HIP-BENCH: A BENCHMARK FOR HIDDEN-PICTURE OPTICAL ILLUSION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual generative models have demonstrated impressive capabilities in synthesizing high-quality images from multimodal inputs. However, their ability to generate hidden-picture optical illusions, where hidden targets are integrated into background images and emerge only under specific viewing conditions, remains underexplored. To fill this gap, we propose a novel benchmark for **Hi**dden-**P**icture Optical Illusion Generation (**HiP-Bench**). Specifically, we construct a dataset of 512 samples and use 35 models to produce such illusions across three paradigms: (i) text-to-image, (ii) instruction-based image editing, and (iii) image-conditioned generation. A panel of human experts then evaluates 13,824 generated images across four dimensions, primarily focusing on illusion fidelity and shape consistency. Through five research questions, our analysis reveals the capability boundaries of existing models. Moreover, we explore **M**ulti-Scale **D**istance **S**imulation (**MDS**), an automated baseline based on multimodal large language models that assesses illusion effectiveness through multi-scale visual inputs. Overall, this study establishes a reliable benchmark to facilitate and evaluate future progress.

1 INTRODUCTION

Hidden-picture optical illusions integrate hidden targets into background images, creating a dual visual experience. At close range, observers primarily perceive the background, while the hidden target becomes apparent when the image is viewed from a distance (Gao et al., 2025). This mechanism supports applications such as human-specific CAPTCHAs for information security (Ding et al., 2025), hidden-content moderation (Chen & Soydaner, 2026), and controllable visual stimuli for cognitive training and medical diagnosis (Gomez-Villa et al., 2025). However, handcrafting them requires fine-grained tuning of texture, color, and shape, making automatic generation valuable for scalable customization. It also provides a stringent capability test for visual generative models, since the target must be encoded into the background, requiring simultaneous control of local texture realism and global shape structure.

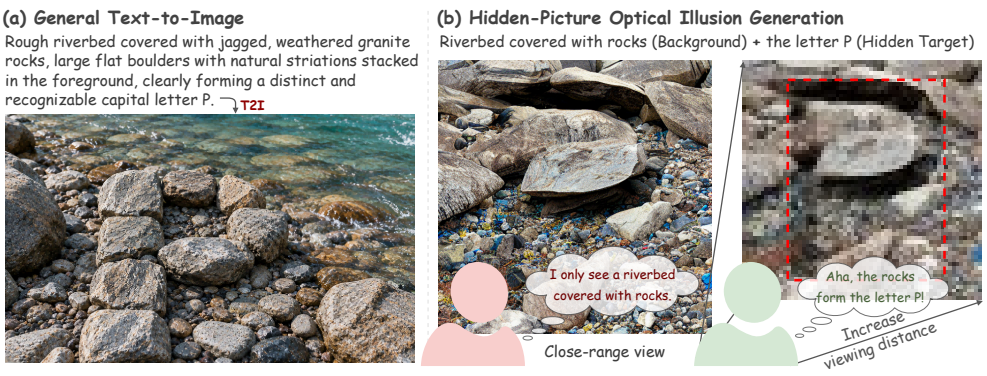

Figure 1: Comparison of general text-to-image and hidden-picture optical illusion generation. Rather than rendering concepts independently, hidden targets are concealed within background images only when viewed from a distance, requiring precise spatial control over conflicting semantics.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

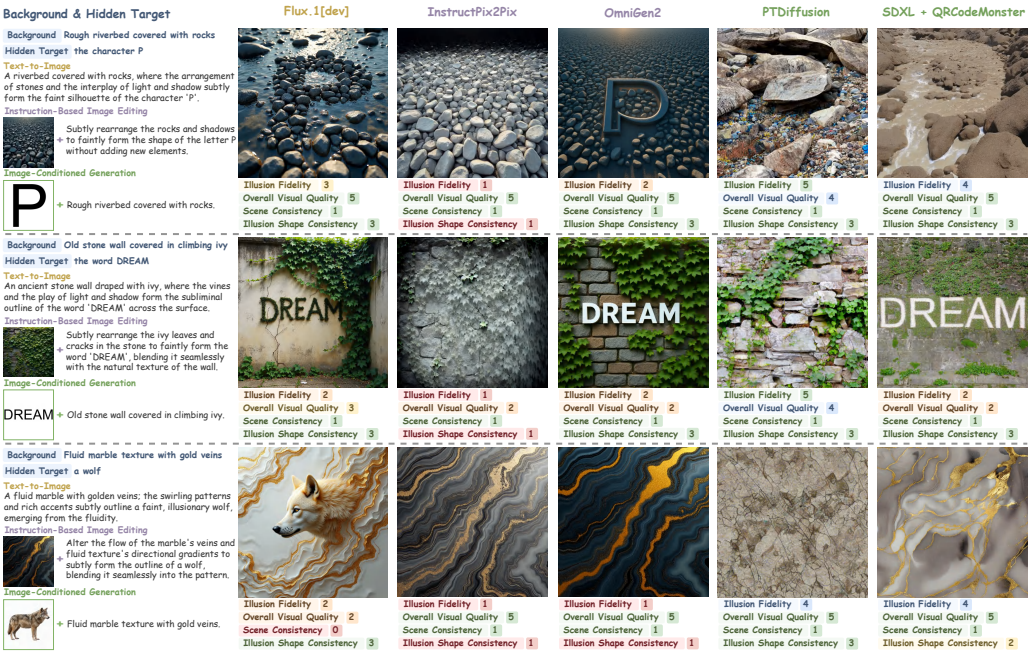

Figure 2: Examples of HiP-Bench (see more in Appendix A.1). For each row, we present a background–hidden target pair and corresponding inputs for three generative paradigms: (i) text-to-image, (ii) instruction-based image editing, and (iii) image-conditioned generation, accompanied by output images from distinct models with human-annotated scores on four dimensions.

Recent advances in visual generative models (Rombach et al., 2022; Black Forest Labs, 2024; Chen et al., 2025b) have driven breakthroughs in various benchmarks evaluating visual fidelity, semantic alignment, and compositional reasoning (Li et al., 2026). As Figure 1 shows, unlike general text-to-image generation, which renders distinct concepts independently, hidden-picture optical illusion generation requires precise spatial control (Zhang et al., 2023a; Mo et al., 2024) to encode conflicting semantics in one space (Burgert et al., 2024; Geng et al., 2024) without compromising the visual coherence of either element. Although existing visual illusion benchmarks span diverse illusion categories and tasks (Zhang et al., 2023b; Shahgir et al., 2024; Guan et al., 2024; Zhang et al., 2025; Sun et al., 2026), benchmarks dedicated to hidden-picture optical illusions remain limited to understanding-oriented tasks, including recognition, classification, and OCR (Hemmat et al., 2024; Rostamkhani et al., 2025; Li et al., 2025). Consequently, the absence of a dedicated benchmark poses unique challenges for assessing this task, which is the focus of this paper.

Essential evaluations for visual generation include image quality (e.g., NIQE (Mittal et al., 2013), LAION-Aes (Schuhmann et al., 2022)) and text-image alignment (e.g., CLIP (Radford et al., 2021), VQAScore (Lin et al., 2025)). However, these metrics lack mechanisms to evaluate concealment quality and fail to accurately measure text-image alignment, as their underlying Vision-Language Models struggle to perceive hidden targets (Hemmat et al., 2024; Rostamkhani et al., 2025; Li et al., 2025). Specialized evaluations for hidden targets include concealment quality (e.g., HVS-based assessment models (Sun et al., 2025) and Context-measure (Wang et al., 2026)) and concealment accuracy (e.g., F-measure (Margolin et al., 2014), S-measure and E-measure (Fan et al., 2017; 2018)). However, these metrics are neither specifically designed for hidden-picture optical illusions nor practical for some generative tasks, as they require ground-truth masks or clean backgrounds, which are rarely provided. This necessitates a benchmark for hidden-picture optical illusion generation.

To address this gap, we propose a novel evaluation dataset **HiP-Bench** (Figure 2), which aims to measure the capabilities of visual generative models in synthesizing hidden-picture optical illusions. The dataset comprises 512 pairs of backgrounds and hidden targets, with inputs tailored for three generative paradigms (text-to-image (T2I), instruction-based image editing (Editing), and image-conditioned generation (CondGen)). We use these inputs to generate images with 35 models. For evaluation, an expert team conducts parallel human scoring on the outputs of 27 core models across four key dimensions (**I**llusion **F**idelity (**IF**), Overall **V**isual **Q**uality (**VQ**), **S**cene **C**onsistency (**SC**),

Dataset	Size	Input	Output	Task
Hidden-picture optical illusion				
IllusionBench (Hemmat et al., 2024)	32.5K images	image, question	answer	abstract shape recognition
IllusoryVQA (Rostamkhani et al., 2025)	27.3K images	image, question	answer	hidden target classification and OCR
HC-Bench (Li et al., 2025)	112 images	image, question	answer	hidden target recognition
HiP-Bench (ours)	512 samples, 17.9K images	text (+image)	image, human scores	generation
Other types of visual illusion				
GVIL (Zhang et al., 2023b)	100 images, 1.6K instances	image, question	answer, bbox	illusion recognition and language grounding
IllusionVQA (Shahgir et al., 2024)	1.4K images, 1.4K QA pairs	image, question	answer	illusion comprehension and soft localization
HallusionBench (Guan et al., 2024)	346 images, 1.1K QA pairs	image, question	binary answer	image-context reasoning
IllusionBench (Zhang et al., 2025)	1K images, 5.5K QA pairs	image, question	answer, description	illusion perception and description
VI-Probe (Sun et al., 2026)	28.7K images (from 27 cases)	image, question	binary answer	illusion VQA

Table 1: Comparison of recent visual illusion datasets. Although existing benchmarks span diverse visual illusion categories and tasks, those dedicated to hidden-picture optical illusions have so far been limited to understanding-oriented tasks, including recognition, classification, and OCR. HiP-Bench is the first benchmark dedicated to evaluating hidden-picture optical illusion generation.

and **I**llusion **S**hape **C**onsistency (**ISC**)). We detail the analysis of HiP-Bench through five RQs: *(i) generative paradigm analysis; (ii) intra-paradigm variations; (iii) the impact of textures and hidden targets; (iv) error analysis; (v) the alignment of automatic metrics.* Furthermore, because existing evaluation models perform poorly on illusion-specific dimensions, we explore **M**ulti-Scale **D**istance **S**imulation (**MDS**). This baseline, based on multimodal large language models (MLLMs), utilizes multi-scale visual inputs to account for viewing distance in the automated assessment of IF and ISC.

Our key contributions are: (1) We introduce HiP-Bench, the first benchmark to evaluate hidden-picture optical illusion generation, a novel task uniquely challenging current generative models. (2) We establish multiple assessment dimensions, introducing a reference-free evaluation of Illusion Fidelity that aligns with human cognition. (3) We propose an automated evaluation baseline to address the failure of existing metrics on this task. (4) Our experimental findings are as follows: (i) No paradigm or model fully solves this task, exhibiting distinct trade-offs and failure modes. (ii) High-frequency and directional textures may aid concealment, whereas low-frequency backgrounds tend to expose hidden targets and structured patterns may introduce visual artifacts. (iii) Abstract shapes appear to be harder to conceal than physical entities; directional textures with physical entities show the highest observed overall results. (iv) 8 observed types of errors reveal the limitations of current models. (v) Conventional metrics and MLLM evaluators correlate poorly with human judgments.

2 Related Works

We compare HiP-Bench with recent visual illusion datasets in Table 1, to highlight the distinctions between existing studies and ours. See Appendix B for detailed related work.

3 HiP-Bench: Benchmark Design

3.1 Overview of HiP-Bench

We propose HiP-Bench to evaluate visual generative models' ability to synthesize hidden-picture optical illusions. Specifically, we (1) curate background–hidden target combinations and prepare inputs for three generative paradigms, (2) generate illusions using diverse models, and (3) evaluate outputs through human annotation along four dimensions. Thus, HiP-Bench establishes a standardized evaluation framework across diverse generative paradigms grounded in rigorous human judgments. Figure 3 shows a complete HiP-Bench sample.

3.2 Data Curation Process

(1) Taxonomy and Input Preparation. We first construct a taxonomy of background textures and hidden targets (Table 2; See Appendix C.1 for detailed rationale). Next, we collect 32 instances for each of the 4 background and 4 hidden target types, pairing them to produce a balanced dataset of 512 combinations. For these samples, we use GPT-4o (OpenAI, 2024) to generate T2I and Editing prompts. We also synthesize the corresponding background images for the Editing paradigm and extract control images of the hidden targets for the CondGen paradigm. See Appendix D for details on background–hidden target pairing strategy, prompt generation, prompt audit and image preparation.

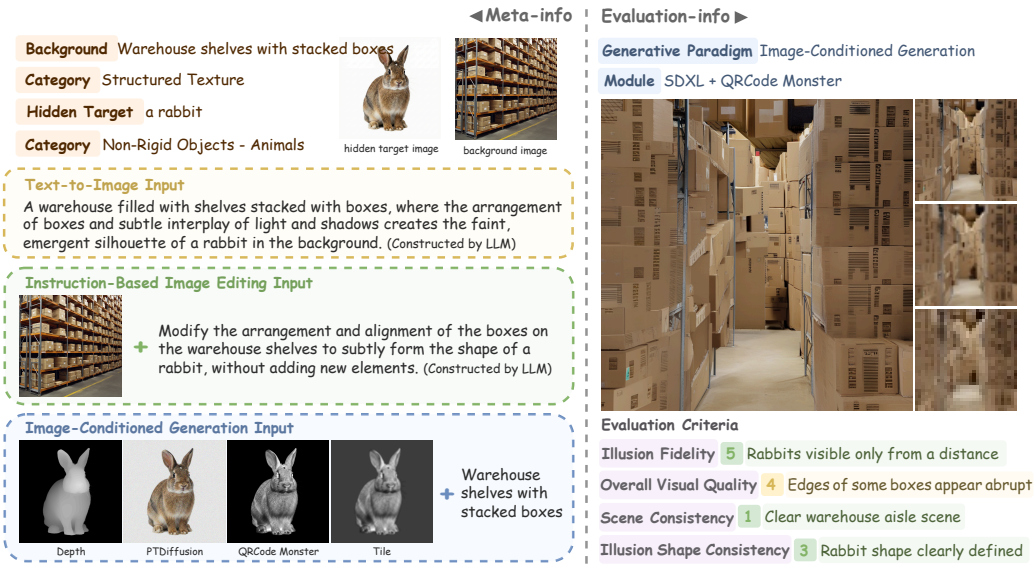

Figure 3: A complete data sample in HiP-Bench, with background–hidden target metadata (top), inputs for three paradigms (middle), and generated images with human scores (bottom).

Attribute	Description	Value	Specific Values
Background Type	Categories of background textures	4	High-Frequency, Structured, Directional, Low-Frequency
Hidden Target Type	Broad categories of hidden targets	4	Textual Elements, Geometry & Shapes, Rigid Objects, Non-Rigid Objects
Target Sub-category	Specific classes of hidden targets	12	Character, Short Word, Long Word, Non-Latin Character, Basic Shape, Complex Shape & Icon, Vehicle, Architectural Structure, Rigid Appliance, Animal, Plant, Human Body
Input Format	Formats of the prepared inputs	3	Prompt, Background Image + Editing Instruction, Control Image + Background Description

Table 2: Overview of the metadata attributes. See Appendix C.2 for visual examples.

(2) Illusion Image Generation. We synthesize illusion images across three generative paradigms using 35 representative models: 22 for T2I, 5 for Editing, and 8 for CondGen. (i) T2I models span diverse architectures, including U-Nets (e.g., SD 1.5 (Rombach et al., 2022)), Diffusion Transformers (e.g., FLUX.1 [dev] (Black Forest Labs, 2024)), and multimodal frameworks (e.g., Janus Pro (Chen et al., 2025b)). (ii) For Editing, we select diffusion models fine-tuned for editing (e.g., InstructPix2Pix (Brooks et al., 2023)) and multimodal frameworks (e.g., OmniGen2 (Wu et al., 2026)). (iii) For CondGen, we evaluate representative spatial control mechanisms (e.g., depth, local detail refinement, and pattern control (Zhang et al., 2023a)) and specialized illusion synthesis methods (e.g., PTDiffusion (Gao et al., 2025)). We apply these models to the 512 samples and obtain a total of 17,920 images. Model lists, repositories, tuning details, and parameters are provided in Appendix F.

(3) Human Annotation. We select 13,824 synthetic images (27 models $\times$ 512 images per model) for human evaluation. The selected models include all Editing and CondGen models. For T2I, we select all Stable Diffusion variants and representative models from other series. We recruit eight computer science graduate students and four annotators from non-computing disciplines. Table 3 details the evaluation dimensions and inter-annotator agreement. Each dimension is independently

Dimension	Specific Elements	Used By (related-work)	Spearman
Illusion Fidelity (IF, 1–5)	concealment quality, distance-dependent perceptibility, target re-identification.	-	0.73
Overall Visual Quality (VQ, 1–5)	Structural integrity, texture naturalness, absence of deliberate layout artifacts.	(Lee et al., 2023)	0.71
Scene Consistency (SC, 0/1)	Semantic alignment between generated background scenes and user inputs.	(Saharia et al., 2022)	0.81
Illusion Shape Consistency (ISC, 1–3)	Semantic correctness and structural accuracy of the hidden targets.	(Hu et al., 2023)	0.78

Table 3: Dimensions used in this paper (see detailed definitions and scoring criteria in Appendix F.1). "-" means no prior papers used it, which shows that Illusion Fidelity is first considered. The Spearman value (0–1) (detailed in Appendix F.2) represents the agreement among human annotations. The relatively strong positive correlations of 0.71–0.81 suggest that the definitions and scoring guidelines are well understood by annotators.

assessed by two randomly assigned annotators, and the final score is obtained by averaging their ratings (annotation details including the interface, workflow, and annotator workload allocation are provided in Appendix G.1). We estimate the annotation cost at roughly 3,600 USD, based on 12 annotators working 20 hours each at 15 USD/hour.

3.3 Dataset Statistics

HiP-Bench contains 512 inputs per paradigm and 17,920 generated images, of which 13,824 are evaluated by humans. Appendix E provides detailed statistics (Table 11), the distribution of hidden targets (Figure 17), and their combinations with background images (Figure 18).

3.4 Automated Evaluation

While human annotation provides a reliable reference, it is costly at scale. We thus explore MLLMs as automated evaluators by prompting them with generated images and rubrics (see Appendix H). To measure alignment with human scores, Schober et al. (2018) suggests that a correlation coefficient (SRCC or PLCC; See Appendix G.2) approaching 0.70 indicates a strong positive correlation. Initial evaluations show that for conventional dimensions (VQ: $\sim$0.68, SC: $\sim$0.69), existing MLLMs approach this level. However, on illusion-specific dimensions (IF and ISC), their correlations drop to a weak range. This indicates that MLLMs generally underperform on illusion-specific dimensions.

To address this limitation, we propose Multi-Scale Distance Simulation (MDS), an automated evaluation baseline inspired by human perception of hidden-picture optical illusions. When viewing such illusions, humans often increase their viewing distance to suppress local texture details and perceive global structures. MDS simulates this scale change through a two-turn prompting strategy (Figure 4). In the first turn, the MLLM evaluates the original image using the scoring rubrics. In the second turn, the MLLM is additionally provided with three downsampled versions of the same image, and is prompted to refine its initial judgment by jointly considering local details from the original image and global structures revealed in the downsampled views. Detailed settings are provided in Appendix H.4.

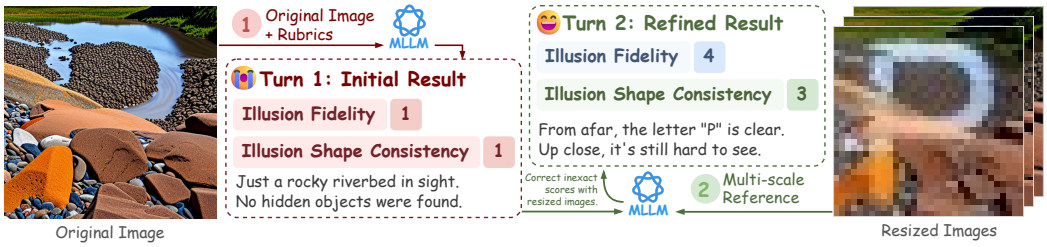

Figure 4: Overview of the MDS strategy. Its two-turn evaluation first judges the original image, then refines the score using three downsampled views that suppress local texture details.

3.5 Merits of HiP-Bench

As a first study, HiP-Bench evaluates visual generative models on hidden-picture optical illusion synthesis. Its advantages are as follows: (1) **Quality control.** We employ expert annotators to conduct human evaluations based on multidimensional rubrics, establishing a reliable human benchmark for assessing hidden-picture optical illusion generation. (2) **Data diversity.** HiP-Bench includes diverse background textures and hidden targets, and provides specific inputs for three generative paradigms. (3) **Model agnosticism.** We assess images generated by a wide range of models, demonstrating HiP-Bench's value and adaptability. (4) **Automated evaluation.** We reveal the limitations of existing automated metrics on illusion-specific dimensions and propose an MLLM-based evaluation strategy. (5) **Application value.** HiP-Bench can support downstream applications, including human-specific CAPTCHAs, marketing and advertising, fashion design, digital entertainment, and cognitive training, while serving as a challenging testbed for visual generative models.

4 Insightful Analysis and Findings

To assess HiP-Bench's effectiveness, we address five research questions (**RQs**): (1) How do different paradigms perform in hidden-picture optical illusion generation? (2) How does performance vary across models within each paradigm? (3) How do background textures and hidden target categories affect generated illusion quality? (4) What failure modes arise in hidden-picture optical illusion generation? (5) How well do automated metrics align with human judgments?

RQ1: Generative Paradigm Analysis. Figure 5 illustrates the corresponding qualitative patterns: Alignment without Hiding, Structure over Quality, Instruction Ignoring, and Destructive Pasting. Table 4 shows that each paradigm exhibits distinct strengths and trade-offs.

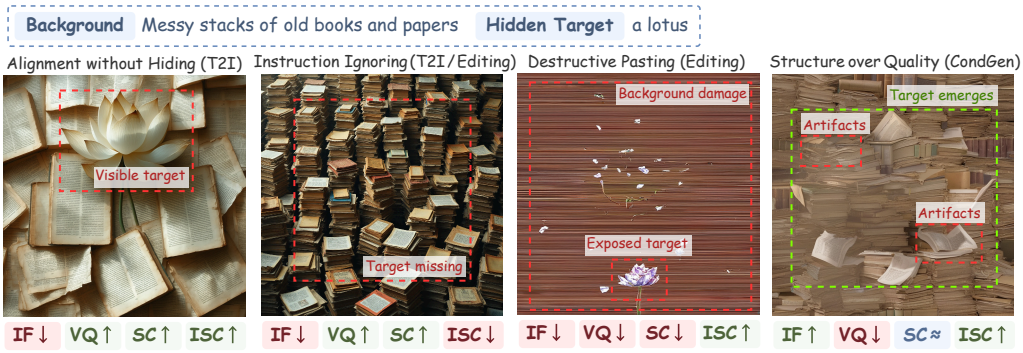

Figure 5: Representative outputs for the same background–hidden target pair across three generation paradigms. In-image annotations highlight paradigm-specific visual behaviors and trade-offs.

(i) *Instruction Ignoring or Alignment without Hiding (T2I).* Models in this paradigm achieve high VQ (4.29) and SC (0.97) but low IF (1.47). This suggests that T2I models either ignore the instruction to hide the target in the background or treat the task as ordinary object rendering rather than illusion generation. A likely reason is that standard T2I training encourages prompted concepts to appear clearly, whereas this task requires the target to be present yet visually hidden.

(ii) *Instruction Ignoring or Destructive Pasting (Editing).* Despite using a background image, this paradigm yields the lowest SC (0.84) but the highest VQ (4.29). Since editing instructions are often ignored, introducing the hidden target requires higher instruction weights, which may overwrite the background. Unlike conventional editing tasks involving localized, visible changes, this task requires subtle texture-level modifications across the background to embed the hidden target instead of inserting a separate foreground object.

(iii) *Structure over Quality (CondGen).* Spatial guidance improves hidden-target structure formation, with high IF (2.03) and ISC (2.41) but the lowest VQ (3.40). This is because current conditioning methods can specify the target's spatial layout, but provide limited guidance on how to encode that structure through coherent background textures. This often leads to unnatural artifacts.

Overall, *no paradigm fully resolves the task.* The highest IF among all models is only 2.42/5.0 (SDXL + QR Code Monster), highlighting the inherent difficulty in generating such illusions.

Paradigm / Model	IF	VQ	SC	ISC	Paradigm / Model	IF	VQ	SC	ISC
Text-to-image	1.47	4.29	0.97	1.86	**Instruction-based image editing**	1.30	4.29	0.84	1.56
Architecture Averages					CosXL Edit	1.22	**4.70**	0.96	1.42
U-Net	1.27	**4.49**	0.94	1.51	FLUX.1-Kontext	1.26	4.52	**0.99**	1.49
Unified Multimodal	1.43	4.03	0.95	1.78	OmniGen2	**1.58**	3.96	0.85	**2.09**
Diffusion Transformer	**1.58**	4.36	**0.99**	**2.05**	**Image-conditioned generation**	2.03	3.40	0.87	2.41
Model Scaling (SD 3.5 Family)					SD 1.5 + Depth	1.91	3.24	0.89	**2.61**
SD 3.5 Medium	1.58	**4.31**	**0.99**	2.01	SD 1.5 + QR Code Monster V2	1.76	**3.90**	0.91	1.97
SD 3.5 Large Turbo	1.68	4.19	**0.99**	2.25	SD 1.5 + PTDiffusion (training free)	2.16	3.34	0.85	2.04
SD 3.5 Large	**1.71**	**4.31**	**0.99**	**2.31**	SDXL + QR Code Monster	**2.42**	3.80	**0.98**	2.29

Table 4: Performance at the paradigm and model levels, including paradigm averages and scores for representative models. See Appendix I.1 for full results and Appendix I.2 for statistical tests.

RQ2: Intra-paradigm Variations. Table 4 details intra-paradigm comparisons. We then examine how performance varies with model architecture, editing behavior, model capacity, and control mechanism.

(i) *Diffusion Transformers better balance concealment and visual quality than U-Nets.* U-Nets maintain high visual quality (VQ: 4.49) but often omit the target or render it overtly, resulting in the lowest IF (1.27) and ISC (1.51). In contrast, Diffusion Transformers achieve higher concealment (IF: 1.58) with a modest reduction in visual quality (VQ: 4.36). Meanwhile, unified multimodal models struggle with this balance; although they achieve better concealment than U-Nets, their visual quality drops to the lowest (VQ: 4.03).

(ii) *Editing models tend to under-edit or over-edit.* Diffusion-based editors largely preserve the background, often failing to introduce the hidden target (average IF: 1.22). By contrast, multimodal editors (OmniGen2) better follow the instruction, but tend to damage the background when introducing the target (VQ: 3.96). This suggests that successful editing for this task has a narrow operating range: the model must modify the background enough to encode the target while preserving it enough to avoid a pasted appearance.

(iii) *QR Code Monster outperforms other control mechanisms.* In CondGen, different control mechanisms show clear trade-offs. ControlNet Depth reliably manifests the hidden shape (average ISC: 2.60) but severely degrades visual quality (average VQ: 3.27). PTDiffusion improves concealment (IF: 2.16) but still suffers from low visual quality (VQ: 3.34). In contrast, QR Code Monster with SDXL achieves the best balance, forming the target structure while preserving background semantics and texture naturalness.

(iv) *Stronger foundation models improve illusion synthesis.* In T2I, scaling SD 3.5 from Medium to Large improves concealment (IF from 1.58 to 1.71), suggesting increased model capacity improves prompt adherence. The Large Turbo model remains competitive, offering an efficient alternative. Similarly, in CondGen: upgrading SD 1.5 to SDXL with their respective QR Code Monster adapters raises IF from 1.76 to 2.42. Thus, stronger models benefit concealment, although scaling alone remains insufficient to solve the task.

RQ3: Attribute Influence Analysis. Figure 6 presents the average evaluation results across 16 background–hidden target combinations, from which several patterns can be observed.

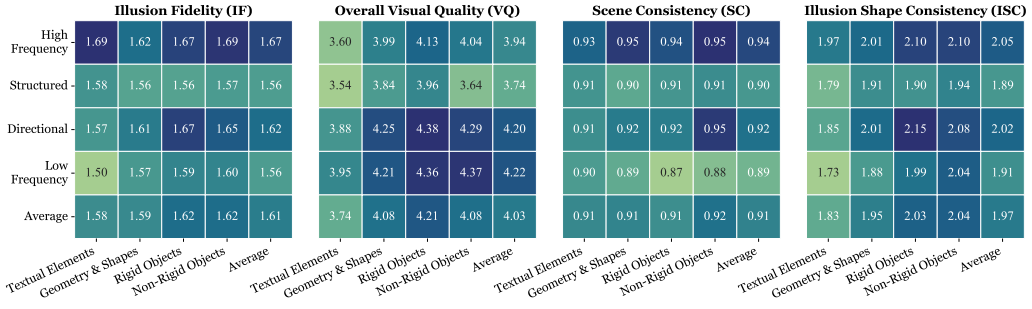

Figure 6: Overall performance across 16 background texture–hidden target category combinations, averaged across all evaluated models. See Appendix I.2 for statistical tests and Appendix I.3 for paradigm-specific results and analysis.

(i) High-Frequency textures achieve the highest IF (1.67) but lower VQ (3.94), as models generate chaotic details to hide targets. Conversely, Directional textures yield comparable IF (1.62) but higher VQ (4.20), since directional patterns help hide target boundaries while preserving visual harmony. Thus, *High-Frequency textures aid concealment, while Directional textures better balance it with image quality.*

(ii) Low-Frequency and Structured textures tend to yield lower IF, but fail differently. Low-Frequency textures maintain the highest VQ (4.22), indicating clean scene rendering but insufficient detail for concealment. In contrast, Structured textures yield the lowest VQ (3.74), as embedding

hidden targets into regular patterns disrupts structural integrity and produces artifacts. Therefore, *Low-Frequency textures expose hidden targets, while Structured textures introduce artifacts.*

(iii) Textual Elements and Geometry & Shapes tend to perform worse. Their exact outlines and lack of internal textures may make slight distortions harder to accommodate, sometimes resulting in pasted overlays. Conversely, the natural textures and flexible boundaries of Rigid and Non-Rigid Objects may help models deform or occlude them so they blend into the background. Overall, *abstract shapes appear to be harder to hide than physical objects*.

(iv) Directional textures paired with Rigid or Non-Rigid Objects appear to be relatively favorable combinations, whereas Structured textures paired with Textual Elements or Geometry & Shapes may be less favorable. This suggests that *some attribute choices may be more conducive to generating effective illusions than others.*

RQ4: Failure Mode Analysis. We analyze low-scoring generated images and identify eight primary error types that reveal the limitations of current models in Figure 7. If an image contains multiple errors, we assign it to the most severe category (see Appendix I.4 for detailed rules).



Figure 7: Case studies of errors observed in HiP-Bench (see more in Appendix A.2). We present one example for each of the eight major error types, sorted in descending order of severity from Type 1 (indicating the most serious errors) to Type 8 (indicating the least serious errors).

Type 1: Severe Structural Corruption. The image contains unrecognizable content or severe geometric distortion, e.g., distorted suspension bridge cables and bridge supports (Case 1). **Type 2: Scene Semantic Mismatch.** Generated background conflicts with the text prompt, e.g., failing to depict "ramen with green onions" and showing green texture instead (Case 2). **Type 3: Concept Omission.** Model ignores instruction, yielding an image without hidden objects, e.g., failing to generate the "square shape" within the mountain scene (Case 3). **Type 4: Illusion Structural Deviation.** The hidden target is generated with an inaccurate or distorted structure, e.g., the character deviating from the target outline (Case 4). **Type 5: Blatant Pasting Effect.** The hidden target lacks visual blending, appearing as a pasted foreground element, e.g., the anchor directly overlaid as a silhouette on the background (Case 5). **Type 6: Independent Object Manifestation.** The model renders the hidden object as an independent physical entity rather than integrating it into the background, e.g., a standalone maple leaf placed on the LEGO bricks (Case 6). **Type 7: Traceable Manipulation.** The background exhibits artificial traces arranged to form the target outline, e.g., unnatural shadows along the arches and cathedral columns (Case 7). **Type 8: Insufficient Concealment.** The hidden target's outline remains too visible, reducing illusion fidelity, e.g., the distinct cat outline and visible "cat" lettering in the latte art.

We analyze the distribution of failure modes for each paradigm in Table 5: (i) T2I and Editing paradigms frequently exhibit Concept Omission, largely due to insufficient explicit spatial control; (ii) The CondGen paradigm reduces omission errors but is highly prone to the Blatant Pasting Effect, as strict spatial constraints hinder visual blending; (iii) The T2I paradigm often exhibits Independent Object Manifestation (24.28%), likely because models fail to understand how the hidden target should be integrated into the background and instead render it as a standalone object.

Failure Mode	T2I	Editing	CondGen	Failure Mode	T2I	Editing	CondGen
(1) Severe Structural Corruption	1.59	4.45	3.15	(5) Blatant Pasting Effect	12.51	9.46	**30.82**
(2) Scene Semantic Mismatch	3.06	12.54	11.97	(6) Independent Object Manifestation	24.28	9.75	13.15
(3) Concept Omission	**52.36**	**60.88**	24.10	(7) Traceable Manipulation	0.33	0.04	4.03
(4) Illusion Structural Deviation	4.24	2.66	6.93	(8) Insufficient Concealment	1.62	0.21	5.84

Table 5: Distribution (%) of failure modes across three generative paradigms. Each low-scoring image is assigned to its most severe failure mode. See Tables 25, 26, and 27 in Appendix I.5 for model-level results and analysis.

Metric / Model	IF	VQ	SC	ISC	Metric / Model	IF	VQ	SC	ISC
(i) Conventional visual metrics					(ii) Direct MLLM-as-a-Judge evaluators				
ΔVQAScore	**27.3 / 24.3**	-	-	-	Qwen3-VL-8B	-23.9 / -25.9	31.4 / 30.3	66.1 / 64.1	15.0 / 15.3
NIQE	-	**45.3 / 42.8**	-	-	GPT-5.4	3.2 / -0.4	60.5 / 55.7	64.2 / 63.7	39.0 / 40.9
MUSIQ	-	19.3 / 18.5	-	-	Gemini-3-Pro	**39.9 / 37.5**	**68.4 / 67.0**	**69.3 / 69.9**	**48.9 / 50.3**
LAION-Aes	-	43.5 / 37.3	-	-	(iii) MLLMs with MDS (Ours)				
CLIP	-	-	66.7 / **65.5**	11.1 / 7.2	Qwen3-VL-8B	6.9 / 7.1	-	-	30.7 / 34.8
BLIP	-	-	58.9 / 60.0	**21.7** / 13.6	GPT-5.4	32.5 / 31.1	-	-	49.4 / 51.0
VQAScore	-	-	**67.5** / 64.4	20.3 / **21.4**	Gemini-3-Pro	**53.8 / 53.1**	-	-	**63.2 / 62.5**

Table 6: Correlation (PLCC / SRCC, % ↑) between automated scores and human judgments. "-" indicates that the metric is not applicable. Bold indicates the best result within each group, and underlining indicates the best result overall.

RQ5: Automated evaluation alignment. To justify our human annotation campaign and evaluate the effectiveness of our proposed automated evaluation strategy, we evaluate three groups of methods on HiP-Bench: (i) conventional visual metrics (ΔVQAScore (Appendix H.1), NIQE (Mittal et al., 2013), MUSIQ (Ke et al., 2021), LAION-Aes (Schuhmann et al., 2022), CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and VQAScore (Lin et al., 2025)), (ii) direct MLLM-as-a-Judge (Qwen3-VL-8B (Yang et al., 2025), GPT-5.4 (OpenAI, 2026), and Gemini-3-Pro (Google DeepMind, 2026)), and (iii) these MLLMs equipped with MDS. For each dimension, we compute the Pearson Linear (PLCC) and Spearman Rank (SRCC) correlation coefficients (detailed in Appendix G.2) between automated scores and human judgments.

Table 6 shows that: (i) conventional visual metrics perform adequately on explicit dimensions (VQ and SC) but struggle with illusion-specific dimensions (IF and ISC); (ii) direct MLLM-as-a-Judge evaluators improve overall correlation but still yield low alignment on IF and ISC; (iii) MDS consistently improves MLLM correlations on both IF and ISC, offering a baseline for automated illusion evaluation (see Appendix I.6 for the ablation study supporting the MDS design, and Appendix J for future use of HiP-Bench)

5 Conclusions

We present HiP-Bench, the first benchmark for hidden-picture optical illusion generation. Our results show that current models struggle with Illusion Fidelity, often exposing, omitting, or crudely pasting hidden objects rather than blending them into backgrounds. With a standardized benchmark, an automated baseline, and analysis, HiP-Bench lays a foundation for hidden-picture optical illusion synthesis, with potential applications such as human-specific CAPTCHAs, marketing and advertising, aesthetic design in fashion, digital entertainmen, and cognitive training. It also serves as a stringent testbed for visual generative models.

6 Limitations and Future Work

First, MDS better aligns with human judgments but cannot replace human evaluation. *More reliable illusion-specific metrics are needed for evaluation and optimization at scale.* Second, evaluating hidden-picture illusions is perceptually demanding. Calibration rounds, a standardized annotation setup, and inter-annotator consistency checks improve reliability, but 12 annotators cannot represent the general population. *Future studies should use a larger and more diverse annotator pool.* Third, HiP-Bench covers diverse backgrounds and hidden targets but lacks real-world illusions. *We have collected such images and are preparing to test the real-world generalization of automated metrics.*

AI USE STATEMENT

In this work, we used generative AI tools for the construction of the HiP-Bench dataset (including prompt generation and synthetic image generation) and for automated evaluation experiments. The generated prompts were manually reviewed for appropriateness, and the generated images were subject to human review and quality control. The automated evaluation results were validated against human judgments. We did not use generative AI tools for any other tasks requiring disclosure. We reviewed all AI-assisted work and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support the reproducibility of image generation experiments, we document the complete construction and input-preparation process of HiP-Bench in Appendix D. For each evaluated model, we provide its implementation source, inference configuration, hyperparameter settings, random seeds, and computational environment, including the relevant hardware information, in Appendix F. These details are intended to enable the reproduction of the generated images and the corresponding model-level and paradigm-level results.

To facilitate the reproduction of our human evaluation, we provide the complete annotation protocol in Appendix G. This includes the definitions and scoring criteria for all evaluation dimensions, annotator instructions, the annotation interface and image presentation settings, viewing-distance requirements, score aggregation procedures, and inter-annotator agreement analyses. These materials specify both the evaluation environment and the procedures used to obtain the reported human judgments.

For automated evaluation, Appendix H documents the evaluated metrics and models, evaluation prompts and rubrics, input-processing procedures, implementation details, and the configuration of the proposed MDS strategy. We additionally report complete experimental results, statistical tests, correlation calculations, and extended analyses in Appendix I, together with practical recommendations for using HiP-Bench in Appendix J.

Upon acceptance, we will publicly release the HiP-Bench dataset, including its metadata, prepared inputs, generated images, and human annotations. We will also release the code and evaluation materials required for dataset construction, image generation, human-score processing, automated evaluation, MDS, and the reproduction of the principal analyses and results reported in this paper.

REFERENCES

Black Forest Labs. Announcing Black Forest Labs. https://bfl.ai/blog/24-08-01-bfl, 2024. URL https://bfl.ai/blog/24-08-01-bfl. Accessed: 2026-07-27.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *CoRR*, abs/2506.15742, 2025. doi: 10.48550/ARXIV.2506.15742. URL https://doi.org/10.48550/arXiv.2506.15742.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023. doi: 10.1109/CVPR52729.2023.01764.

Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion Illusions: Hiding Images in Plain Sight. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. ISBN 9798400705250. doi: 10.1145/3641519.3657500. URL https://doi.org/10.1145/3641519.3657500.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller,

Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/chen24h.html.

Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models. *CoRR*, abs/2401.05252, 2024b. URL https://doi.org/10.48550/arXiv.2401.05252.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=eAKmQPe3m1.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PIXART-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 74–91, 2025a. ISBN 978-3-031-73411-3. doi: 10.1007/978-3-031-73411-3_5.

Qianpu Chen and Derya Soydaner. Now You See the Hate: Adaptive View Retrieval for Hidden Hateful Illusions, 2026. URL https://arxiv.org/abs/2607.19061.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, 2025b. URL https://arxiv.org/abs/2501.17811.

Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J. Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4), July 2010. ISSN 0730-0301. doi: 10.1145/1778765.1778788. URL https://doi.org/10.1145/1778765.1778788.

DeepFloyd Lab. DeepFloyd IF, 2023. URL https://github.com/deep-floyd/IF.

Ziqi Ding, Gelei Deng, Yi Liu, Junchen Ding, Jieshan Chen, Yulei Sui, and Yuekang Li. IllusionCAPTCHA: A CAPTCHA based on Visual Illusion. In *Proceedings of the ACM on Web Conference 2025*, pp. 3683–3691, 2025. ISBN 9798400712746. doi: 10.1145/3696410.3714726. URL https://doi.org/10.1145/3696410.3714726.

Werner Ehm and Jiří Wackermann. Modeling geometric–optical illusions: A variational approach. *Journal of Mathematical Psychology*, 56(6):404–416, 2012. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2012.12.001. URL https://www.sciencedirect.com/science/article/pii/S002224961200123X.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-Measure: A New Way to Evaluate Foreground Maps. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4558–4567, 2017. doi: 10.1109/ICCV.2017.487.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018. ISBN 9780999241127. doi: 10.24963/ijcai.2018/97.

William T. Freeman, Edward H. Adelson, and David J. Heeger. Motion without movement. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '91, pp. 27–30, New York, NY, USA, 1991. Association for Computing Machinery. ISBN 0897914368. doi: 10.1145/122718.122721. URL https://doi.org/10.1145/122718.122721.

Xiang Gao, Shuai Yang, and Jiaying Liu. PTDiffusion: Free Lunch for Generating Optical Illusion Hidden Pictures with Phase-Transferred Diffusion Model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18240–18249, June 2025. doi: 10.1109/CVPR52734.2025.01700.

Daniel Geng, Inbum Park, and Andrew Owens. Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24154–24163, 2024. doi: 10.1109/CVPR52733.2024.02280.

Alexandra Gomez-Villa, Kai Wang, C. Alejandro Parraga, Bartłomiej Twardowski, Jesus Malo, Javier Vazquez-Corral, and Joost Van De Weijer. The Art of Deception: Color Visual Illusions and Diffusion Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18642–18652, 2025. doi: 10.1109/CVPR52734.2025.01737.

Google. Nano Banana 2 - Gemini AI image generator & photo editor, 2026. URL https://gemini.google/overview/image-generation/.

Google DeepMind. Gemini 3 — Google DeepMind. https://deepmind.google/models/gemini/, 2026. URL https://deepmind.google/models/gemini/. Accessed: 2026-05-10; page subsequently updated.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024. doi: 10.1109/CVPR52733.2024.01363.

Arshia Hemmat, Adam Davies, Tom A. Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in Plain Sight: Evaluating Abstract Shape Recognition in Vision-Language Models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 88527–88556, 2024. doi: 10.52202/079017-2808. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a13ff984831deea39e6132bafdfdd6d5-Paper-Datasets_and_Benchmarks_Track.pdf.

Elad Hirsch and Ayellet Tal. Color Visual Illusions: A Statistics-based Computational Model. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9447–9458. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b39183e7053a0106e4376f4e9c5c74d-Paper.pdf.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20349–20360, 2023. doi: 10.1109/ICCV51070.2023.01866.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137, 2021. doi: 10.1109/ICCV48922.2021.00510.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy S Liang. Holistic Evaluation of Text-to-Image Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 69981–70011, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dd83eada2c3c74db3c7fe1c087513756-Paper-Datasets_and_Benchmarks.pdf.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier Painting Than Thinking: Can Text-to-Image Models Set the Stage, but Not Direct the Play? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iqAFhWistW.

Sifan Li, Yujun Cai, and Yiwei Wang. SemVink: Advancing VLMs' Semantic Understanding of Optical Illusions via Visual Global Thinking. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27167–27177, November 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1381. URL https://aclanthology.org/2025.emnlp-main.1381/.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *CoRR*, abs/2405.08748, 2024. doi: 10.48550/ARXIV.2405.08748. URL https://doi.org/10.48550/arXiv.2405.08748.

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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 366–384, 2025. ISBN 978-3-031-72673-6. doi: 10.1007/978-3-031-72673-6_20.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, December 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751, June 2025. doi: 10.1109/CVPR52734.2025.00725.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to Evaluate Foreground Maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014. doi: 10.1109/CVPR.2014.39.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. FreeControl: Training-Free Spatial Control of Any Text-to-Image Diffusion Model with Any Condition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7465–7475, 2024. doi: 10.1109/CVPR52733.2024.00713.

Monster labs. Controlnet QR Code Monster v2 For SD-1.5, 2023a. URL https://huggingface.co/monster-labs/control_v1p_sd15_qrcode_monster.

Monster labs. Controlnet QR Code Monster v1 For SDXL, 2023b. URL https://huggingface.co/monster-labs/control_v1p_sdxl_qrcode_monster.

Aude Oliva, Antonio Torralba, and Philippe G. Schyns. Hybrid images. *ACM Trans. Graph.*, 25 (3):527–532, July 2006. ISSN 0730-0301. doi: 10.1145/1141911.1141919. URL https://doi.org/10.1145/1141911.1141919.

OpenAI. GPT-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026. URL https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-05-10.

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

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. Lumina-Image 2.0: a Unified and Efficient Image Generative Framework. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20031–20042, 2025. doi: 10.1109/ICCV51701.2025.01863.

Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly Accurate Dichotomous Image Segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 38–56, 2022. ISBN 978-3-031-19797-0. doi: 10.1007/978-3-031-19797-0_3.

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 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 18–24 Jul 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, June 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. Illusory VQA: Benchmarking and Enhancing Multimodal Models on Visual Illusions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2986–2995, 2025. doi: 10.1109/CVPRW67362.2025.00282.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial Diffusion Distillation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 87–103, 2025. ISBN 978-3-031-73016-0. doi: 10.1007/978-3-031-73016-0_6.

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia and analgesia*, 126(5):1763–1768, May 2018. ISSN 0003-2999. doi: 10.1213/ane.0000000000002864. URL https://journals.lww.com/anesthesia-analgesia/Fulltext/2018/05000/Correlation_Coefficients__Appropriate_Use_and.50.aspx.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk,

and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. IllusionVQA: A Challenging Optical Illusion Dataset for Vision Language Models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=7ysaJGs7zY.

Shakker Labs. FLUX.1-dev-ControlNet-Union-Pro, 2024. URL https://huggingface.co/Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro.

Stability AI. Introducing Stable Diffusion 3.5, 2024a. URL https://stability.ai/news/introducing-stable-diffusion-3-5.

Stability AI. Stable Diffusion 3.5 Large Controlnet - Depth, 2024b. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-large-controlnet-depth.

Stability AI. Cos Stable Diffusion XL 1.0 and Cos Stable Diffusion XL 1.0 Edit, 2024c. URL https://huggingface.co/stabilityai/cosxl.

Qiyang Sun, Xia Wang, Changda Yan, Xin Zhang, and Shiwei Xu. An HVS-derived network for assessing the quality of camouflaged targets with feature fusion. *Neurocomput.*, 617(C), February 2025. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.129016. URL https://doi.org/10.1016/j.neucom.2024.129016.

Xiaoxiao Sun, Mingyang Li, Kun Yuan, Min Woo Sun, Mark Endo, Shengguang Wu, Changlin Li, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Do VLMs Perceive or Recall? Probing Visual Perception vs. Memory with Classic Visual Illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25861–25870, June 2026.

Chen-Yang Wang, Gepeng Ji, Song Shao, Ming-Ming Cheng, and Deng-Ping Fan. Context-measure: Contextualizing Metric for Camouflage, 2026. URL https://arxiv.org/abs/2512.07076.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Towards Instruction-Aligned Multimodal Generation, 2026. URL https://arxiv.org/abs/2506.18871.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified Image Generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304, June 2025. doi: 10.1109/CVPR52734.2025.01241.

xinsir. ControlNet Depth SDXL, support zoe, midias, 2024. URL https://huggingface.co/xinsir/controlnet-depth-sdxl-1.0.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. LLLaVA-Critic: Learning to Evaluate Multimodal Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13618–13628, 2025. doi: 10.1109/CVPR52734.2025.01271.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui

Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024. doi: 10.52202/079017-0688. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26cfdcd8fe6fd75cc53e92963a656c58-Paper-Conference.pdf.

Guanghui Ye, Huan Zhao, Zixing Zhang, Xupeng Zha, and Zhihua Jiang. LSTDial: Enhancing Dialogue Generation via Long- and Short-Term Measurement Feedback. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5857–5871, June 2024. doi: 10.18653/v1/2024.naacl-long.326. URL https://aclanthology.org/2024.naacl-long.326/.

Guanghui Ye, Huan Zhao, Bo Li, Haijiao Chen, Zhixue Zhao, Zhihua Jiang, and Keqin Li. CCDE: A Compact and Competitive Dialogue Evaluation Framework via Knowledge Distillation of Large Language Models. *IEEE Transactions on Computational Social Systems*, 12(6):5127–5139, 2025a. doi: 10.1109/TCSS.2025.3580272.

Guanghui Ye, Huan Zhao, Zixing Zhang, and Zhihua Jiang. UniDE: A multi-level and low-resource framework for automatic dialogue evaluation via LLM-based data augmentation and multitask learning. *Information Processing & Management*, 62(3):104035, 2025b. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2024.104035. URL https://www.sciencedirect.com/science/article/pii/S0306457324003947.

Guanghui Ye, Huan Zhao, Zhixue Zhao, Xupeng Zha, Yang Liu, and Zhihua Jiang. Knowledge Image Matters: Improving Knowledge-Based Visual Reasoning with Multi-Image Large Language Models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21883–21896, July 2025c. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1063. URL https://aclanthology.org/2025.acl-long.1063/.

Guanghui Ye, Huan Zhao, Yingxue Gao, Zhixue Zhao, Kehan Wang, Xupeng Zha, and Zhihua Jiang. Making Visual Dialogue More Engaging: A New Task, Method, and Metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16145–16153, Mar. 2026. doi: 10.1609/aaai.v40i19.38650. URL https://ojs.aaai.org/index.php/AAAI/article/view/38650.

Jerrold H. Zar. Spearman Rank Correlation. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd, 2005. ISBN 9780470011812. doi: 10.1002/0470011815.b2a15150. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a15150.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023a. doi: 10.1109/ICCV51070.2023.00355. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00355.

Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding Visual Illusions in Language: Do Vision-Language Models Perceive Illusions Like Humans? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5718–5728, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.348. URL https://aclanthology.org/2023.emnlp-main.348/.

Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. IllusionBench: A Large-scale and Comprehensive Benchmark for Visual Illusion Understanding

in Vision-Language Models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11208962.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, Xu Luo, Zehan Wang, Kaipeng Zhang, Lirui Zhao, Si Liu, Xiangyu Yue, Wanli Ouyang, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-Next : Making Luminaß-T2X Stronger and Faster with Next-DiT. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 131278–131315, 2024. doi: 10.52202/079017-4172. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ed2dad593d87ca474a636cba610a29d3-Paper-Conference.pdf.

The appendix is structured as follows:

- More Examples in Section A.
  - More Examples of HiP-Bench in Section A.1.
  - More Examples of Error Types in Section A.2.
- Details of Related Work in Section B.
  - Visual Illusions in Section B.1.
  - Visual Generative Models in Section B.2.
  - Automated Evaluation in Section B.3.
- Taxonomy Design in Section C.
  - Taxonomy Rationale in Section C.1.
  - Visual Examples of Taxonomy Categories in Section C.2.
- Dataset Construction and Input Preparation in Section D.
  - Background-Object Pairing Strategy in Section D.1.
  - Prompt Generation for Text-to-Image Models in Section D.2.
  - Data Preparation for Instruction-Based Image Editing Models in Section D.3.
  - Prompt Validity Audit in Section D.4.
  - Control Image Preprocessing for Image-Conditioned Generation Models in Section D.5.
- More Statistics of HiP-Bench in Section E.
- Evaluated Models and Implementation Details in Section F.
  - Model Specifications in Section F.1.
  - Hyperparameter Tuning in Section F.2.
- Human Annotation Details in Section G.
  - Scoring Guidelines and Inter-annotator Agreements in Section G.1.
  - Definitions of SRCC and PLCC in Section G.2.
- Details of Automated Evaluation in Section H.
  - Construction of ΔVQAScore in Section H.1.
  - Prompts for Direct MLLM-as-a-Judge in Section H.2.
  - Prompts for Ablation Variants in Section H.3.
  - Prompts for MLLM with MDS in Section H.4.
- Evaluation Details, Full Results, Statistical Tests, and Extended Analyses in Section I.
  - Full Results in Section I.1.
  - Uncertainty Estimates and Significance Tests in Section I.2.
  - Extended Analysis on Attribute Influence in Section I.3.
  - Failure Mode Assignment Rules in Section I.4.
  - Extended Analysis on Failure Modes in Section I.5.
  - Ablation Study of Automated Evaluation in Section I.6.
- Future Use of HiP-Bench in Section J.

A More Examples

A.1 More Examples of HiP-Bench

We provide additional examples from HiP-Bench in Figures 8 and 9. The examples include four hidden target categories: Textual Elements, Geometry & Shapes, Rigid Objects, and Non-Rigid Objects. The two figures correspond to the first two and the latter two categories, respectively. For each row, we present a background–hidden target pair and corresponding inputs for three generative paradigms: (i) text-to-image, (ii) instruction-based image editing, and (iii) image-conditioned generation, accompanied by output images from distinct models with human-annotated scores on four dimensions.

A.2 More Examples of Error Types

We provide additional examples of the error types defined in RQ4: Failure Mode Analysis. Figures 10 and 11 present more cases for Type 1–4 and Type 5–8 errors, respectively.

B Details of Related Work

B.1 Visual Illusions

Computational studies have addressed geometric–optical illusions (Ehm & Wackermann, 2012), lightness and color illusions (Hirsch & Tal, 2020), motion illusions (Freeman et al., 1991), and hybrid images whose interpretation changes with viewing distance (Oliva et al., 2006).

In this work, we study hidden-picture optical illusions, in which elements of a primary scene collectively form a secondary hidden target. Camouflage images also embed foreground figures in background images, making them difficult to detect while retaining cues for recognition through focused inspection (Chu et al., 2010). In camouflage images, the foreground figure is found by inspecting cues in the background rather than by viewing the image at a reduced scale. By contrast, hidden-picture optical illusions reveal different content at different viewing scales: at close range or when enlarged, the image appears as a primary scene; at a distance or when reduced, it appears as a hidden target.

As summarized in Table 1, although existing benchmarks span diverse visual illusion categories and tasks, those dedicated to hidden-picture optical illusions have so far been limited to understanding-oriented tasks, including recognition, classification, and OCR.

For benchmarks covering other types of visual illusions, GVIL (Zhang et al., 2023b) covers two broad areas, color and geometric illusions, and evaluates Same-Different Question Answering, Referential Question Answering, Attribute Question Answering, and Referential Localization. IllusionVQA (Shahgir et al., 2024) evaluates comprehension across 12 categories, including Impossible Object, Real Scene, Size Illusion, Hidden Illusion, and Color Illusion; its separate Soft-Localization task asks models to distinguish geometrically impossible objects from ordinary geometric objects. HallusionBench (Guan et al., 2024) evaluates image-context reasoning through two visual-question categories, Visual Dependent and Visual Supplement. The former includes an Illusion subcategory, while the latter includes Chart, Table, Map, and OCR. Its controlled question structure supports analyses of response tendencies and logical consistency, as well as the diagnosis of Language Hallucination and Visual Illusion. IllusionBench (Zhang et al., 2025) evaluates illusion perception and description through true-or-false, multiple-choice, and open-ended description tasks over five image types: Classic Cognitive Illusion, Real Scene Illusion, No Illusion, Ishihara Images, and Trap Illusion. VI-Probe (Sun et al., 2026) uses 27 classical optical illusions grouped by their primary controlling factors into size and length, brightness and color, and orientation (parallelism), with controlled visual and linguistic variations used to probe visual perception and memory in VLMs.

For hidden-picture optical illusion understanding, IllusionBench (Hemmat et al., 2024) evaluates whether VLMs can recognize shapes formed by arrangements of visual elements, while IllusoryVQA (Rostamkhani et al., 2025) explores low-pass filtering as a preprocessing strategy for illusion VQA. SemVink (Li et al., 2025) introduces HC-Bench for evaluating hidden texts, objects, and illusions, and reports that resizing images to low resolutions (32–128 pixels) substantially improves

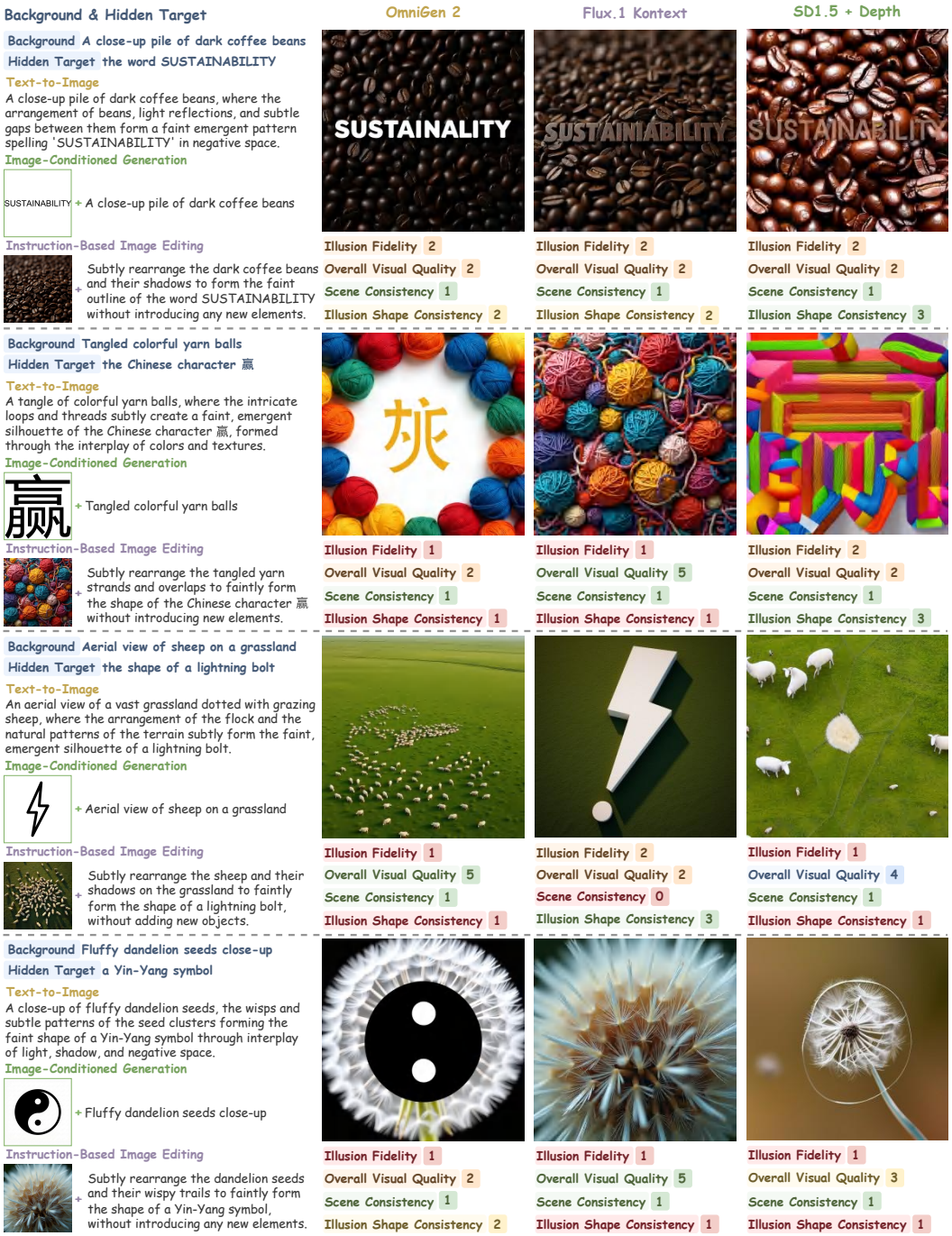

Figure 8: Additional examples from HiP-Bench for Textual Elements and Geometry & Shapes. For each row, we present a background–hidden target pair and corresponding inputs for text-to-image, instruction-based image editing, and image-conditioned generation, accompanied by output images from distinct models with human-annotated scores on four dimensions.

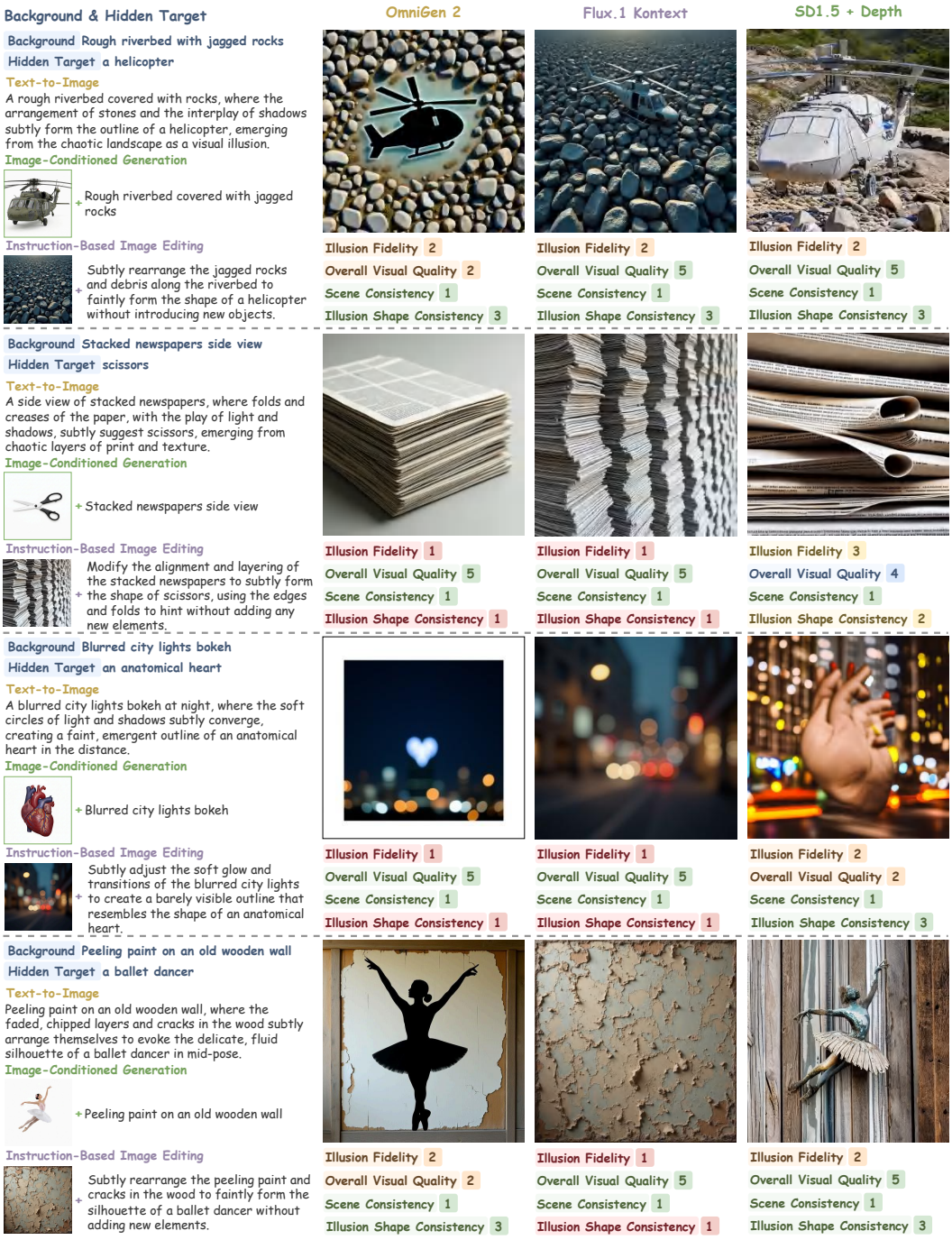

Figure 9: Additional examples from HiP-Bench for Rigid Objects and Non-Rigid Objects. For each row, we present a background–hidden target pair and corresponding inputs for text-to-image, instruction-based image editing, and image-conditioned generation, accompanied by output images from distinct models with human-annotated scores on four dimensions.

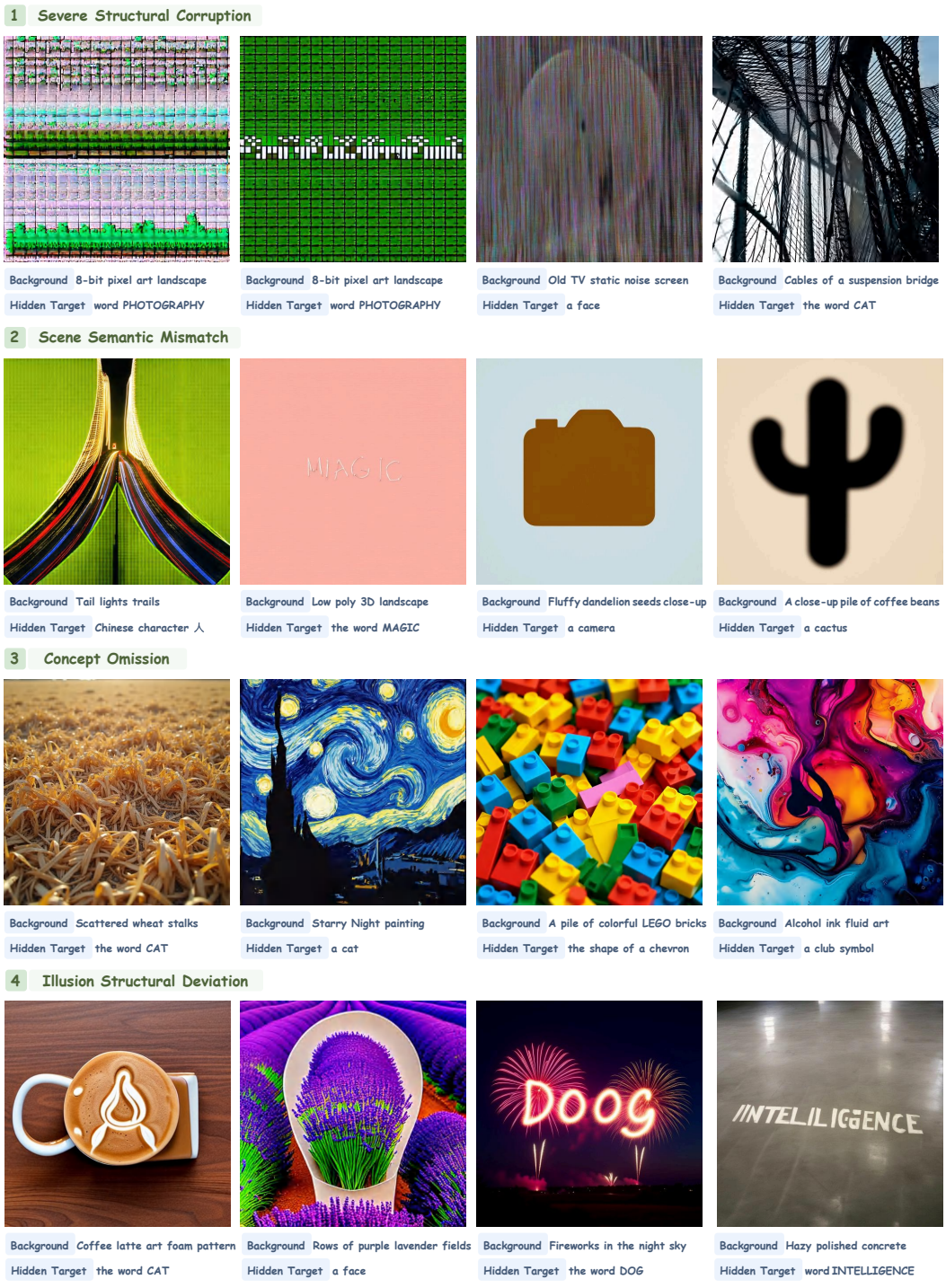

Figure 10: Additional case studies on Type 1–4 errors observed in HiP-Bench, as defined in RQ4: Failure Mode Analysis. Each row presents one error type, with four examples shown across the row to illustrate the corresponding failure mode.

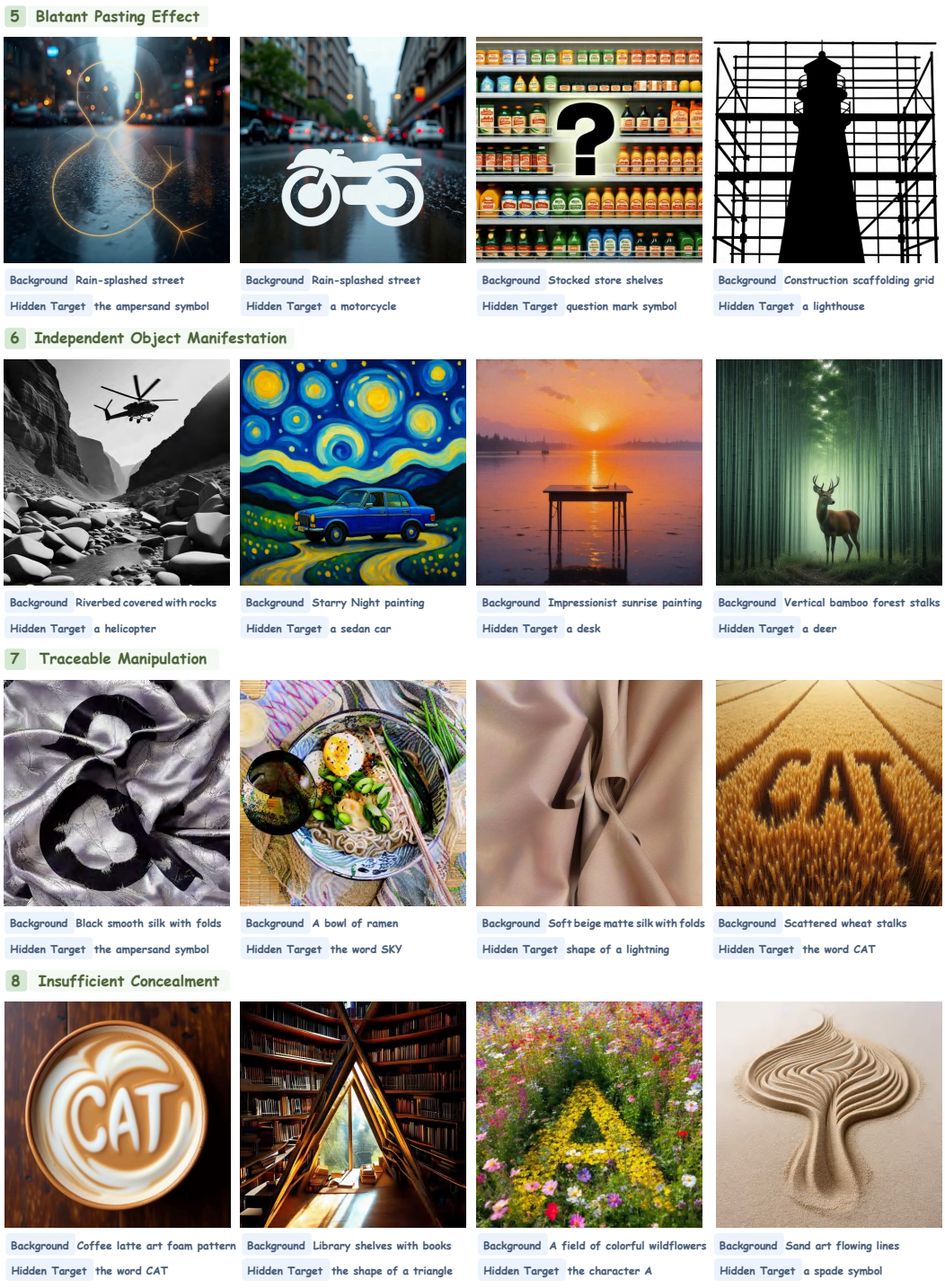

Figure 11: Additional case studies on Type 5–8 errors observed in HiP-Bench, as defined in RQ4: Failure Mode Analysis. Each row presents one error type, with four examples shown across the row to illustrate the corresponding failure mode.

hidden-content recognition by suppressing redundant visual details and exposing global patterns. These findings motivate our use of image resizing as a controlled transformation for evaluating scale-dependent visibility in MDS.

For the generation of other visual illusions, Visual Anagrams (Geng et al., 2024) generates images whose appearance changes through transformations such as flips, rotations, or pixel rearrangements, while Diffusion Illusions (Burgert et al., 2024) optimizes images that produce illusions after specific physical arrangements.

For hidden-picture optical illusion generation, PTDiffusion (Gao et al., 2025) embeds a reference image into a text-described scene via training-free phase transfer, and QR-code-style control models (Monster labs, 2023a;b) embed recognizable patterns into natural images.

B.2 Visual Generative Models

Text-to-image generation has developed from diffusion models to larger systems with improved prompt following and high-resolution synthesis. Latent Diffusion Models, represented by Stable Diffusion (Rombach et al., 2022), reduce the cost of generation by operating in a latent space and have become a common basis for open-source T2I models. SDXL (Podell et al., 2024) extends this line with a larger model and improved conditioning, while DeepFloyd IF (DeepFloyd Lab, 2023) uses a cascaded pipeline that generates low-resolution images and then applies super-resolution stages. Diffusion transformer models, such as PixArt-α (Chen et al., 2024c) and Hunyuan-DiT (Li et al., 2024), apply transformer architectures to high-resolution text-conditioned generation. Recent flow-matching-based models, including FLUX.1 (Black Forest Labs, 2024) and Stable Diffusion 3.5 (Stability AI, 2024a), use flow-matching or related formulations to improve sampling efficiency and image quality. In addition, unified multimodal models such as Janus-Pro (Chen et al., 2025b) and OmniGen (Xiao et al., 2025) incorporate T2I generation into broader frameworks for visual understanding and generation.

Spatial control has been widely studied in diffusion-based image generation. ControlNet (Zhang et al., 2023a) adds trainable control branches to pretrained diffusion models, enabling conditions such as depth maps, edges, and segmentation maps to guide image synthesis. Later methods reduce training cost through training-free control (Mo et al., 2024), extending text-to-image models from prompt-only generation to structurally controlled generation.

Instruction-based image editing uses text commands to guide changes to an existing image. InstructPix2Pix (Brooks et al., 2023) trains a diffusion model on paired instruction-edit data to follow editing instructions. CosXL Edit (Stability AI, 2024c) builds an instruction editor on the SDXL framework and supports higher-resolution editing. OmniGen (Xiao et al., 2025) and OmniGen2 (Wu et al., 2026) incorporate language-guided editing into unified multimodal generation models, where visual inputs and text instructions are handled within one model.

B.3 Automated Evaluation

Automated evaluation of visual generation commonly relies on image quality, aesthetics, and text-image alignment metrics. No-reference quality and aesthetic metrics, such as NIQE (Mittal et al., 2013), MUSIQ (Ke et al., 2021), and LAION-Aes (Schuhmann et al., 2022), estimate perceptual quality or aesthetic preference, while CLIP-based metrics (Radford et al., 2021), BLIP-based metrics (Li et al., 2022), and VQAScore (Lin et al., 2025) measure prompt-image consistency. These metrics are designed for visible semantic content or general perceptual quality, and therefore provide limited evidence for hidden-picture optical illusions.

Specialized evaluation for concealed objects has mainly been developed for camouflage quality assessment and foreground localization. HVS-based assessment models (Sun et al., 2025) estimate camouflage quality by modeling human-vision-related processes, including target localization, feature extraction, and quality scoring, and are designed to correlate with human opinion scores. Context-measure (Wang et al., 2026) further emphasizes the context-dependent nature of camouflage by modeling pixel-wise correlations between the target and its surrounding regions. Another line of work evaluates the accuracy of predicted foreground maps against ground-truth masks. Weighted F-measure (Margolin et al., 2014) improves the standard F-measure for foreground map evaluation, S-measure (Fan et al., 2017) measures region- and object-level structural similarity, and

E-measure (Fan et al., 2018) combines local pixel matching with image-level alignment. These metrics are useful for camouflaged object analysis, but they are not designed for evaluating hidden-picture optical illusions and usually assume ground-truth masks, reference annotations, or controlled foreground–background settings, which are often unavailable in some generative settings.

Recent evaluation work has explored LLM-as-a-Judge as a flexible alternative to fixed automatic metrics. This paradigm was first widely used in open-ended language generation and dialogue evaluation, where strong language models are prompted to score or compare model outputs according to natural-language criteria (Zheng et al., 2023; Liu et al., 2023; Ye et al., 2025a;b). The evaluation signals from such assessments have also been leveraged as quality feedback to guide response generation during training (Ye et al., 2024). Beyond text-only evaluation, this paradigm has been extended to multimodal domains, where MLLMs serve as judges to assess generated images with task-specific rubrics including prompt faithfulness, visual quality, and instruction following (Chen et al., 2024a; Xiong et al., 2025). The landscape of multimodal evaluation continues to be enriched by broader advances in MLLM capabilities and task-specific metric design (Ye et al., 2025c; 2026).

C Taxonomy Design

C.1 Taxonomy Rationale

We use the background and hidden-target categories to construct a balanced benchmark rather than to define mutually exclusive perceptual classes. Each sample is assigned according to its dominant visual, semantic, or structural property. A structured background may therefore also contain high spatial frequency or strong orientation, while hidden targets from different semantic groups may share geometric or topological properties. To describe these group-level tendencies, we complement the category labels with quantitative attributes computed for all 128 backgrounds and 128 hidden targets.

All images were resized to 512×512 pixels before measurement. Background images were converted to grayscale. For each hidden-target image, we estimated the white background from its border pixels and extracted a binary foreground mask. Small isolated regions were removed, a one-pixel morphological closing was applied, and small holes were filled. The same processing and metric settings were used for all samples. Each category contains 32 samples, and the results below are reported as mean $\pm$ standard deviation.

Background Attributes For the backgrounds, we measured spatial frequency, structure-tensor coherence, and the number of Harris corners. These attributes describe texture variation, gradient orientation, and local structural patterns, respectively. The results are reported in Table 7.

Category	Spatial Frequency	Structure-Tensor Coherence	Harris Corner Count
High-Frequency	0.0920 ± 0.0241	0.6240 ± 0.1226	181.28 ± 160.12
Structured	0.1289 ± 0.0534	0.7904 ± 0.1162	225.00 ± 210.12
Directional	0.0886 ± 0.0595	0.7808 ± 0.1128	102.34 ± 151.38
Low-Frequency	0.0123 ± 0.0094	0.8671 ± 0.1382	5.75 ± 8.46

Table 7: Background attributes. Values are reported as mean $\pm$ standard deviation. Low-Frequency backgrounds are smooth, with low spatial frequency and few corners; High-Frequency backgrounds contain fine, irregular details with low coherence; Structured backgrounds show strong texture variation and abundant local intersections; and Directional backgrounds exhibit coherent oriented patterns with fewer corners.

Low-Frequency backgrounds had lower spatial frequency than the other three groups, consistent with their smoother appearance. Structure-tensor coherence was highest for Low-Frequency backgrounds, similar for Structured and Directional backgrounds, and lowest for High-Frequency backgrounds. Coherence measures the alignment of image gradients rather than their frequency. Smooth images with consistent gradients can therefore have high coherence, while irregular fine-scale patterns can produce lower coherence; the similar values for Structured and Directional backgrounds indicate that both contain oriented patterns. Structured backgrounds also had a higher mean Harris

corner count than Directional backgrounds. Harris corners capture local intersections and changes in edge direction, which occur more often in rectilinear structures than in continuous directional patterns. Together, these attributes distinguish the smoothness of Low-Frequency backgrounds, the fine and irregular detail of High-Frequency backgrounds, and the different forms of orientation in Structured and Directional backgrounds.

Hidden Target Attributes For the hidden targets, we measured circularity, solidity, and the number of connected foreground components from their binary masks. The results are reported in Table 8.

Category	Circularity	Solidity	Connected Components
Textual Elements	0.0785 ± 0.0612	0.4117 ± 0.0706	4.84 ± 4.62
Geometry & Shapes	0.0663 ± 0.0283	0.3037 ± 0.0897	1.25 ± 0.67
Rigid Objects	0.2663 ± 0.2509	0.6714 ± 0.2282	1.09 ± 0.39
Non-Rigid Objects	0.3820 ± 0.2615	0.7948 ± 0.1472	1.03 ± 0.18

Table 8: Hidden-target attributes. Values are reported as mean $\pm$ standard deviation. Circularity and solidity were highest for Non-Rigid Objects, followed by Rigid Objects, Textual Elements, and Geometry & Shapes. Textual Elements had 4.84 connected components on average, whereas Geometry & Shapes, Rigid Objects, and Non-Rigid Objects had approximately one component per target. These differences are consistent with the visual forms of the four categories.

Circularity was highest for Non-Rigid Objects, followed by Rigid Objects, while Textual Elements and Geometry & Shapes had lower values. Solidity followed the same order: Non-Rigid Objects had the highest value, followed by Rigid Objects, Textual Elements, and Geometry & Shapes. Textual Elements contained an average of 4.84 connected components, whereas the other three categories contained approximately one component per target. These group-level differences are consistent with the visual forms represented by the four semantic categories.

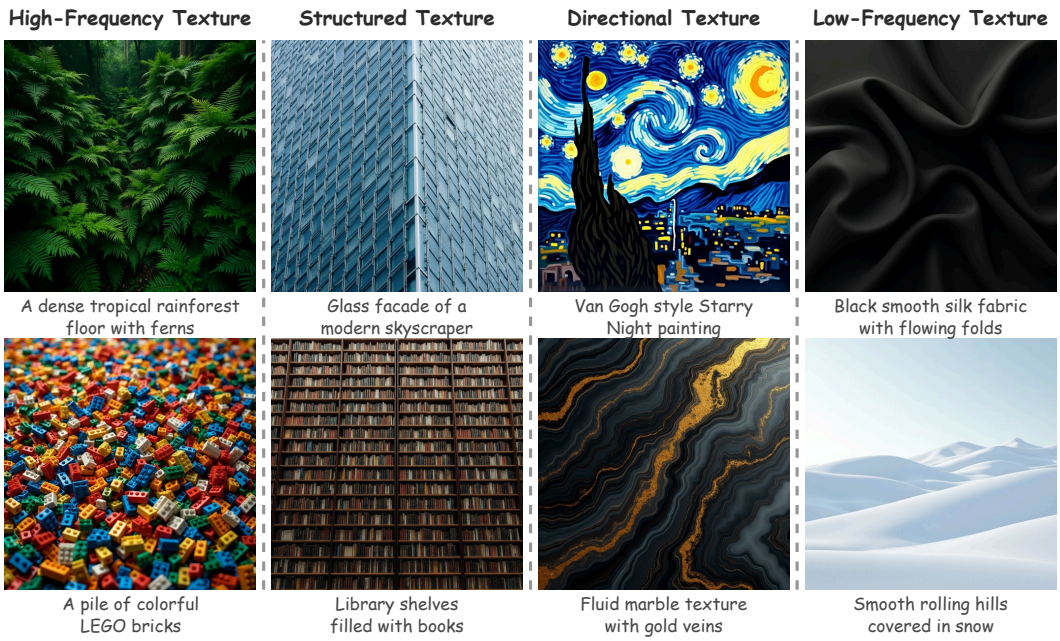

Figure 12: Visual examples of the 4 background types. Each column shows representative examples of one type.

C.2 Visual Examples of Taxonomy Categories

This section provides visual examples of the taxonomy defined in the main text. Figure 12 displays the 4 background types, while Figure 13 presents the 12 hidden target sub-categories organized under the 4 hidden target types.

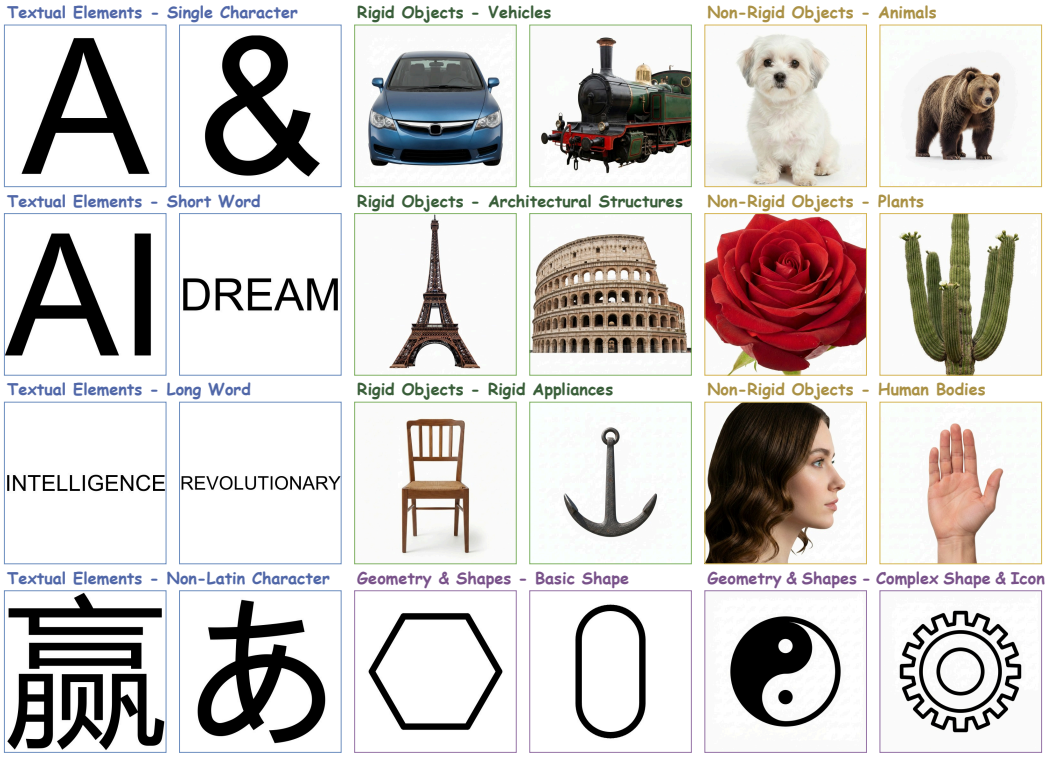

Figure 13: Visual examples of the hidden target taxonomy. The 12 hidden target sub-categories are grouped under 4 high-level types: Textual Elements, Geometry & Shapes, Rigid Objects, and Non-Rigid Objects.

D Dataset Construction and Input Preparation

D.1 Background-Object Pairing Strategy

Figure 14 illustrates the background-object pairing strategy used to construct the dataset. The dataset contains four background categories and four hidden target categories, with 32 entries in each category. This gives 128 background descriptions and 128 hidden targets in total. Pairing is performed by matching entries with the same within-category index. For the i-th entry, where $i \in \{1, \ldots, 32\}$, the i-th hidden target from each hidden target category is paired with the i-th background description from each background category. Thus, each index produces $4 \times 4 = 16$ object-background combinations. Repeating this process for all 32 indices gives $16 \times 32 = 512$ combinations. This strategy ensures that every hidden target is evaluated across all four background categories and that all 16 background–target category pairs are equally represented.

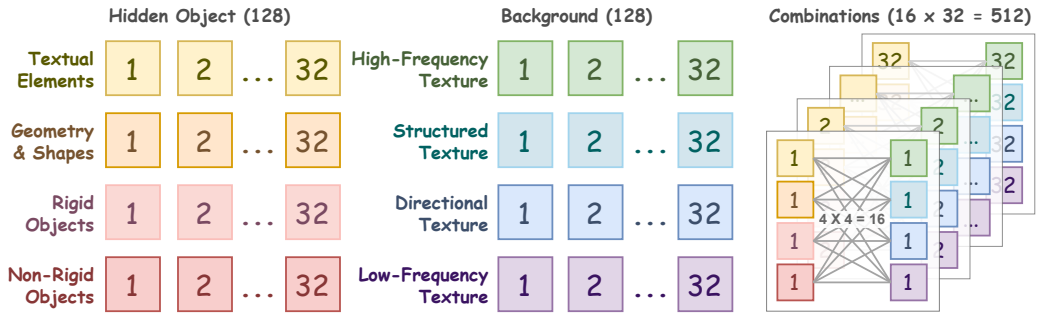

Figure 14: Background-object pairing strategy used for dataset construction. Both hidden targets and backgrounds are organized into four categories, with 32 entries per category. Pairing is based on the within-category index: for each index i, the i-th object from each hidden target category is combined with the i-th background from each background category, producing $4 \times 4 = 16$ combinations. Repeating this process over 32 indices yields $16 \times 32 = 512$ balanced object-background combinations.

D.2 Prompt Generation for Text-to-Image Models

We use the following system prompt with GPT-4o (OpenAI, 2024) (temperature set to 0.7) to generate specific text-to-image prompts based on different combinations of backgrounds and hidden targets.

Text-to-Image Prompt Generation Template

System Role: You are an expert Prompt Engineer specializing in Generative AI and Optical Illusions. Your task is to fuse a "Background Scene" and a "Hidden Concept" into a single, cohesive Text-to-Image (T2I) prompt that triggers *Pareidolia* (the perception of a specific image in a random or ambiguous visual pattern).

Input Data:

- < Background Scene >: A description of the main visual environment (e.g., "A dense tropical rainforest").
- < Hidden Concept >: The object or shape that needs to be subliminally hidden within the background (e.g., "Cat").

Critical Instructions:

1. **Primary Focus:** The image must *primarily* look like the < Background Scene >. The < Hidden Concept > must effectively be an *optical illusion* formed by lighting, textures, negative space, or the arrangement of objects.
2. **Avoid Literalism:** NEVER write "A cat standing in the forest." The hidden concept must *not* be a physical object present in the scene. It must be an *emergent shape*.
3. **Vocabulary Strategy:** Use phrases that suggest subtlety and emergence, such as: "subtly forming the silhouette of...", "casting shadows that resemble...", "arranged in a pattern suggesting...", "a subliminal illusion of...", "negative space creating the outline of...".
4. **Style Consistency:** Maintain the artistic style or texture implied by the Background Scene (e.g., if the background is "Oil Painting," the hidden shape must be formed by brushstrokes).

Response Format: Output ONLY a valid JSON object containing a single key `"prompt"`. Do not output any other text.

Examples:

Input:

- < Background Scene >: "A messy pile of rusty mechanical parts and gears."
- < Hidden Concept >: "Skull"

Output: `{"prompt": "A chaotic pile of rusty mechanical parts and gears, where the arrangement of metal components and dark shadows inadvertently creates a grim, subliminal silhouette of a skull."}`

Input:

- < Background Scene >: "Vast Sahara sand dunes casting long shadows at sunset."
- < Hidden Concept >: "Pyramids"

Output: `{"prompt": "Vast Sahara sand dunes under a setting sun, with the interplay of golden light and deep shadows forming the distinct, illusionary triangular shapes of Pyramids on the desert floor."}`

Input:

- < Background Scene >: "Van Gogh style Starry Night painting."
- < Hidden Concept >: "Spiral"

Output: `{"prompt": "A Van Gogh style oil painting of a starry night, where the swirling blue and yellow brushstrokes naturally converge to form a prominent, hypnotic Spiral pattern in the sky."}`

D.3 Data Preparation for Instruction-Based Image Editing Models

To generate background images, we utilize the FLUX.1 [dev] (Black Forest Labs, 2024). To control the frequency and structural properties of the generated textures, we append category-specific modifiers to the base background descriptions. The detailed texture modifiers and generation hyperparameters are provided in the Prompt Box below.

Texture Modifiers and Generation Settings

High-Frequency:
"extremely crowded, cluttered composition, edge to edge fill, no empty space, visual chaos, hyper-maximalist, highly detailed granular texture, full frame coverage, intricate noise everywhere, density variation, rugged surface, 8k raw photo."

Structured High-Frequency:
"rigid geometric pattern, repetitive grid, architectural symmetry, sharp straight lines, mathematical precision, tessellated layout, structured alignment, clear defined edges, rhythmic visual order, full frame geometric fill, no negative space."

Mid-Frequency Directional:
"strong directional flow, sweeping motion across entire frame, organic curves, deep flowing lines, heavy textured strokes, coherent orientation, fluid composition, volumetric lighting, dense and continuous pattern, no gaps, artistic motion blur, highly textured surface."

Low-Frequency Smooth:
"high fidelity material render, smooth polished surface, soft studio lighting, gentle ambient occlusion, clean lines, shallow depth of field, sharp focus on subject, subtle tonal gradations, pristine condition, 8k realistic photography, not abstract."

FLUX.1 [dev] Settings: Resolution: 1024 × 1024 | Inference Steps: 25 | Guidance Scale: 3.5 | Candidates per Prompt: 8

For each background prompt, eight candidate images are generated, and two human annotators independently use a custom interface (Figure 15) to select the most suitable one. When the two selections agree, the selected image is retained; otherwise, a third human annotator resolves the disagreement by choosing between the two initially selected candidates.

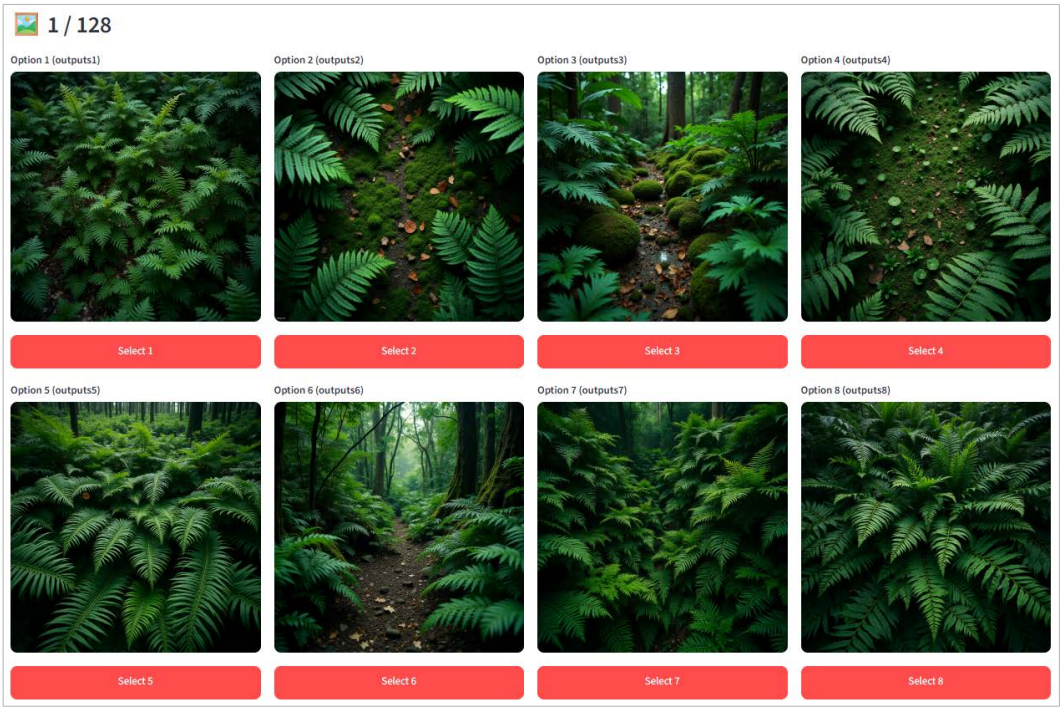

Figure 15: Human curation interface for background image selection. For each background prompt, the interface displays eight candidate images simultaneously. Annotators inspect the candidates and select the image that best matches the intended background description and texture category for subsequent dataset construction.

We use the following system prompt with GPT-4o (temperature set to 0.7) to generate specific editing instructions based on different background–hidden target combinations.

Editing Instruction Generation Template

System Role: You are an expert Prompt Engineer specializing in Instruction-based Image Editing and Optical Illusions. Your task is to generate a precise editing instruction that tells an AI model (like InstructPix2Pix) how to modify an existing source image to hide a specific concept within it.

Input Data:

- < Source Image Description >: What the current image looks like (e.g., "A dense tropical rainforest").
- < Texture Type >: The texture category of the source image (High-Frequency Texture, Structured Texture, Directional Texture, Low-Frequency Smooth Texture).
- < Hidden Concept >: The object or shape that needs to be subliminally inserted (e.g., "Cat").

Critical Instructions:

1. **Instruction Format:** The output must be a direct *command* or *request* to the AI editor. Start with verbs like "Edit...", "Change...", "Modify...", "Subtly rearrange...", "Alter...".
2. **Respect the Source:** Do NOT ask to replace the entire image. The instruction must explicitly mention using the *existing elements* (textures, shadows, lines, or patterns) of the source image to form the illusion.
3. **Illusion Strategy based on Texture:**
   - If *High-Frequency Texture*: Ask to reorganize the noise or debris.
   - If *Structured Texture*: Ask to alter the grid or alignment.
   - If *Directional Texture*: Ask to bend the flow or lines.
   - If *Low-Frequency Smooth Texture*: Ask to use subtle shading or lighting changes.
4. **Subtlety is Key:** Use adverbs like "subtly", "barely visible", "faintly", "disguised as".

5. **No Hallucinations:** Do not ask to add objects that don't belong (e.g., don't say "add a cat statue"). Say "make the ferns look like a cat".

Response Format: Output ONLY a valid JSON object containing a single key `"instruction"`. Do not output any other text.

Examples:

Input:
- < Source Image Description >: "A messy pile of rusty mechanical parts and gears."
- < Texture Type >: "High-Frequency Texture"
- < Hidden Concept >: "Skull"

Output: `{"instruction": "Subtly rearrange the rusty gears and shadows to faintly form the silhouette of a skull without adding new objects."}`

Input:
- < Source Image Description >: "Glass facade of a modern skyscraper."
- < Texture Type >: "Structured Texture"
- < Hidden Concept >: "Letter A"

Output: `{"instruction": "Modify the grid lines of the windows to create a subliminal geometric pattern that resembles the letter A."}`

Input:
- < Source Image Description >: "Vast Sahara sand dunes casting long shadows."
- < Texture Type >: "Directional Texture"
- < Hidden Concept >: "Pyramids"

Output: `{"instruction": "Alter the curves of the sand dunes and their shadows to suggest the triangular shapes of Pyramids."}`

Input:
- < Source Image Description >: "Pristine white snowfield."
- < Texture Type >: "Low-Frequency Smooth Texture"
- < Hidden Concept >: "Bear"

Output: `{"instruction": "Add very faint shading to the snow surface to create a barely visible impression of a bear's outline."}`

D.4 Prompt Validity Audit

Since the T2I prompts and Editing instructions were prepared with LLM assistance, we conduct a manual validity audit of all final textual inputs used in HiP-Bench. The purpose of this audit is to verify that each input correctly specifies its intended benchmark instance and expresses the hidden-picture generation objective. Two human annotators independently examine all 512 T2I prompts and all 512 Editing instructions using the following criteria.

For T2I Prompts:

Background Correctness: The prompt correctly describes the intended background scene.

Target Correctness: The prompt correctly identifies the designated hidden target.

Target–Background Integration: The prompt requires the target to be integrated into the background rather than rendered as an independent foreground object.

For Editing Instructions:

Target Correctness: The instruction correctly identifies the designated hidden target.

Background-Based Integration: The instruction requires the target to be formed through modifications to the provided background rather than added as an independent foreground object.

As summarized in Table 9, all final textual inputs satisfy every applicable criterion according to both annotators.

Audit Criterion	Passed by Both
Text-to-image	
Background Correctness	512 / 512
Target Correctness	512 / 512
Target–Background Integration	512 / 512
Instruction-based image editing	
Target Correctness	512 / 512
Background-Based Integration	512 / 512

Table 9: Results of manual validity audit of the final LLM-generated textual inputs. Two annotators independently examine every input. An input is passed only when both annotators confirm that it satisfies the corresponding criterion.

D.5 Control Image Preprocessing for Image-Conditioned Generation Models

The hidden targets consist of four categories: Textual Elements and Geometry & Shapes are programmatically generated using rendering scripts, while Rigid Objects and Non-Rigid Objects are synthesized using the Nano Banana 2 image model(Google, 2026). We apply distinct preprocessing pipelines to convert these source images into four specific control mechanisms. The exact operation and hyperparameters for each control condition are detailed in Table 10, with visual examples shown in Figure 16.

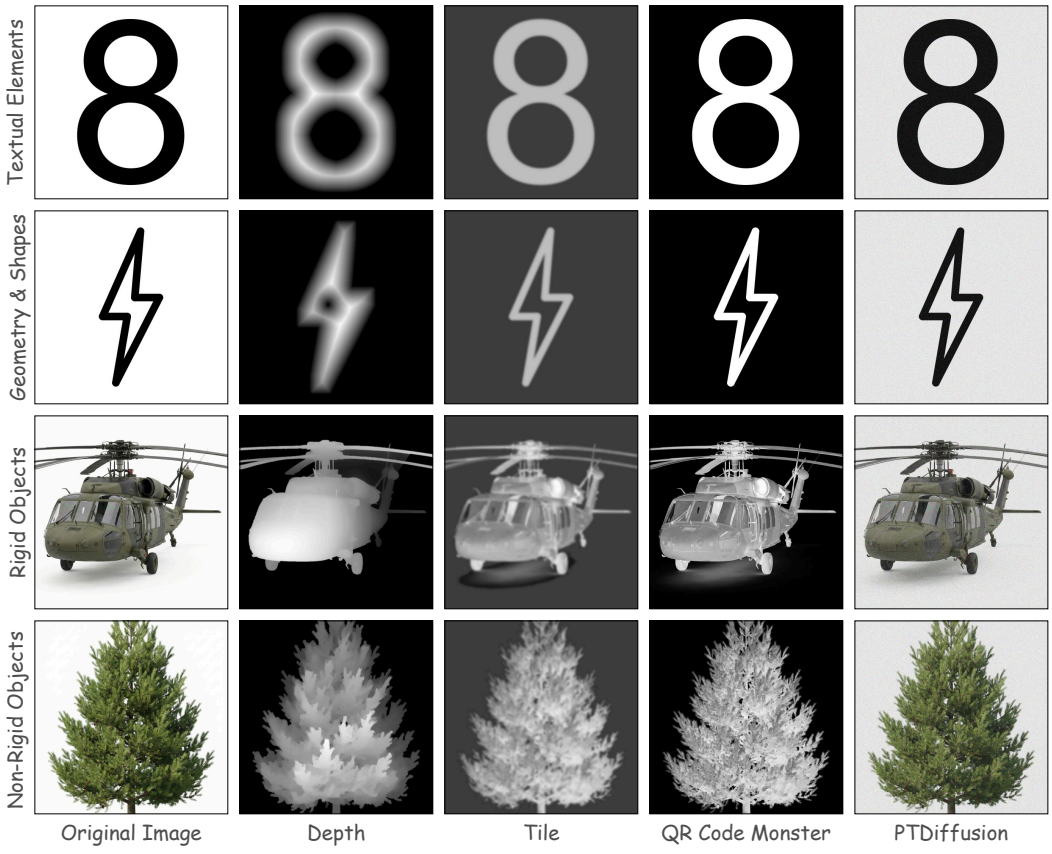

Figure 16: Examples of preprocessed control images for ControlNet variants and PTDiffusion. Hidden target source images are transformed into model-specific control conditions using the preprocessing pipelines summarized in Table 10.

Control Mechanism	Textual Elements, Geometry & Shapes	Rigid & Non-Rigid Objects
Depth	Apply dilation (1 to 5 iterations depending on thickness). Perform L2 distance transform. Apply gamma correction ($\gamma = 0.8$). Apply Gaussian blur (9×9).	Matte to black background via IS-Net (Qin et al., 2022). Extract depth via Depth-Anything-V2-Large (Yang et al., 2024). Normalize to $[0, 255]$. Mask contours. Apply Gaussian blur (5×5).
Tile	Invert binary inputs. Linearly map values from $[0, 255]$ to $[60, 190]$. Apply Gaussian blur ($\sigma = 5.0$).	Convert to 8-bit grayscale and invert. Lift pixels (< 20) to 60. Apply Gaussian blur ($\sigma = 5.0$).
QR Code Monster	Invert binary inputs.	Convert to 8-bit grayscale. Invert the image.
PTDiffusion	Compress dynamic range from $[0, 255]$ to $[20, 235]$. Add Gaussian noise ($\mu = 0, \sigma = 25$) for texture. Clip values to $[0, 255]$.	

Table 10: Preprocessing pipelines and hyperparameters for converting hidden target source images into specific control inputs.

E More Statistics of HiP-Bench

Table 11 summarizes the prepared inputs and generated images across the three generative paradigms, including the subset selected for human evaluation. Figure 17 details the hierarchical composition of the 128 hidden targets across four evenly divided super-categories and 12 subcategories. Figure 18 illustrates the 512 background–hidden target combinations, where each hidden target is paired with all four background textures, ensuring a balanced distribution within each subcategory.

Item	Count
Prepared Inputs (for 3 paradigms)	
Text-to-image prompts	512
Background images + Editing instructions	512
Hidden target images + Background descriptions	512
Generated Images (Human-Evaluated / All)	
Overall	13,824 / 17,920
Text-to-image	7,168 / 11,264
Instruction-based image editing	2,560 / 2,560
Image-conditioned generation	4,096 / 4,096

Table 11: Statistics of prepared inputs and generated images in HiP-Bench.

F Evaluated Models and Implementation Details

F.1 Model Specifications

The evaluated models are selected according to the three input paradigms defined in HiP-Bench. Tables 12, 13, and 14 detail the specifications of the 35 models across text-to-image, instruction-based image editing, and image-conditioned generation. To ensure reproducibility, we fix the random seed to 42 for all tasks. For each evaluated model, the complete evaluation run took less than 5 hours on 4 local NVIDIA RTX 3090 24GB GPUs.

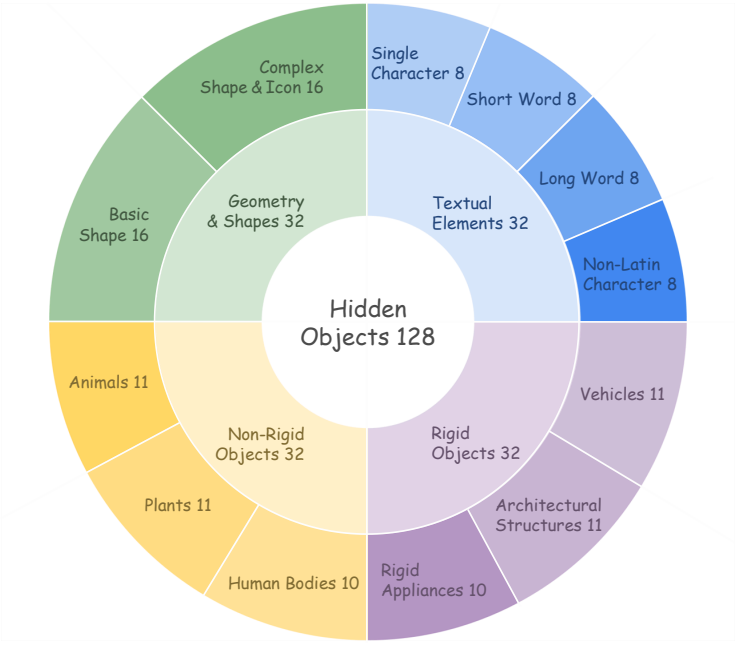

Figure 17: The hierarchical composition of the 128 hidden targets. The inner ring represents the four super-categories, each with 32 targets, and the outer ring indicates the instance counts for the 12 sub-categories.

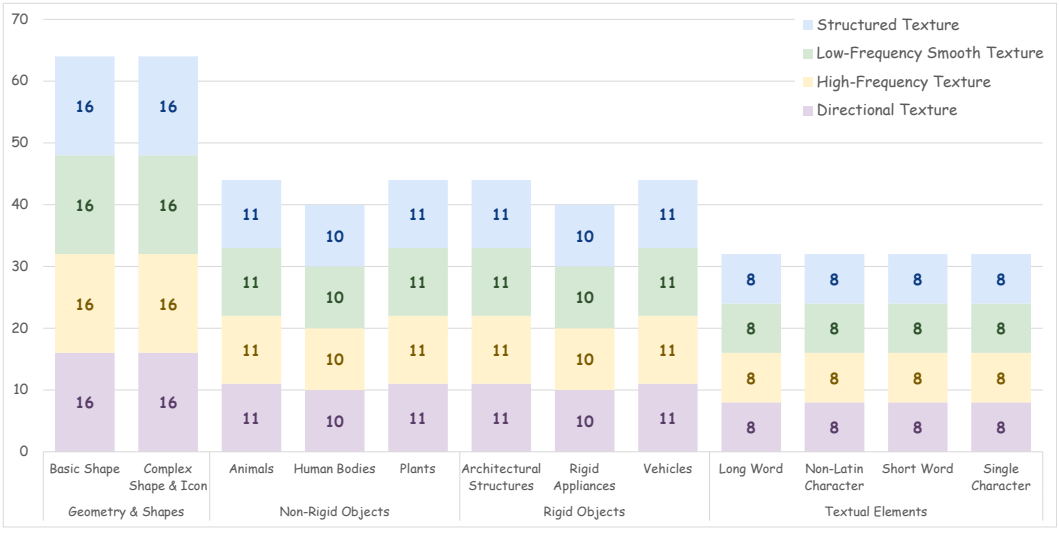

Figure 18: The distribution of the 512 background-object combinations. The x-axis groups the 12 object sub-categories by super-category. The stacked bars demonstrate the equal distribution of the four background textures within each sub-category.

Model	Model ID & Settings
FLUX.1 [dev] (Black Forest Labs, 2024)	`black-forest-labs/FLUX.1-dev` Steps: 50, CFG scale: 3.5, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
FLUX.1 [schnell] (Black Forest Labs, 2024)	`black-forest-labs/FLUX.1-schnell` Steps: 4, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
SD 3.5 Large (Stability AI, 2024a)	`stabilityai/stable-diffusion-3.5-large` Steps: 28, CFG scale: 3.5, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
SD 3.5 Large Turbo (Stability AI, 2024a)	`stabilityai/stable-diffusion-3.5-large-turbo` Steps: 4, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
SD 3.5 Medium (Stability AI, 2024a)	`stabilityai/stable-diffusion-3.5-medium` Steps: 40, CFG scale: 4.5, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
SDXL Base (Podell et al., 2024)	`stabilityai/stable-diffusion-xl-base-1.0` Steps: 40, CFG scale: 5.0, Sampler: EulerDiscreteScheduler, Resolution: 1024×1024
SDXL Turbo (Sauer et al., 2025)	`stabilityai/sdxl-turbo` Steps: 2, Sampler: EulerAncestralDiscreteScheduler, Resolution: 1024×1024
SD 1.5 (Rombach et al., 2022)	`stable-diffusion-v1-5/stable-diffusion-v1-5` Steps: 50, CFG scale: 7.5, Sampler: PNDMScheduler, Resolution: 512×512

DeepFloyd IF (XL-L-x4) (DeepFloyd Lab, 2023; Rombach et al., 2022)	`DeepFloyd/IF-I-XL-v1.0`
	Steps: 100, CFG scale: 7.0, Sampler: DDPMScheduler, Resolution: 64×64
	`DeepFloyd/IF-II-L-v1.0`
	Steps: 50, CFG scale: 4.0, Sampler: DDPMScheduler, Resolution: 256×256
	`stabilityai/stable-diffusion-x4-upscaler`
	Steps: 75, CFG scale: 9.0, Sampler: DDIMScheduler, Resolution: 1024×1024

DeepFloyd IF (L-L-x4) (DeepFloyd Lab, 2023; Rombach et al., 2022)	`DeepFloyd/IF-I-L-v1.0`
	Steps: 100, CFG scale: 7.0, Sampler: DDPMScheduler, Resolution: 64×64
	`DeepFloyd/IF-II-L-v1.0`
	Steps: 50, CFG scale: 4.0, Sampler: DDPMScheduler, Resolution: 256×256
	`stabilityai/stable-diffusion-x4-upscaler`
	Steps: 75, CFG scale: 9.0, Sampler: DDIMScheduler, Resolution: 1024×1024

DeepFloyd IF (M-M-x4) (DeepFloyd Lab, 2023; Rombach et al., 2022)	`DeepFloyd/IF-I-M-v1.0`
	Steps: 100, CFG scale: 7.0, Sampler: DDPMScheduler, Resolution: 64×64
	`DeepFloyd/IF-II-M-v1.0`
	Steps: 50, CFG scale: 4.0, Sampler: DDPMScheduler, Resolution: 256×256
	`stabilityai/stable-diffusion-x4-upscaler`
	Steps: 75, CFG scale: 9.0, Sampler: DDIMScheduler, Resolution: 1024×1024

Hunyuan-DiT v1.2 (Li et al., 2024)	`Tencent-Hunyuan/HunyuanDiT-v1.2` Steps: 50, CFG scale: 5.0, Sampler: DDPMScheduler, Resolution: 1024×1024
PixArt-α (Chen et al., 2024c)	`PixArt-alpha/PixArt-XL-2-1024-MS` Steps: 20, CFG scale: 4.5, Sampler: DPMSolverMultistepScheduler, Resolution: 1024×1024
PixArt-δ (Chen et al., 2024b)	`PixArt-alpha/PixArt-LCM-XL-2-1024-MS` Steps: 4, CFG scale: 4.5, Sampler: DPMSolverMultistepScheduler, Resolution: 1024×1024
PixArt-Σ (Chen et al., 2025a)	`PixArt-alpha/PixArt-Sigma-XL-2-1024-MS` Steps: 20, CFG scale: 4.5, Sampler: DPMSolverMultistepScheduler, Resolution: 1024×1024
Janus-Pro (1B) (Chen et al., 2025b)	`deepseek-ai/Janus-Pro-1B`, Temperature: 1.0, CFG scale: 5.0, Resolution: 384×384
Janus-Pro (7B) (Chen et al., 2025b)	`deepseek-ai/Janus-Pro-7B`, Temperature: 1.0, CFG scale: 5.0, Resolution: 384×384
JanusFlow (Ma et al., 2025)	`deepseek-ai/JanusFlow-1.3B` Steps: 30, CFG scale: 2.0, Resolution: 384×384
OmniGen v1 (Xiao et al., 2025)	`Shitao/OmniGen-v1` Steps: 50, CFG scale: 2.5, Sampler: Euler (OmniGenScheduler), Resolution: 1024×1024
OmniGen2 (Wu et al., 2026)	`OmniGen2/OmniGen2` Steps: 50, CFG scale: 5, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
Lumina-Image 2.0 (Qin et al., 2025)	`Alpha-VLLM/Lumina-Image-2.0` Steps: 50, CFG scale: 4.0, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
Lumina-Next-SFT (Zhuo et al., 2024)	`Alpha-VLLM/Lumina-Next-SFT` Steps: 30, CFG scale: 4.0, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024

Table 12: Configurations of the text-to-image generation models. Model IDs refer to Hugging Face repositories when applicable. For multi-stage pipelines, stages are listed in execution order. All generations use a fixed random seed of 42.

Model	Model ID & Settings
InstructPix2Pix (Brooks et al., 2023)	`timbrooks/instruct-pix2pix` Steps: 100, Sampler: EulerAncestralDiscreteScheduler, Resolution: 512×512
CosXL Edit (Stability AI, 2024c)	`stabilityai/cosxl` Steps: 40, Sampler: EDMEuler, Resolution: 1024×1024
FLUX.1 Kontext (Black Forest Labs et al., 2025)	`black-forest-labs/FLUX.1-Kontext-dev` Steps: 28, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
OmniGen v1 (Xiao et al., 2025)	`Shitao/OmniGen-v1` Steps: 50, Sampler: Euler (OmniGenScheduler), Resolution: 1024×1024
OmniGen2 (Wu et al., 2026)	`OmniGen2/OmniGen2` Steps: 50, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024

Table 13: Configurations of the instruction-based image editing models. Model IDs refer to Hugging Face repositories when applicable. Text Guidance Scale and Image Guidance Scale values are reported in Table 15. All generations use a fixed random seed of 42.

Model	Model ID & Settings
FLUX.1 [dev] + ControlNet Depth (Black Forest Labs, 2024; Shakker Labs, 2024)	`black-forest-labs/FLUX.1-dev` + `Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro Depth` Steps: 50, CFG scale: 3.5, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
SD 1.5 + ControlNet QR Code Monster v2 (Rombach et al., 2022; Monster labs, 2023a)	`stable-diffusion-v1-5/stable-diffusion-v1-5` + `monster-labs/control_v1p_sd15_qrcode_monster` Steps: 50, CFG scale: 7.5, Sampler: PNDMScheduler, Resolution: 512×512
SD 1.5 + ControlNet Depth (Rombach et al., 2022; Zhang et al., 2023a)	`stable-diffusion-v1-5/stable-diffusion-v1-5` + `lllyasviel/control_v11f1p_sd15_depth` Steps: 50, CFG scale: 7.5, Sampler: PNDMScheduler, Resolution: 512×512
SD 1.5 + ControlNet Tile (Rombach et al., 2022; Zhang et al., 2023a)	`stable-diffusion-v1-5/stable-diffusion-v1-5` + `lllyasviel/control_v11f1e_sd15_tile` Steps: 50, CFG scale: 7.5, Sampler: PNDMScheduler, Resolution: 512×512
SDXL Base + ControlNet QR Code Monster (Podell et al., 2024; Monster labs, 2023b)	`stabilityai/stable-diffusion-xl-base-1.0` + `monster-labs/control_v1p_sdxl_qrcode_monster` Steps: 40, CFG scale: 5.0, Sampler: EulerDiscreteScheduler, Resolution: 1024×1024
SDXL Base + ControlNet Depth (Podell et al., 2024; xinsir, 2024)	`stabilityai/stable-diffusion-xl-base-1.0` + `xinsir/controlnet-depth-sdxl-1.0` Steps: 40, CFG scale: 5.0, Sampler: EulerDiscreteScheduler, Resolution: 1024×1024
SD 3.5 Large + ControlNet Depth (Stability AI, 2024a;b)	`stabilityai/stable-diffusion-3.5-large` + `stabilityai/stable-diffusion-3.5-large-controlnet-depth` Steps: 28, CFG scale: 3.5, Sampler: FlowMatchEulerDiscreteScheduler, Resolution: 1024×1024
PTDiffusion (on SD 1.5) (Rombach et al., 2022; Gao et al., 2025)	`stable-diffusion-v1-5/stable-diffusion-v1-5` (Training-free) Steps: 50, CFG scale: 7.5, Sampler: PNDMScheduler, Resolution: 512×512

Table 14: Configurations of the image-conditioned generation models. Model IDs refer to Hugging Face repositories when applicable. Control Scale / Control Start for ControlNet-based models and Phase Ratio for PTDiffusion are reported in Table 16. All generations use a fixed random seed of 42.

F.2 HYPERPARAMETER TUNING

To ensure a fair evaluation, we conduct a targeted grid search to determine the optimal guidance parameters for both instruction-based image editing and image-conditioned generation models.

For instruction-based image editing models, the text prompts already specify the type and intensity of the transformation. Therefore, a single set of optimal parameters is sufficient for each model across all tasks. We sample a validation set of 16 representative cases, covering all 4×4 background and hidden target combinations defined in Table 2 in the main text. Using this subset, we conduct a grid search for InstructPix2Pix, CosXL Edit, and OmniGen2. For FLUX.1 Kontext [dev] and OmniGen v1, we use their default settings to evaluate out-of-the-box performance.

For image-conditioned generation models, the inputs do not specify the intensity of the transformation. Therefore, the parameters must adapt to different background and hidden target types. Based on Table 2, backgrounds are grouped into High-Frequency Texture (HF), Low-Frequency Texture (LF), and Structured/Directional Textures (SD); hidden targets are grouped into Textual Elements/Geometry & Shapes (TG), and Rigid/Non-Rigid Objects (RN). These combinations form six tuning groups: HF-TG, HF-RN, SD-TG, SD-RN, LF-TG, and LF-RN. We sample 3 cases per group (18 total) to conduct the grid search.

We select the final parameters based on four metrics (Table 17): Illusion Fidelity, Overall Visual Quality, Scene Consistency, and Illusion Shape Consistency. Our primary goal is to maximize Illusion Fidelity (effectively hiding the target) without compromising the other three metrics. In failure cases where all searched parameters perform poorly, we directly apply theoretically optimal settings for the specific failure (e.g., maximizing text guidance for ignored instructions, or minimizing control strength for severe structural artifacts).

Tables 15 and 16 report the optimal parameters for instruction-based image editing and image-conditioned generation models. A visual example of the parameter search process is provided in Figure 19.

Model	**Optimal Parameters**
InstructPix2Pix	Text Guidance Scale = 7.5, Image Guidance Scale = 1.5
CosXL Edit	Text Guidance Scale = 7.0, Image Guidance Scale = 1.5
FLUX.1 Kontext [dev]	Guidance Scale = 2.5 (Default)
OmniGen v1	Text Guidance Scale = 2.5, Image Guidance Scale = 1.6 (Default)
OmniGen2	Text Guidance Scale = 5.0, Image Guidance Scale = 1.5

Table 15: Optimal parameter configurations for instruction-based image editing models. The same parameters are fixed across all editing tasks.

Module	**HF-TG**	**HF-RN**	**SD-TG**	**SD-RN**	**LF-TG**	**LF-RN**
FLUX.1 [dev] + ControlNet Depth	0.7 / 0.1	0.5 / 0.0	0.6 / 0.1	0.5 / 0.0	0.6 / 0.1	0.6 / 0.0
SD 1.5 + ControlNet QR Code Monster v2	1.4 / 0.1	1.4 / 0.1	1.2 / 0.1	1.3 / 0.0	0.8 / 0.1	1.3 / 0.0
SD 1.5 + ControlNet Depth	0.8 / 0.1	0.8 / 0.2	1.0 / 0.2	0.6 / 0.1	0.8 / 0.2	0.6 / 0.2
SD 1.5 + ControlNet Tile	0.8 / 0.0	0.6 / 0.1	0.6 / 0.0	0.6 / 0.1	0.4 / 0.1	0.4 / 0.0
SDXL Base + ControlNet QR Code Monster	1.0 / 0.0	1.4 / 0.1	1.0 / 0.1	1.4 / 0.1	1.0 / 0.1	1.4 / 0.1
SDXL Base + ControlNet Depth	0.6 / 0.0	0.4 / 0.0	0.8 / 0.1	0.4 / 0.1	0.6 / 0.1	0.6 / 0.1
SD 3.5 Large + ControlNet Depth	0.7 / 0.1	0.6 / 0.1	0.8 / 0.0	0.7 / 0.1	0.7 / 0.1	0.7 / 0.1
PTDiffusion	Ratio = -10	Ratio = -5	Ratio = -10	Ratio = -2	Ratio = -10	Ratio = 5

Table 16: Optimal parameter configurations for image-conditioned generation models across the six tuning groups. HF, LF, and SD denote high-frequency, low-frequency, and structured / directional backgrounds, respectively; TG and RN denote textual elements / geometry & shapes and rigid / non-rigid objects. For ControlNet-based models, entries denote Control Scale / Control Start. For PTDiffusion, the Phase Ratio is reported.

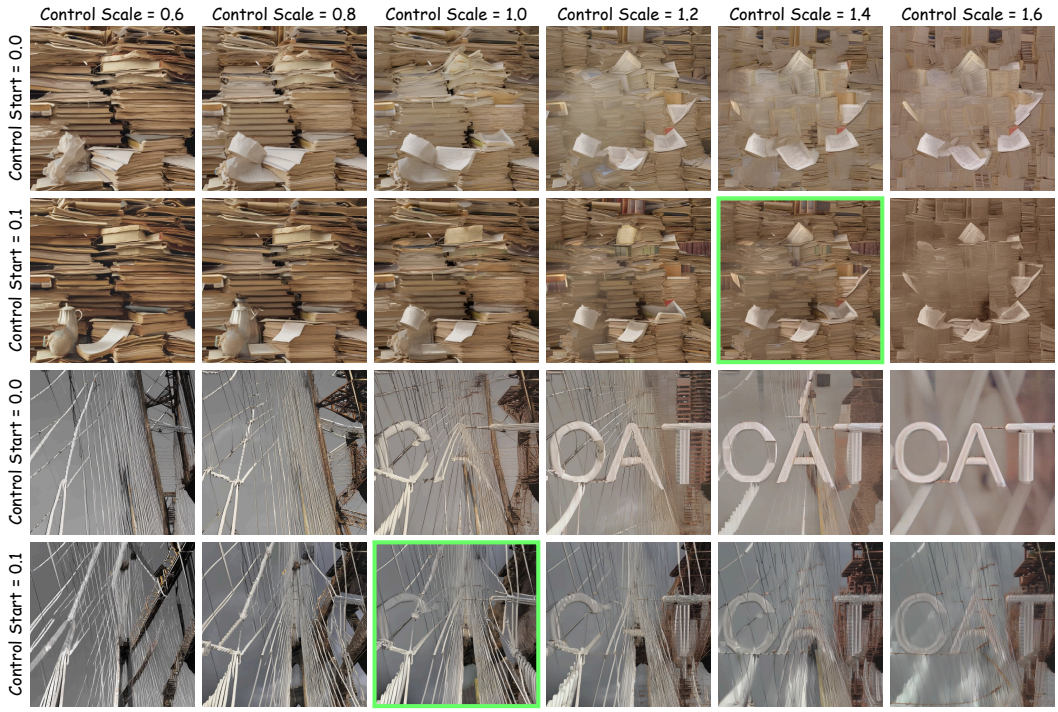

Figure 19: Visual examples of the parameter grid search for image-conditioned generation. Columns vary Control Scale, and rows vary Control Start. The selected optimal configuration for each case is highlighted with a green box.

G Human Annotation Details

G.1 Scoring Guidelines and Inter-annotator Agreements

As Table 17 shows, we provide detailed dimension definitions and scoring guidelines used in HiP-Bench. Each selected image is independently evaluated by a pair of annotators. The annotation workload is evenly distributed across annotators, with each annotator assigned to no more than five separate annotation sessions to limit repeated exposure to the benchmark content.

Score	Description
	Illusion Fidelity (IF) – Ensure the illusion is effectively concealed, exhibiting view-dependent perceptibility and re-identification difficulty.
5	Once the illusion is recognized, it is difficult to quickly relocate it by continuous staring when returning to the normal view.
4	The illusion is hard to identify in a normal view and requires specific conditions (e.g., zooming, distancing, or prompt guidance) to be clearly recognized.
3	The image attempts to integrate the illusion, but its outline can be directly observed in a normal view.
2	Illusion content exists, but it feels disconnected from the overall scene, appearing pasted or isolated.
1	The image only presents scene content; no valid illusion can be identified.
	Overall Visual Quality (VQ) – Ensure the image maintains structural integrity and texture naturalness, free from deliberate layout artifacts.
5	There are no obvious abnormalities or unnaturalness in the overall visual presentation.
4	Only slight, localized unnatural arrangements caused by deliberate illusion crafting are visible, with limited impact on overall visual appeal.
3	Some textures or shapes are clearly arranged to serve a specific effect, noticeably impacting the visual appeal.
2	Significant image quality issues exist, such as abrupt boundaries, scene disconnections, or sharp drops in local quality, severely ruining the visual appeal.
1	Massive structural damage or geometric distortion; inability to form recognizable, coherent content.
	Scene Consistency (SC) – Ensure the semantic alignment between the generated background scene and the provided background description.
1	The scene content of the generated image is basically consistent with the background description.
0	The scene content of the generated image is obviously inconsistent with the background description.
	Illusion Shape Consistency (ISC) – Ensure the semantic correctness and structural accuracy of the generated illusion based on the hidden target description.
3	The illusion content accurately restores the morphological features of the hidden target description in both semantics and structure.
2	The semantics of the illusion are correct, but its structural performance is non-standard or has obvious deviations.
1	The illusion content is completely unrecognizable, with altered semantics or severe structural distortion.

Table 17: Dimension definitions and scoring guidelines for data annotation of HiP-Bench.

We develop a graphical user interface (Figure 20) for human annotation. The interface presents the evaluated images in a randomized order, alongside their corresponding background and hidden-target descriptions. The background and hidden-target descriptions are hidden at the beginning of each annotation trial. Annotators first inspect the original image and its multi-resolution previews without seeing the descriptions. They then explicitly click to reveal the textual descriptions for rubric-based scoring. The interface does not display the hidden-target image, segmentation mask, bounding box, or a highlighted target location.

All annotations were conducted using the same display setup (a 27-inch monitor with a resolution of 2560×1440), with participants seated at an approximately constant viewing distance of 70 cm. To simulate distant viewing, it simultaneously presents the images resized to 128×128, 64×64, and 32×32 resolutions.

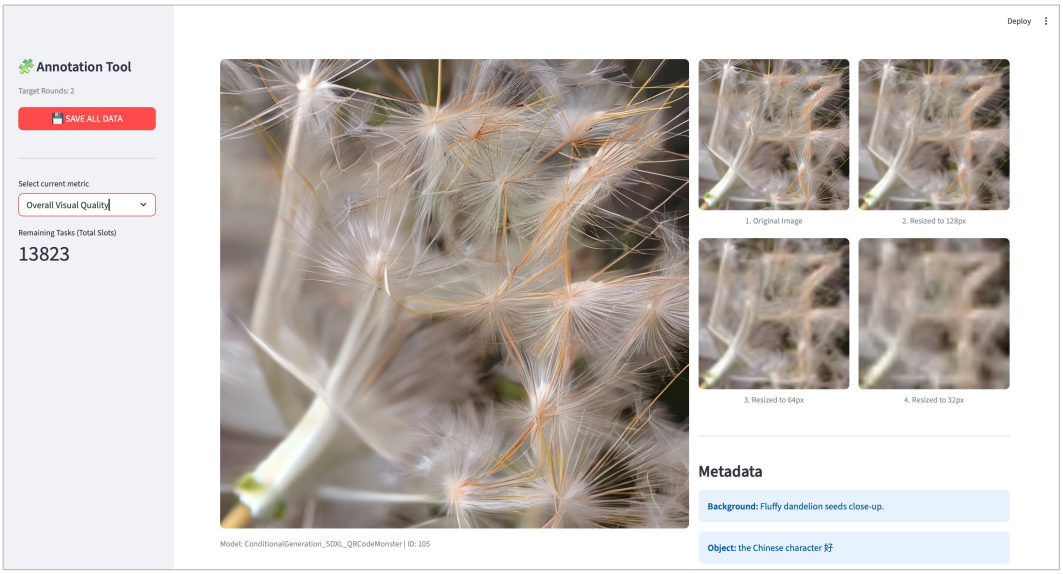

Figure 20: Graphical interface for human annotation in HiP-Bench. Each image is shown with multi-resolution previews at 128×128, 64×64, and 32×32 to simulate distant viewing. The background and hidden-target descriptions are initially collapsed and can be revealed on demand after the initial visual inspection.

We report inter-annotator Spearman correlations computed on a sampled subset of 100 instances in Table 3. Relatively strong positive correlations ranging from 0.71 to 0.81 are observed across all evaluation criteria, suggesting that the assessment dimensions and scoring guidelines can be well understood by the annotators. We also calculate the average weighted kappa value over the sampled instances and obtain a high score of 0.77, demonstrating good agreement among data specialists.

G.2 Definitions of SRCC and PLCC

For correlation calculation, we use two widely-used correlation measures (Zar, 2005): SRCC and PLCC, which measure Spearman Ranking and Pearson Linear Correlation Coefficients between predicted scores and human scores. Correlation coefficients aim to evaluate the reliability of automated metrics by calculating how well automated scores ($x = (x_1, ..., x_n)$) correlate with human judgments ($y = (y_1, ..., y_n)$) using correlation functions. A stronger correlation indicates greater agreement between prediction and reference. Specifically, SRCC measures the monotonic relationship between two variables on the basis of their ranked values, while PLCC quantifies the strength of a linear relationship between two continuous variables.

$$\text{SRCC} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{1}$$

where $d_i = r_{xi} - r_{yi}$, and r_{xi}, r_{yi} denote the ranks of x_i and y_i.

$$\text{PLCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ denote the mean values of x and y, respectively.

H Details of Automated Evaluation

H.1 Construction of Δ VQAScore

VQAScore (Lin et al., 2025) evaluates image-text alignment by computing the probability of a vision-language model answering "yes" to a prompt asking whether an image contains a specific text concept. To automatically quantify Illusion Fidelity (IF, see Table 17) without human annotation, we propose ΔVQAScore.

A successful illusion hides the target at normal resolution but reveals it when down-sampled to simulate distant viewing. Therefore, we evaluate IF by calculating the VQAScore difference between the down-sampled image and the original image.

Let I_{orig} denote the original generated image, I_{low} denote its down-sampled version, and T_{obj} represent the text prompt describing the hidden target. The metric is formulated as:

$$\begin{aligned}\Delta\text{VQAScore} = \text{VQAScore}(I_{low}, T_{obj}) \\ - \text{VQAScore}(I_{orig}, T_{obj})\end{aligned} \tag{3}$$

The value of Equation 3 directly reflects IF.

High ΔVQAScore**:** The hidden target scores higher in the down-sampled image than in the original. This means the illusion works. The target is successfully hidden at normal size but becomes visible when viewed from a distance.

Low or Negative ΔVQAScore**:** The hidden target scores the same in both, or higher in the original image. This means that the concealment fails. The target is clearly visible at normal size and does not blend into the background.

H.2 Prompts for Direct MLLM-as-a-Judge

The prompts for the direct MLLM-as-a-Judge evaluation are provided below. We set the temperature to 0.0 to improve the determinism and reproducibility of the scoring process.

MLLM-as-a-Judge Prompt

System Role: You are an expert evaluator specializing in AI image generation and optical illusions (hidden-picture). Your task is to evaluate a generated optical illusion image based on the provided background description and hidden target description. You will assess the image across four specific dimensions. Assign a score based on the provided criteria and give a brief reason (1-2 sentences) for each.

Input Elements:

- < Background Description >: Description of the intended background scene.
- < Hidden Target Description >: Description of the intended hidden or hidden target.
- < Generated Image >: The AI-generated optical illusion image to be evaluated.

Evaluation Dimensions and Scoring Criteria:

Illusion Fidelity (Score 1-5)
Assess the presentation of the visual illusion under different viewing conditions.

- **5:** Once the illusion is recognized, it is difficult to quickly relocate it by continuous staring when returning to the normal view.
- **4:** The illusion is hard to identify in a normal view and requires specific conditions (e.g., zooming, distancing, or prompt guidance) to be clearly recognized.
- **3:** The image attempts to integrate the illusion, but its outline can be directly observed in a normal view.

- **2:** Illusion content exists, but it feels disconnected from the overall scene, appearing pasted or isolated.
- **1:** The image only presents scene content; no valid illusion can be identified.

Overall Visual Quality (Score 1-5)
Evaluate whether there are obvious abnormalities or unnaturalness in the overall visual presentation.

- **5:** No obvious abnormalities or unnaturalness in the overall visual presentation.
- **4:** Only slight, localized unnatural arrangements caused by deliberate illusion crafting are visible, with limited impact on overall visual appeal.
- **3:** Some textures or shapes are clearly arranged to serve a specific effect, noticeably impacting the visual appeal.
- **2:** Significant image quality issues exist, such as abrupt boundaries, scene disconnections, or sharp drops in local quality, severely ruining the visual appeal.
- **1:** Massive structural damage or geometric distortion; cannot form recognizable, coherent content.

Scene Consistency (Score 0-1)
Evaluate whether the overall scene matches the Background Description.

- **1:** The scene content of the generated image is basically consistent with the background description.
- **0:** The scene content of the generated image is obviously inconsistent with the background description.

Illusion Shape Consistency (Score 1-3)
Evaluate how accurately the illusion content represents the semantics and structure of the Hidden Target Description.

- **3:** The illusion content accurately restores the morphological features of the hidden target description in both semantics and structure.
- **2:** The semantics of the illusion are correct, but its structural performance is non-standard or has obvious deviations.
- **1:** The illusion content is completely unrecognizable, its semantics change, or severe structural distortion occurs.

Output Format: Output ONLY a valid JSON object. Do not include any conversational text, explanations, or markdown formatting (such as "'json). Please use the following JSON template for your output:

JSON Template:

```
{
"Illusion Fidelity": {"score": null, "reason": ""},
"Overall Visual Quality": {"score": null, "reason": ""},
"Scene Consistency": {"score": null, "reason": ""},
"Illusion Shape Consistency": {"score": null, "reason": ""}
}
```

Constraint: Do NOT copy the scoring criteria in your reason. You MUST describe specific reasons from the image.

H.3 Prompts for Ablation Variants

The prompts for the ablation variants are provided below. We set the temperature to 0.0 to improve the determinism and reproducibility of the scoring process.

Prompts for Ablation Variants

System Role: You are an expert evaluator specializing in AI image generation and optical illusions (hidden-picture). Your task is to evaluate a generated optical illusion image based on the provided background description and hidden target description.
[For w/ Single-scale:] You will be provided with two images: the original generated image and a resized version. The resized image is provided to help simulate the effect of humans squinting or viewing from a distance, thereby helping you see the hidden content clearly.
[For w/ Multi-scale:] You will be provided with four images: the original generated image and three resized versions (decreasing in size: 128x128, 64x64, 32x32). The resized images are provided to simulate the effect of a human gradually squinting or viewing the image from increasingly further distances. AI

vision models often process hidden-picture optical illusion differently than human perception; therefore, you must fully utilize these three resized images to help you spot hidden structures and boundaries that are typically only perceptible to the human eye. Please evaluate strictly from the perspective of human visual cognition.
You will assess the image across four specific dimensions. Assign a score based on the provided criteria and give a brief reason (1-2 sentences) for each.

Input Elements:

- < Background Description >: Description of the intended background scene.
- < Hidden Target Description >: Description of the intended hidden or hidden target.
- < Original Generated Image >: The AI-generated optical illusion image to be evaluated.
- *[For w/ Single-scale:]* < Resized Generated Image >: A resized version (64x64) of the generated image to simulate squinting or distant viewing.
- *[For w/ Multi-scale:]* < Resized Generated Images >: three resized versions of the generated image to simulate squinting or distant viewing.

Evaluation Dimensions and Scoring Criteria:

Illusion Fidelity (Score 1-5)
Assess the presentation of the visual illusion under different viewing conditions.

- **5:** Once the illusion is recognized, it is difficult to quickly relocate it by continuous staring when returning to the normal view.
- **4:** The illusion is hard to identify in a normal view and requires specific conditions (e.g., zooming, distancing, or prompt guidance) to be clearly recognized.
- **3:** The image attempts to integrate the illusion, but its outline can be directly observed in a normal view.
- **2:** Illusion content exists, but it feels disconnected from the overall scene, appearing pasted or isolated.
- **1:** The image only presents scene content; no valid illusion can be identified.

Illusion Shape Consistency (Score 1-3)
Evaluate how accurately the illusion content represents the semantics and structure of the Hidden Target Description.

- **3:** The illusion content accurately restores the morphological features of the hidden target description in both semantics and structure.
- **2:** The semantics of the illusion are correct, but its structural performance is non-standard or has obvious deviations.
- **1:** The illusion content is completely unrecognizable, its semantics change, or severe structural distortion occurs.

Output Format: Output ONLY a valid JSON object. Do not include any conversational text, explanations, or markdown formatting (such as "'json). Please use the following JSON template for your output:

JSON Template:

```
{
"Illusion Fidelity": {"score": null, "reason": ""},
"Illusion Shape Consistency": {"score": null, "reason": ""}
}
```

Constraint: Do NOT copy the scoring criteria in your reason. You MUST describe specific reasons from the image.

H.4 Prompts for MLLM with MDS

The Multi-scale Distance Simulation (MDS) strategy involves two turns: Turn 1 evaluates the original image using prompts in Appendix H.2, and Turn 2 applies the correction prompts below. We set the temperature to 0.0 to improve the determinism and reproducibility of the scoring process.

MLLM with MDS (Turn 2) Prompt

Here are three resized versions of the previously evaluated image (decreasing in size: 128x128, 64x64, 32x32). These resized images are provided to simulate the effect of a human gradually squinting or viewing the image from increasingly further distances.
AI vision models often process hidden-picture optical illusion differently than human perception; therefore, you must fully utilize these three resized images to help you spot hidden structures and boundaries that are typically only perceptible to the human eye. Please evaluate strictly from the perspective of human visual cognition.
Review your previous score and reason. If a correction is exactly necessary based on these new distant views (for example, if the hidden target is invisible in the original view but emerges in the resized views, meaning it's a highly effective concealment, or if the global morphology becomes much clearer in the smaller images, masking the minor generative flaws.), please revise your judgment.

Output Format: Output ONLY a valid JSON object. Do not include any conversational text, explanations, or markdown formatting (such as "'json). Please use the following JSON template for your output:

JSON Template:

```
{
"Illusion Fidelity": {"score": null, "reason": ""},
"Illusion Shape Consistency": {"score": null, "reason": ""}
}
```

Constraint: Do NOT copy the scoring criteria in your reason. You MUST describe specific reasons from the image.

I Evaluation Details, Full Results, Statistical Tests, and Extended Analyses

I.1 Full Results

Tables 18, 19, and 20 detail the full evaluation results for all evaluated models.

Model	IF	VQ	SC	ISC
U-Net	1.27	<u>4.49</u>	0.94	1.51
SD 1.5	1.10	4.59	0.90	1.19
SDXL Base	1.29	**4.64**	**0.96**	1.56
DeepFloyd IF (XL-L-x4)	**1.43**	4.23	0.95	**1.78**
Unified Multimodal	1.43	4.03	0.96	1.78
Janus-Pro (7B)	1.52	3.83	**0.99**	1.87
JanusFlow	1.06	4.05	0.90	1.10
OmniGen v1	1.46	**4.38**	0.98	1.81
OmniGen2	**1.70**	3.85	0.95	**2.33**
Diffusion Transformer	<u>1.58</u>	4.36	<u>0.99</u>	<u>2.05</u>
FLUX.1 [dev]	1.68	4.45	**0.99**	2.25
SD 3.5 Large	**1.71**	4.31	**0.99**	**2.31**
SD 3.5 Large Turbo	1.68	4.19	**0.99**	2.25
SD 3.5 Medium	1.58	4.31	**0.99**	2.01
Hunyuan-DiT v1.2	1.42	**4.58**	0.97	1.82
PixArt-Σ	1.48	4.42	**0.99**	1.87
Lumina-Image 2.0	1.52	4.25	**0.99**	1.85
Overall Average	1.47	4.29	0.97	1.86

Table 18: Full text-to-image evaluation results. Bold entries within each model family indicate the best-performing model for that metric. Rows with bold family names report averages over models in the corresponding family, and underlined entries indicate the best family-level average.

Model	IF	VQ	SC	ISC
InstructPix2Pix (on SD 1.5)	1.18	4.54	0.82	1.34
CosXL Edit (on SDXL)	1.22	**4.70**	0.96	1.42
FLUX.1 Kontext (on FLUX.1 [dev])	1.26	4.52	**0.99**	1.49
OmniGen v1	1.25	3.74	0.58	1.44
OmniGen2	**1.58**	3.96	0.85	**2.09**
Overall Average	1.30	4.29	0.84	1.56

Table 19: Full instruction-based image editing evaluation results. Bold entries indicate the best result for each metric among all evaluated editing models.

Model	IF	VQ	SC	ISC
SD 1.5 + Depth	1.91	3.24	0.89	2.61
SD 1.5 + Tile	2.10	3.35	0.72	2.49
SD 1.5 + QR Code Monster v2	1.76	**3.90**	0.91	1.97
SDXL Base + Depth	1.96	3.29	0.95	2.59
SDXL Base + QR Code Monster	**2.42**	3.80	**0.98**	2.29
SD 3.5 Large + Depth	2.04	3.09	0.82	2.62
FLUX.1 [dev] + Depth	1.86	3.20	0.86	**2.68**
PTDiffusion (on SD 1.5)	2.16	3.34	0.85	2.04
Overall Average	2.03	3.40	0.87	2.41

Table 20: Full image-conditioned generation evaluation results. Bold entries indicate the best result for each metric among all evaluated image-conditioned models.

I.2 Uncertainty Estimates and Significance Tests

We conduct uncertainty estimation and significance testing for RQ1–RQ3 using the 512 benchmark pairs. For RQ1 and RQ2, scores are first averaged across the models in each compared group for every benchmark pair. For RQ3, scores are averaged across all 27 evaluated models for every pair. We estimate 95% confidence intervals using 10,000 stratified bootstrap resamples, with the 16 background–hidden target combinations as strata. RQ1 and RQ2 use paired randomization tests over benchmark pairs. RQ3 uses Freedman–Lane permutation tests for planned comparisons and a two-way permutation ANOVA for the background–target interaction. All tests use 50,000 permutations. We apply Holm correction across all comparisons within each RQ and report the resulting adjusted p-values. A positive difference (Δ) indicates that the first group has a higher score than the second group.

RQ1: Generative Paradigm Analysis Table 21 confirms the differences among the three paradigms in RQ1. CondGen achieves higher IF and ISC but lower VQ than both T2I and Editing. Its SC is lower than T2I but higher than Editing. Editing therefore has the lowest SC, while T2I has lower IF and ISC than CondGen. All reported differences remain significant after Holm correction.

RQ2: Intra-Paradigm Variations Table 22 supports the main comparisons in RQ2. Among the evaluated T2I models, the DiT group achieves higher IF and ISC but slightly lower VQ than the U-Net group. The unified multimodal group also exceeds the U-Net group in IF and ISC, but has lower VQ than both groups. Native editing models achieve higher IF but lower VQ than T2I-derived editors, consistent with the qualitative editing behaviors discussed in the main text. SDXL + QR Code Monster exceeds the mean of the other CondGen models in IF, VQ, and SC. SD1.5 + Tile achieves higher IF but lower SC than SD1.5 + Depth, while PTDiffusion also achieves higher IF than SD1.5 + Depth. Within the SD 3.5 family, Large achieves higher IF than Medium. The SDXL + QR Code Monster also exceeds SD1.5 + QR Code Monster V2 in IF. All reported differences remain significant after Holm correction.

RQ3: Attribute Influence Analysis Table 23 supports the attribute findings in RQ3. High-Frequency textures achieve higher IF than the average of the other backgrounds, but their differ-

ence from Directional textures is not significant. Directional textures have higher VQ than High-Frequency textures. As groups, High-Frequency and Directional textures achieve higher IF than Low-Frequency and Structured textures. Low-Frequency textures have higher VQ than the other backgrounds, whereas Structured textures have lower VQ. Physical targets show a small descriptive advantage over abstract targets, but this difference does not remain significant after Holm correction. The selected combination differences are also not significant after correction, and no background–target interaction is detected.

Comparison	**Metric**	**Δ (95% CI)**	p_{Holm}
CondGen − T2I	IF	$0.553\ [0.514,\ 0.592]$	< 0.001
CondGen − T2I	VQ	$-0.890\ [-0.956,\ -0.824]$	< 0.001
CondGen − T2I	SC	$-0.093\ [-0.105,\ -0.081]$	< 0.001
CondGen − T2I	ISC	$0.553\ [0.501,\ 0.604]$	< 0.001
CondGen − Editing	IF	$0.729\ [0.691,\ 0.767]$	< 0.001
CondGen − Editing	VQ	$-0.890\ [-0.961,\ -0.818]$	< 0.001
CondGen − Editing	SC	$0.033\ [0.015,\ 0.050]$	< 0.001
CondGen − Editing	ISC	$0.855\ [0.805,\ 0.904]$	< 0.001
T2I − Editing	SC	$0.126\ [0.112,\ 0.140]$	< 0.001

Table 21: Statistical comparisons between the three generative paradigms. Confidence intervals are obtained by stratified bootstrap. The reported p-values are adjusted across all RQ1 comparisons using the Holm method.

Comparison	**Metric**	**Δ (95% CI)**	p_{Holm}
Architecture			
DiT − U-Net	IF	$0.313\ [0.284,\ 0.341]$	< 0.001
DiT − U-Net	VQ	$-0.131\ [-0.204,\ -0.060]$	< 0.001
DiT − U-Net	ISC	$0.545\ [0.495,\ 0.595]$	< 0.001
Unified Multimodal − U-Net	IF	$0.164\ [0.137,\ 0.191]$	< 0.001
Unified Multimodal − U-Net	ISC	$0.268\ [0.222,\ 0.315]$	< 0.001
U-Net − Unified Multimodal	VQ	$0.463\ [0.387,\ 0.540]$	< 0.001
DiT − Unified Multimodal	VQ	$0.332\ [0.266,\ 0.398]$	< 0.001
Editing behavior			
Native Editing − T2I-derived Editing	IF	$0.197\ [0.163,\ 0.231]$	< 0.001
Native Editing − T2I-derived Editing	VQ	$-0.735\ [-0.830,\ -0.640]$	< 0.001
Control mechanism			
SDXL + QR Code Monster − other CondGen models	IF	$0.444\ [0.343,\ 0.542]$	< 0.001
SDXL + QR Code Monster − other CondGen models	VQ	$0.456\ [0.362,\ 0.552]$	< 0.001
SDXL + QR Code Monster − other CondGen models	SC	$0.123\ [0.108,\ 0.139]$	< 0.001
SD1.5 + Tile − SD1.5 + Depth	IF	$0.190\ [0.099,\ 0.280]$	< 0.001
SD1.5 + Depth − SD1.5 + Tile	SC	$0.171\ [0.138,\ 0.205]$	< 0.001
PTDiffusion − SD1.5 + Depth	IF	$0.251\ [0.140,\ 0.361]$	< 0.001
SDXL + QR Code Monster − PTDiffusion	VQ	$0.459\ [0.328,\ 0.592]$	< 0.001
Model scaling and configuration			
SD 3.5 Large − SD 3.5 Medium	IF	$0.131\ [0.081,\ 0.180]$	< 0.001
SDXL + QR Code Monster − SD1.5 + QR Code Monster V2	IF	$0.656\ [0.542,\ 0.769]$	< 0.001

Table 22: Statistical comparisons for the intra-paradigm findings in RQ2. "Other CondGen models" denotes the mean of the other seven evaluated CondGen models. The reported p-values are Holm-adjusted across all RQ2 comparisons.

Comparison	Metric	Δ (95% CI)	p_{Holm}
Background category			
High-Frequency − Directional	IF	0.043 [−0.003, 0.090]	0.209
Directional − High-Frequency	VQ	0.258 [0.140, 0.376]	< 0.001
High-Frequency − other backgrounds	IF	0.083 [0.045, 0.122]	< 0.001
High-Frequency/Directional − Low-Frequency/Structured	IF	0.082 [0.049, 0.114]	< 0.001
Low-Frequency − other backgrounds	VQ	0.260 [0.173, 0.348]	< 0.001
Structured − other backgrounds	VQ	−0.379 [−0.465, −0.291]	< 0.001
Target category			
Physical − Abstract targets	IF	0.037 [0.005, 0.070]	0.105
Background–target combination			
Directional + Physical − other combinations	IF	0.062 [0.016, 0.107]	0.072
Structured + Abstract − other combinations	IF	−0.045 [−0.090, 0.002]	0.209
Background × Target interaction	IF	−−	0.484

Table 23: Statistical comparisons for the attribute findings in RQ3. Physical targets include Rigid and Non-Rigid Objects; abstract targets include Textual Elements and Geometry & Shapes. The reported p-values are Holm-adjusted across all RQ3 comparisons.

I.3 Extended Analysis on Attribute Influence

While the main text (RQ3: Attribute Influence Analysis) discusses the general trends averaged across all models, breaking down the results by generative paradigm reveals distinct behaviors. The performance matrices are presented in Figures 21, 22, and 23.

(1) Text-to-Image. Results in this paradigm (Figure 21) align with the overall trends in the main text. High-Frequency and Directional textures are most effective for hiding targets, whereas Structured textures and Textual Elements consistently yield the lowest scores across all metrics.

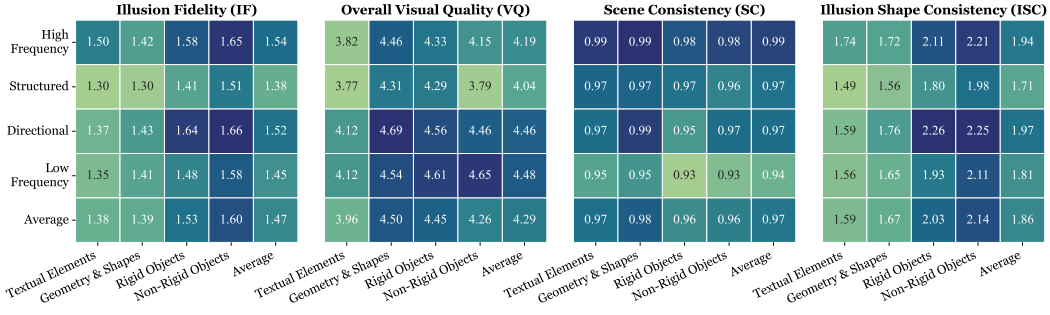

Figure 21: Performance across 16 combinations of background textures and hidden target categories for the text-to-image paradigm. Darker colors indicate higher scores.

(2) Instruction-Based Image Editing. This paradigm (Figure 22) shows distinct deviations from the general trends.

(i) Unlike the general trend, only High-Frequency textures effectively hide targets here (IF: 1.36). However, they yield the lowest VQ (4.11) and SC (0.82), indicating that editing models struggle to modify complex textures without altering the original reference image.

(ii) Directional textures drop to a low IF (1.27) in this paradigm, performing similarly to Low-Frequency textures. Yet, they maintain a high VQ (4.42) and the highest SC (0.88), suggesting editing models prioritize preserving the original structural flow over embedding the hidden target.

(iii) High-Frequency textures paired with Rigid Objects perform best (IF: 1.46, ISC: 1.89), replacing the directional combinations seen in the general trend.

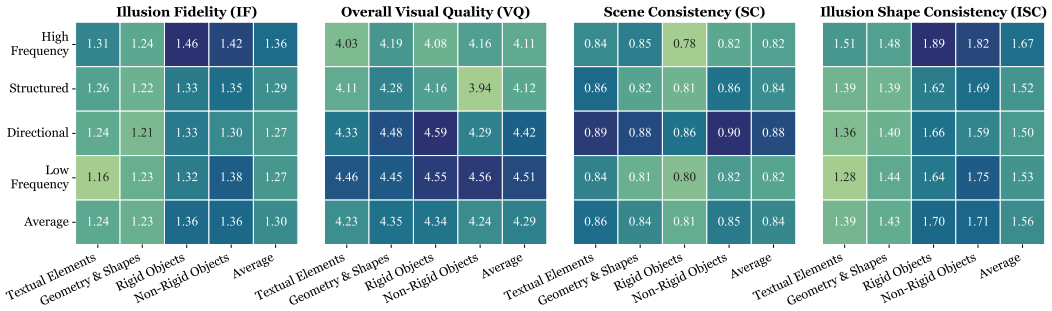

Figure 22: Performance across 16 combinations of background textures and hidden target categories for the instruction-based image editing paradigm. Darker colors indicate higher scores.

(3) Image-Conditioned Generation. Explicit spatial controls (Figure 23) alter the attribute interactions.

(i) High-Frequency, Structured, and Directional textures all achieve high IF (>2.02), leaving only Low-Frequency as ineffective. However, this forced spatial control reduces overall image quality, particularly on Structured textures, where enforcing exact alignment on regular patterns creates obvious artifacts.

(ii) Reversing the general trend, abstract shapes (Textual Elements and Geometry & Shapes) achieve higher IF and ISC than physical objects (Rigid and Non-Rigid Objects). Because spatial adapters force exact outlines into the image, these shapes with strict boundaries are rendered more accurately.

(iii) Unlike editing models, conditional models maintain the highest SC (0.94) on High-Frequency textures. The dense details of these textures accommodate forced structural boundaries better than smooth or regular patterns.

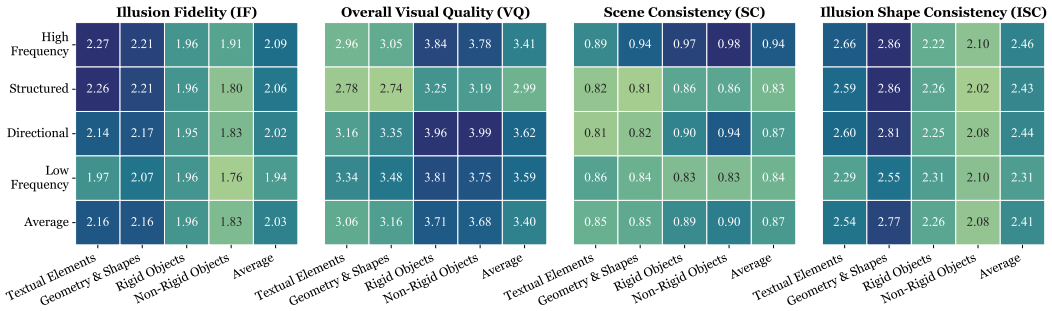

Figure 23: Performance across 16 combinations of background textures and hidden target categories for the image-conditioned generation paradigm. Darker colors indicate higher scores.

I.4 Failure Mode Assignment Rules

Table 24 lists the criteria used to assign the eight failure modes in RQ4 from the human annotation scores for IF, VQ, SC, and ISC. The criteria are evaluated in severity order, from Type 1 to Type 8, and each annotation set is assigned to the first matching type. Each image is scored by two human annotators on each dimension, yielding two annotation sets: $(\mathrm{IF}_1, \mathrm{VQ}_1, \mathrm{SC}_1, \mathrm{ISC}_1)$ and $(\mathrm{IF}_2, \mathrm{VQ}_2, \mathrm{SC}_2, \mathrm{ISC}_2)$.

We classify the two annotation sets independently and count both results. For each model or paradigm, the percentage of each failure mode is the number of annotation sets assigned to that type divided by the number assigned to any of the eight types. Annotation sets that match none of the eight types are excluded. Two annotators second-stage manual audit each candidate assignment. Disagreements were resolved by discussion, and the final statistics were computed from the verified assignments.

Failure Mode	Condition
(1) Severe Structural Corruption	$\mathrm{VQ} = 1$
(2) Scene Semantic Mismatch	$\mathrm{SC} = 0$
(3) Concept Omission	$\mathrm{IF} = 1$
(4) Illusion Structural Deviation	$\mathrm{ISC} = 2$
(5) Blatant Pasting Effect	$\mathrm{IF} = 2,\ \mathrm{VQ} \in \{2, 3\}$
(6) Independent Object Manifestation	$\mathrm{IF} = 2,\ \mathrm{VQ} \in \{4, 5\}$
(7) Traceable Manipulation	$\mathrm{VQ} = 3$
(8) Insufficient Concealment	$\mathrm{IF} = 3$

Table 24: Criteria for failure mode classification in RQ4. The criteria are evaluated in the listed order.

I.5 Extended Analysis on Failure Modes

Tables 25, 26, and 27 present the detailed error rates of failure modes defined in RQ4: Failure Mode Analysis.

The results in Table 25 show that: (i) Concept Omission (Type 3) is the most prevalent error across all models (52.36% average), peaking in Legacy U-Net models (68.25% average). This indicates a general tendency to ignore concealment instructions when generating scenes. (ii) Independent Object Manifestation (Type 6) is the second most common failure (24.28% average). Notably, Diffusion Transformers exhibit significantly higher rates of this error (32.01% average) compared to U-Nets (14.69% average). This suggests that while newer architectures better recognize the hidden target, they tend to render it as a distinct physical entity rather than embedding it into the background texture.

Model	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8
Legacy U-Net	2.28	5.80	**68.25**	2.22	6.16	14.69	0.13	0.49
SD 1.5	4.30	8.59	**79.39**	0.88	0.98	5.86	0.00	0.00
SDXL Base	1.76	3.81	**69.24**	0.39	5.47	18.55	0.00	0.78
DeepFloyd IF (XL-L-x4)	0.78	4.99	**56.11**	5.38	12.02	19.65	0.39	0.68
Unified Multimodal	3.22	3.89	**53.72**	5.23	14.36	17.95	0.27	1.37
Janus-Pro (7B)	2.74	1.37	**50.64**	9.29	16.81	15.84	0.59	2.74
JanusFlow	9.96	7.13	**78.42**	1.37	0.78	2.25	0.00	0.10
OmniGen v1	0.20	2.25	**55.96**	6.35	12.01	22.17	0.00	1.07
OmniGen2	0.00	4.79	29.88	3.91	27.83	**31.54**	0.49	1.56
Diffusion Transformer	0.36	1.42	**44.76**	4.55	14.19	32.01	0.45	2.25
FLUX.1 [dev]	0.20	0.78	36.23	2.54	14.45	**41.50**	0.39	3.91
SD 3.5 Large	0.00	1.37	32.39	3.72	18.30	**41.00**	0.68	2.54
SD 3.5 Large Turbo	0.20	1.37	34.86	4.69	20.02	**36.33**	0.78	1.76
SD 3.5 Medium	0.00	0.88	**45.89**	7.24	14.19	29.35	0.49	1.96
Hunyuan-DiT v1.2	0.20	2.83	**56.74**	2.93	8.40	28.52	0.00	0.39
PixArt-Σ	1.17	1.37	**54.30**	3.42	11.13	26.27	0.00	2.34
Lumina-Image 2.0	0.78	1.37	**52.94**	7.34	12.82	21.14	0.78	2.84
Overall Average	1.59	3.06	**52.36**	4.24	12.51	24.28	0.33	1.62

Table 25: Detailed error rates (%) of failure modes for the text-to-image paradigm. Failure mode definitions are provided in RQ4: Failure Mode Analysis. Bold values indicate the highest error rate among all failure types for each model.

The results in Table 26 show that: (i) Concept Omission (Type 3) dominates the errors (60.88% average), exceeding 70% in diffusion-based editing models like InstructPix2Pix, CosXL Edit, and FLUX.1 Kontext. This indicates that these models fail to execute concealment instructions, often leaving the original image unchanged. (ii) Unified Multimodal models show a different error distribution. While OmniGen2 lowers Concept Omission to 40.12%, the Blatant Pasting Effect (Type 5, 22.02%) and Independent Object Manifestation (Type 6, 16.63%) increase. This reveals a trade-off: generating the hidden target often comes at the expense of concealing it effectively.

Model	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8
InstructPix2Pix (on SD 1.5)	2.34	16.31	**70.12**	0.88	4.39	5.96	0.00	0.00
CosXL Edit (on SDXL)	0.39	4.00	**76.07**	1.66	4.98	12.60	0.00	0.29
FLUX.1 Kontext (on FLUX.1 [dev])	0.20	1.17	**73.73**	2.93	12.50	9.38	0.00	0.10
OmniGen v1	19.14	25.98	**44.34**	2.44	3.42	4.20	0.20	0.29
OmniGen2	0.20	15.26	**40.12**	5.38	22.02	16.63	0.00	0.39
Overall Average	4.45	12.54	**60.88**	2.66	9.46	9.75	0.04	0.21

Table 26: Detailed error rates (%) of failure modes for the instruction-based image editing paradigm. Failure mode definitions are provided in RQ4: Failure Mode Analysis. Bold values indicate the highest error rate among all failure types for each model.

The results in Table 27 show that: (i) The Blatant Pasting Effect (Type 5) is the primary failure (30.82% average), consistently exceeding 35% in Depth-guided models. This indicates that enforcing strict geometric boundaries degrades natural blending. (ii) Compared to Depth, QR Code Monster models reduce the Pasting Effect (Type 5) but exhibit higher Concept Omission (Type 3, 34.92%–43.72%), Traceable Manipulation (Type 7), and Insufficient Concealment (Type 8). This indicates that looser spatial guidance improves blending but fails to enforce the hidden target's structural integrity.

Model	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8
SD 1.5 + Depth	0.59	10.97	16.16	6.17	**42.61**	18.81	1.08	3.62
SDXL Base + Depth	2.56	4.63	17.42	6.10	**42.72**	18.21	2.56	5.81
SD 3.5 Large + Depth	1.77	16.91	12.88	10.32	**35.10**	15.14	1.67	6.19
FLUX.1 [dev] + Depth	0.59	13.96	13.77	1.86	**48.93**	20.12	0.29	0.49
SD 1.5 + QR Code Monster v2	1.81	8.14	**43.72**	9.05	17.99	10.75	2.51	6.03
SDXL Base + QR Code Monster	3.31	1.66	**34.92**	9.17	17.90	6.63	12.15	14.25
SD 1.5 + Tile	4.09	**27.53**	19.34	6.35	23.75	11.26	2.87	4.81
PTDiffusion (on SD 1.5)	11.47	11.25	**37.60**	6.73	13.56	2.21	10.69	6.50
Overall Average	3.15	11.97	24.10	6.93	**30.82**	13.15	4.03	5.84

Table 27: Detailed error rates (%) of failure modes for the image-conditioned generation paradigm. Failure mode definitions are provided in RQ4: Failure Mode Analysis. Bold values indicate the highest error rate among all failure types for each model.

I.6 Ablation Study of Automated Evaluation

We conduct an ablation study on GPT-5.4 (OpenAI, 2026) and Gemini-3-Pro (Google DeepMind, 2026) to evaluate the two core components of MDS: multi-scale visual inputs and two-turn prompting. We compare four configurations: (1) **Direct (Original)**: single-turn evaluation using only the original image; (2) **w/ Single-scale**: single-turn evaluation with the original and one down-sampled image; (3) **w/ Multi-scale**: single-turn evaluation with the original and three down-sampled images; (4) **MDS (Ours)**: two-turn evaluation, where the model first scores the original image, then refines its judgment using the three down-sampled images. The exact prompts used for all configurations are detailed in Appendices H.2, H.3, and H.4.

Table 28 shows that providing down-sampled images improves the correlation on IF and ISC, indicating that distant viewing helps MLLMs recognize hidden structures. However, using multiple scales in a single turn (w/ Multi-scale) decreases the IF correlation for GPT-5.4 compared to using a single scale (w/ Single-scale). This suggests that processing multiple images at once reduces the model's ability to evaluate concealment. MDS avoids this issue by separating the initial evaluation from the multi-scale refinement, thereby achieving the highest overall correlation for both models.

Variant	GPT-5.4		Gemini-3-Pro	
	IF	ISC	IF	ISC
Direct (Original)	3.2 / -0.4	39.0 / 40.9	39.9 / 37.5	48.9 / 50.3
w/ Single-scale	31.1 / 26.7	40.4 / 43.9	50.4 / 47.0	53.2 / 51.7
w/ Multi-scale	15.2 / 11.8	45.4 / 49.7	53.6 / 51.6	59.5 / 59.4
MDS (Ours)	**32.5 / 31.1**	**49.4 / 51.0**	**53.8 / 53.1**	**63.2 / 62.5**

Table 28: Ablation study of the MDS strategy for automated evaluation. We report correlations (PLCC / SRCC, % ↑) with human judgments on IF and ISC.

J Future Use of HiP-Bench

We recommend different evaluation protocols according to the purpose of the evaluation and the available annotation resources.

Human evaluation. Human evaluation remains the reference protocol for HiP-Bench, especially for the illusion-specific dimensions (IF and ISC) and fine-grained model comparison. Future users should follow the same scoring criteria and annotation procedure described in Appendix G.1. Human evaluation is recommended for final model rankings, comparisons between similarly performing models, and claims based on small performance differences.

Automatic evaluation. MDS provides a lower-cost and reproducible option for preliminary evaluation of IF and ISC. It can be applied to the full test set for initial model comparison, large-scale analysis, and screening before human evaluation. However, MDS is not intended to replace human evaluation. Results obtained with MDS should be clearly reported as automatic scores and should not be treated as equivalent to human scores.

Hybrid evaluation. When full human evaluation is not feasible, users may apply MDS to the full evaluation set and conduct human evaluation on a subset that covers the background and hidden-target categories in HiP-Bench. When comparing multiple models, the same subset should be used for each model. The automatic and human results should be reported separately rather than combined into a single score. In this setting, MDS supports broad comparison, while human evaluation is used to confirm the conclusions.