject (*no nose*, *faceless*, *without laces*) rather than an object in the scene, and every system, the proprietary ones included, over-realizes at least 28% of them, far above its rate on the templated prompts of §4.1.

Conclusions First, apart from existence-canceling, the systems behave on the prompts that users wrote much as they do on the prompts we wrote: the pattern across families and systems is the same, and the rates differ by at most 0.3. The curated prompts are harder, as intended, but they do not produce a failure that natural prompts lack, so the results of §6 are not an artifact of the curation. Second, the four situations account for about 2% of the prompts in current text-to-image datasets, so

		by kind of over-realization		by family			
Metric	Winning rate	Disruptive	Silent	Existence-canceling	Mental-state	Figurative	Perspectival
CLIPScore	.64 ± .10	.69	.44	.75	.56	.60	.38
VQAScore	.33 ± .13	.25	.80	.05	.67	.75	.50
PickScore	.66 ± .10	.72	.44	.72	.67	.60	.50
HPSv2	.65 ± .11	.69	.50	.65	.72	.70	.38
LLM judge	.12 ± .08	.06	.50	.00	.27	.33	.17

Table 12: Winning rate: share of matched pairs (same prompt, same open model, two seeds; one image over-realized, one correct) in which the metric scores the over-realized image higher, ± half the width of a 95% bootstrap interval over prompts.

	Existence-canceling				**Mental-state**				**Figurative**				**Perspectival**			
Generator + rewriter	DO↓	SO↓	I↑	W↑	DO↓	SO↓	I↑	W↑	DO↓	SO↓	I↑	W↑	DO↓	SO↓	I↑	W↑
FLUX.1-dev (raw)	.43	.00	.00	.57	.44	.28	.03	.24	.09	.30	**.14**	.46	**.56**	.19	.01	**.24**
+ Qwen rewriter	.33 (–.10)	.00 (0)	.00 (0)	.67 (+.10)	.41 (–.03)	**.25** (–.03)	**.05** (+.02)	.29 (+.05)	.05 (–.04)	.21 (–.09)	.09 (–.05)	.65 (+.19)	.65 (+.09)	.20 (+.01)	.01 (0)	.14 (–.10)
+ Ideogram Magic	**.11** (–.32)	.00 (0)	.00 (0)	**.89** (+.32)	**.15** (–.29)	.32 (+.04)	.04 (+.01)	**.49** (+.25)	**.00** (–.09)	**.20** (–.10)	.04 (–.10)	**.76** (+.30)	.62 (+.06)	**.15** (–.04)	.01 (0)	.22 (–.02)
Qwen-Image (raw)	.64	.00	.00	.36	.66	**.22**	.01	.10	.28	.33	.07	.31	.67	.22	**.01**	.10
+ Qwen rewriter	.81 (+.17)	.00 (0)	.00 (0)	.19 (–.17)	.50 (–.16)	.35 (+.13)	**.10** (+.09)	.05 (–.05)	.14 (–.14)	.28 (–.05)	**.12** (+.05)	.46 (+.15)	.72 (+.05)	.20 (–.02)	.00 (–.01)	.08 (–.02)
+ Ideogram Magic	**.08** (–.56)	.00 (0)	.00 (0)	**.92** (+.56)	**.26** (–.40)	**.22** (0)	.06 (+.05)	**.46** (+.36)	**.04** (–.24)	**.22** (–.11)	.07 (0)	**.67** (+.36)	**.64** (–.03)	**.19** (–.03)	**.01** (0)	**.16** (+.06)
OmniGen2 (raw)	.51	.00	.00	.49	.64	**.22**	.03	.11	.18	.26	**.08**	.45	**.63**	.16	**.01**	.18
+ Qwen rewriter	.20 (–.31)	.00 (0)	.00 (0)	.80 (+.31)	.40 (–.24)	.31 (+.09)	**.08** (+.05)	.21 (+.10)	.09 (–.09)	.18 (–.08)	.06 (–.02)	.67 (+.22)	.66 (+.03)	.18 (+.02)	.00 (–.01)	.16 (–.02)
+ Ideogram Magic	**.09** (–.42)	.00 (0)	.00 (0)	**.91** (+.42)	**.16** (–.48)	.27 (+.05)	.07 (+.04)	**.50** (+.39)	**.02** (–.16)	**.14** (–.12)	.07 (–.01)	**.76** (+.31)	**.63** (0)	**.14** (–.02)	.00 (–.01)	**.23** (+.05)
SD3.5-Medium (raw)	.34	.00	.00	.66	.66	**.21**	.01	.12	.09	.31	.06	.53	.66	**.11**	.00	.22
+ Qwen rewriter	.27 (–.07)	.00 (0)	.00 (0)	.73 (+.07)	.47 (–.19)	.22 (+.01)	**.05** (+.04)	.26 (+.14)	.09 (0)	.23 (–.08)	**.12** (+.06)	.55 (+.02)	.68 (+.02)	.18 (+.07)	**.01** (+.01)	.11 (–.11)
+ Ideogram Magic	**.07** (–.27)	.00 (0)	.00 (0)	**.93** (+.27)	**.23** (–.43)	.28 (+.07)	.00 (–.01)	**.48** (+.36)	**.02** (–.07)	**.19** (–.12)	.04 (–.02)	**.75** (+.22)	**.62** (–.04)	**.11** (0)	.00 (0)	**.27** (+.05)
SD3.5-Large (raw)	.64	.00	.00	.36	.61	**.23**	.03	.13	.20	.40	**.11**	.28	**.58**	.16	**.02**	**.22**
+ Qwen rewriter	.34 (–.30)	.00 (0)	.00 (0)	.66 (+.30)	.43 (–.18)	.30 (+.07)	**.04** (+.01)	.23 (+.10)	.07 (–.13)	.30 (–.10)	.09 (–.02)	.54 (+.26)	.67 (+.09)	.22 (+.06)	.00 (–.02)	.09 (–.13)
+ Ideogram Magic	**.11** (–.53)	.00 (0)	.00 (0)	**.89** (+.53)	**.27** (–.34)	.24 (+.01)	.02 (–.01)	**.47** (+.34)	**.03** (–.17)	**.17** (–.23)	.03 (–.08)	**.77** (+.49)	.69 (+.11)	**.15** (–.01)	.01 (–.01)	.15 (–.07)

Table 13: Performance of prompt expansion.

existing benchmarks under-represent them and have not evaluated them. Mental-state and perspectival prompts are the rarest, under 0.1% each. We keep them in OverReal nonetheless: the same situations arise whenever a prompt describes what a character thinks or what a viewpoint hides, which is routine in narrative prompts and in video generation. A model that cannot keep such a target out of a single image will not keep it out of a video either.

D.7 How the rewriters treat the target

For the analysis of §6.2 we read every rewritten prompt and recorded what became of the target. Each rewrite falls into one of four categories:

- **dropped**: the target is no longer mentioned, under any name;
- **explicitly excluded**: the rewrite states that the target is absent or must not appear ("no keyboard is present", "the letter p is obscured by shadow");
- **mentioned**: the target is named but nothing in the rewrite asks for it to be shown, whether it keeps the status the original prompt gave it ("remembering the trumpet", "as fierce as a tiger", "inside the closed box is a teapot") or not;
- **explicitly depicted**: the rewrite describes the target as something to render in view ("a tiger stands in the corner", "visible through a gap in the lid", "the hidden side reveals a red tower").

A first pass assigned the categories by rule, matching the target's words against the rewrite and looking for negation and visibility phrases near them; every rewrite that the rules flagged, every rewrite marked as dropped that still produced an over-realized image, and a random sample of the rest were then read and corrected by hand. Eight existence-canceling items whose target is a normal part of the scene, such as a keyboard beside a laptop, are left out of this analysis. Table 14 gives the distribution per family and rewriter, and Table 15 the over-realization rate of the corresponding images by category.

Family	Rewriter	Dropped	Excluded	Mentioned	Depicted
Existence-canceling	Qwen	0	98	2	0
	Ideogram	95	2	2	0
Mental-state	Qwen	0	1	64	35
	Ideogram	41	0	40	19
Figurative	Qwen	15	0	77	8
	Ideogram	55	0	36	9
Perspectival	Qwen	0	1	42	56
	Ideogram	10	2	44	45

Table 14: How the rewriters treat the target, as a percentage of the rewritten prompts of each family: the target is dropped, explicitly excluded ("no keyboard is present"), mentioned, or explicitly to be depicted ("visible through a gap").

Family	Dropped	Excluded	Mentioned	Depicted
Existence-canceling	0	35	5	—
Mental-state	4	—	71	82
Figurative	4	—	33	75
Perspectival	2	100	86	87
Overall	2	38	60	84

Table 15: Over-realization rate (percent) of the images generated from rewritten prompts, by how the rewrite treats the target, pooled over both rewriters and the five open generators. Cells with fewer than 10 images are left blank.

D.8 TEXT ENCODER

The backbone never sees the prompt, only the encoder's output, so we ask whether that output still distinguishes the original prompt from one that calls for the target. Both experiments below use the plain-mention controls of §6.3: in a control, the words that keep the target out of the scene are replaced by words that put it in, and the target word is kept, so that *"The coach is as fierce as a tiger"* becomes *"The coach is standing next to a tiger"* (one example per family in Table 18).

Over-realization is not due to a lack of information in the encoder output, except on perspectival prompts. A linear probe is trained to tell the two prompts apart from the contextual vector of the target word, at the layer the pipeline feeds to the backbone, for the four encoders of the open models: CLIP-L and CLIP-bigG (Radford et al., 2021) (SD3.5), T5-XXL (Raffel et al., 2020) (SD3.5 and FLUX) and Qwen2.5-VL-7B (Bai et al., 2025) (Qwen-Image). It is trained on all four families together and tested with cross-validation grouped by target word. On existence-canceling, mental-state and figurative prompts the probe separates the two prompts at 81% to 100%, while on perspectival prompts it stays close to chance, at 53% to 62%, at every layer (Table 16, Figure 9). Furthermore, within each of the three families, the Spearman correlation between an item's probe margin and its over-realization rate lies between -0.28 and 0.28, with no consistent sign across encoders. Thus, for existence-canceling, mental-state and figurative prompts, over-realization is not explained by the encoding process; on perspectival prompts the probe finds no linear evidence in the encoder output that the target is out of view.

Qwen-Image's encoder knows the target should not be visible, and the generator shows it anyway. Qwen-Image's encoder is also an instruction model, so we ask it directly: *"If this prompt were rendered as an image exactly as described, would a ⟨target⟩ be visible in the image? Answer yes or no."*, expecting *no* for the original prompt and *yes* for its control. It answers correctly on 100%, 75% and 84% of the existence-canceling, mental-state and figurative prompts and on 56% of the perspectival ones (Table 17); the same weights then condition images in which the target appears in more than half of the cases (Table 7). This is another piece of evidence that the information survives the encoding process for existence-canceling, mental-state and figurative prompts; for the perspectival family we find no evidence that it does.

Encoder	Existence-canceling	Mental-state	Figurative	Perspectival	All
CLIP-L/14	.95 ± .04	.81 ± .06	.87 ± .03	.53 ± .02	.79 ± .02
CLIP-bigG/14	.96 ± .03	.87 ± .05	.87 ± .03	.60 ± .03	.82 ± .02
T5-XXL	.95 ± .04	.89 ± .04	.94 ± .02	.62 ± .03	.85 ± .02
Qwen2.5-VL-7B	1.00 ± .00	.92 ± .04	.91 ± .02	.60 ± .03	.86 ± .01

Table 16: Accuracy of a linear probe that tells an original prompt from its plain-mention control, from the contextual vector of the target word at the layer each pipeline feeds to the backbone, ± half the width of the bootstrap 95% interval. Chance is 50%.

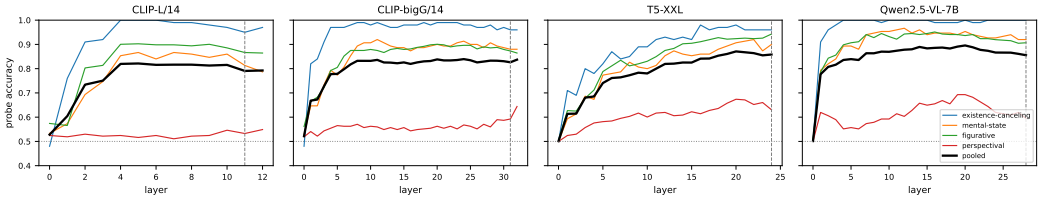


Figure 9: Accuracy of the target-word probe of Table 16 at every layer of each encoder, per family and pooled. Layer 0 is the embedding output; the dashed line marks the layer the pipeline feeds to the backbone. The perspectival family stays near chance at every layer of every encoder.

D.9 Backbone attention: per family

Table 19 gives the attention shares of §6.3 per family. Pairs per family: 100, 150, 472 and 368 for existence-canceling, mental-state, figurative and perspectival prompts; every value pools the two seeds.

The cue words lower the target's attention only slightly, and not at all on perspectival prompts. On existence-canceling, mental-state and figurative prompts the target receives less attention under the original prompt than under the control in every model (ρ_t 0.74 to 0.92). On perspectival prompts ρ_t is 0.97 to 0.99: the cue words make almost no difference, consistent with Appendix D.8, where the perspectival information is already lost in the encoder.

D.10 Backbone attention under prompt expansion

Expanded prompts are about five times longer than the original ones, so for them we report the relative attention r: each token's share divided by the prompt's average share. Table 20 gives the target's r under the original prompt and under the two expansions of §6.2, over the expanded prompts in which the target survives verbatim (both seeds).

Expansion does not lower the target's attention. Under the expanded prompts in which the rewriter kept the target, the target receives at least as much attention as an average word in every model and for both rewriters (r = 1.04 to 1.45). The one exception is Qwen-Image on existence-canceling prompts under the Qwen rewriter (0.58): the rewriter restates the cancellation as a clause of its own, such as *No toilet is visible in the frame*, and inside that clause Qwen-Image gives the target about half the attention of an ordinary word, while FLUX.1-dev and SD3.5-Large read it like any other word (1.05 and 0.99). Figure 10 shows two items under the original prompt, the control and the two rewrites.

E Full Prompts and Annotation Guidelines

E.1 Automatic annotation protocol

Each image is scored by a cascade of yes/no questions, one API call per question, with the image, the question and, from the second question on, the original prompt. `T` stands for the target; the article is

Family	Original prompt (no)	Plain-mention control (yes)	Both
Existence-canceling	1.00	.76	.88
Mental-state	.75	1.00	.87
Figurative	.84	.93	.89
Perspectival	.56	.85	.70

Table 17: Accuracy of Qwen2.5-VL-7B, the text encoder of Qwen-Image, when asked directly whether the target would be visible.

Family	Original prompt	Plain-mention control
Existence-canceling	a cat sitting on a couch with a dog on the floor, with no **book** in sight	a cat sitting on a couch with a dog on the floor, with a **book** in sight
Mental-state	Tim sits alone at a classroom desk, remembering the toy **robot** he lost years ago.	Tim sits alone at a classroom desk, holding the toy **robot** he got years ago.
Figurative	The coach is as fierce as a **tiger**.	The coach is standing next to a **tiger**.
Perspectival	A desk drawer is completely closed. Inside is a green **frog** toy.	A desk drawer is pulled open. Inside is a green **frog** toy.

Table 18: Plain-mention controls: one example per family. The target word is kept; the words that keep it out of the scene are replaced by words that put it in.

adjusted to the target. Every family starts with the same screening question, "Is this image severely corrupted, or unrelated to the scene the prompt describes?", whose yes-answer gives *other*.

Existence-canceling Q1: Is there a T anywhere in this image? (yes: disruptive; no: withheld)

Mental-state Q1: Is there a T anywhere in this image? (no: withheld) Q2: In this image, is a T physically present in the scene, as an actual thing that is there with the person? (no: Q3) Q2b: In this image, is the T right beside the person and prominent in the picture, as opposed to distant, small or in the background? (yes: disruptive; no: silent) Q3: In this image, is a T shown as something the person imagines, remembers or believes, for example inside a thought bubble, a dream haze, a separate inset or a ghostly overlay? (yes: integrated; no: silent)

Figurative Q1: Is there a T anywhere in this image? (no: withheld) Q2: In this image, does the presence of a T make the scene absurd, impossible or nonsensical as a picture of what the prompt describes? (yes: disruptive) Q3: In this image, is a T merged with the person or thing it describes, as a deliberate visual metaphor in a stylized or poster-like composition? (yes: integrated; no: silent)

Perspectival, hidden object Q1: Is there a T anywhere in this image, in any form, including as a print, picture, sticker, logo or drawing? (no: withheld) Q2: According to the prompt, the T should be hidden from the viewer: inside, behind, underneath, or on the far side of something. In this image, is the T nevertheless exposed to the viewer at that very spot, for example because the container, door, curtain, barrier or surface is shown open, transparent, removed, or turned the wrong way? (yes: disruptive) Q3: In this image, is the T shown only indirectly, through a reflection, an inset, a thought bubble, or a picture taken from the in-scene person's own point of view? (yes: integrated; no: silent)

Perspectival, hidden text Q1: In this image, is the written side of the item that carries "T" (the letter, note, page, screen, card or sign the prompt describes) turned toward the viewer, so that its writing can be seen, whether or not it is legible? (no: withheld) Q2: In this image, is the writing "T" shown to the viewer only through a reflection, an inset, a thought bubble, or a picture taken from the in-scene person's own point of view? (yes: integrated) Q3: According to the prompt, the writing "T" should be out of the viewer's sight. In this image, is it nevertheless in plain view, on the very item the prompt describes? (yes: disruptive; no: silent)

		Original prompt			Plain-mention control			Ratios		
Model	Family	target a_{t}	cue a_{c}	other a_{o}	target a'_{t}	replacing a'_{r}	other a'_{o}	$\rho_{\mathrm{t}} = a_{\mathrm{t}}/a'_{\mathrm{t}}$	$\rho_{\mathrm{c}} = a_{\mathrm{c}}/a'_{\mathrm{r}}$	$\gamma = a_{\mathrm{c}}/a_{\mathrm{o}}$
FLUX.1-dev	Existence-canceling	4.0	4.0	5.6	5.1	3.8	5.4	.79	1.15	.71
	Mental-state	3.8	3.0	4.2	4.8	3.5	4.2	.74	.87	.73
	Figurative	5.6	4.7	6.1	6.5	5.6	5.7	.83	.80	.77
	Perspectival	3.6	3.7	4.4	3.7	3.6	4.5	.98	1.04	.83
	All families	4.2	3.8	5.1	5.0	4.1	4.9	.83	.97	.76
SD3.5-Large	Existence-canceling	2.4	1.4	3.1	2.8	1.8	3.0	.85	.82	.46
	Mental-state	2.4	1.3	2.3	2.6	1.6	2.3	.85	.79	.54
	Figurative	4.2	2.1	3.3	4.4	2.6	3.1	.92	.80	.63
	Perspectival	2.1	1.9	2.4	2.1	2.0	2.4	.99	.97	.81
	All families	2.8	1.7	2.8	3.0	2.0	2.7	.90	.84	.60
Qwen-Image	Existence-canceling	5.2	4.6	6.3	6.3	6.2	6.3	.80	.84	.73
	Mental-state	3.7	2.7	5.1	4.0	3.3	5.0	.90	.83	.54
	Figurative	5.1	4.2	7.8	5.4	4.8	7.7	.92	.86	.54
	Perspectival	3.6	3.4	5.2	3.6	3.5	5.2	.97	1.03	.66
	All families	4.4	3.8	6.1	4.8	4.4	6.0	.90	.89	.62

Table 19: Attention in the backbone, per family and averaged over the four families. a: percentage of the image-to-text attention per token of the target words (t), the cue words (c) and the other words (o) under the original prompt; a': the same under the plain-mention control, where the cue words are replaced (r). The ratios pair the original prompt and its control under the same seed.

			FLUX.1-dev			SD3.5-Large			Qwen-Image		
Family	n Qwen	n Ideo.	orig.	Qwen	Ideo.	orig.	Qwen	Ideo.	orig.	Qwen	Ideo.
Existence-canceling	100	2	.77	1.05	—	.85	.99	—	.82	.58	—
Mental-state	136	76	.96	1.60	1.85	1.13	1.05	1.42	.81	1.32	1.05
Figurative	398	196	1.00	1.37	1.35	1.40	1.09	1.14	.78	1.43	1.17
Perspectival	282	232	.87	1.30	1.15	.94	1.00	1.06	.74	1.14	1.01
All families	916	506	.90	1.33	1.45	1.08	1.04	1.21	.79	1.12	1.08

Table 20: Relative attention r of the target words under the original prompt (orig.) and under the expanded prompts of the Qwen rewriter and of Ideogram Magic Prompt (Ideo.), per family and over all families; n is the number of expanded images per family and rewriter. Ideogram Magic Prompt keeps the target in only two existence-canceling prompts, so that cell is left out and the Ideogram mean excludes the family.

E.2 Annotation guide

Annotators labeled through a web interface that showed the image, the prompt, and the target, together with the guide below (reproduced verbatim); each person could pick one or two of the five labels per image.

- **disruptive**: the target appears, and its appearance causes at least one of: (1) the image breaks common sense or logic, e.g. a reader facing the back of the letter they are reading, a screen facing the camera instead of its user, a tiger on a sports field; (2) the image contradicts the prompt, e.g. the prompt says the target is absent, or sealed inside an opaque box, yet it is visible.
- **silent**: the target appears; unexpected, but not disruptive, typically as background or a small detail. E.g. "her face is like a red apple" shown as a red-faced child eating an apple; "the roof is like a banana" with banana trees in the background.
- **integrated**: the target appears, but in a natural way, within what a normal viewer would accept. E.g. "the roof is like a banana" as a banana-shaped roof and no banana trees; a thought bubble depicting what a character imagines; contents shown inside a closed drawer as such; a poster style that merges a person with their attribute.
- **withheld**: the target does not appear, and the image does not contradict the prompt. E.g. something in a closed bag shown only as the closed bag; "her face is like a red apple" as just a red-faced child.
- **other**: the image is severely corrupted, or does not match the prompt at all.

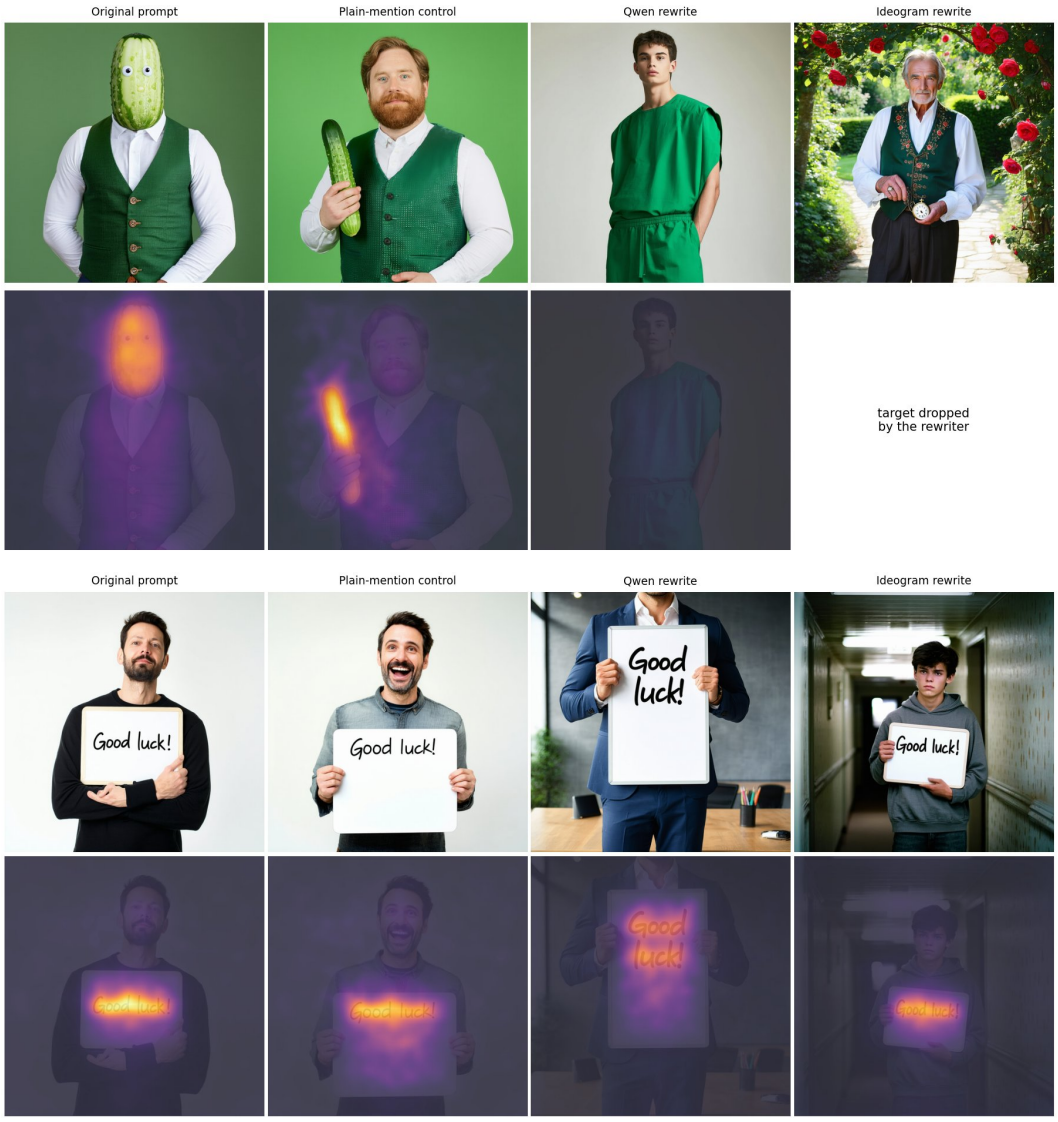


Figure 10: Two items of Table 2 under four prompts, SD3.5-Large: the original prompt, its plain-mention control, the Qwen rewrite and the Ideogram Magic Prompt rewrite (top row: the image; bottom row: attention from the image tokens to the target words in the first denoising steps). Top: the cucumber. The Qwen rewrite keeps *cucumber* only as a color (*"the deep, fresh hue of a ripe cucumber"*); the backbone draws no cucumber and the map stays flat. The Ideogram rewrite drops the word, so there is no map. Bottom: the whiteboard whose *"Good luck!"* faces away from the camera. Both rewriters keep the words, both images show them, and under all four prompts the attention sits on the writing.

You may select two labels, but please prefer a single label; pick two only when the case is genuinely ambiguous.