MIRRORREFINE: MIRROR REFLECTION GENERATION VIA MULTI-AGENT REFINEMENT WITH SPATIAL PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Advancing visual generation toward world modeling requires more than photorealism: generated views must remain consistent with the underlying scene. In mirror-reflection synthesis, this requires inferring reflected content and spatial layout from the source scene, beyond the text–image alignment emphasized by general-purpose prompt-refinement methods. We propose MirrorRefine, a multi-agent framework for mirror-reflection generation through multi-action evolutionary refinement. Starting from a generated or supplied image, the Reflection Optimization Controller analyzes the original request and visible content outside the mirror to infer intended reflected content and construct action-specific synthesis conditions, rather than treating potentially erroneous reflections as reliable targets. The Quality Evaluator provides a vision–language model with coarse reflection-layout priors predicted by an offline-trained Reflection Spatial Geometry (RSG) module. Together with textual conditions, these priors guide the diagnosis of geometric, appearance, and illumination inconsistencies in the synthesized image. Based on this diagnosis, the evaluator selects among image generation, mirror-region inpainting, and targeted editing, and returns structured revision suggestions to the controller. The controller updates the synthesis conditions and coordinates the selected operation, after which the resulting image is reassessed to guide subsequent refinement. All component parameters, including those of RSG, remain fixed during iterative refinement. Experiments on MirrorBenchV2, MSD, and CC-Mirror show complementary strengths: among the compared systems, the online configuration leads the reported geometry, lighting, and appearance scores, while the local configuration achieves the highest PSNR on MSD and CC-Mirror.

1 INTRODUCTION

World models aim to learn predictive representations of environments from observations (Ha & Schmidhuber, 2018). For visual generation, an important aspect of this goal is synthesizing views that remain consistent with the same underlying scene, rather than merely producing individually plausible images (Eslami et al., 2018). Mirror-reflection synthesis offers a concrete setting for studying this requirement: direct and reflected views within an image must depict a shared scene while respecting object identity, mirror configuration, viewpoint, and illumination (Ge et al., 2026; Pharr et al., 2016). Yet text-to-image (T2I) models can produce visually plausible scenes with reflections that mismatch source-object appearance or fine-grained attributes, as illustrated in Figure 1. Other failures include incompatible locations, implausible orientations, and missing expected content (Dhiman et al., 2025a; Ge et al., 2026; Mai et al., 2026). Moreover, the reflected view can reveal surfaces not visible in the direct view, so consistency cannot generally be achieved by copying or horizontally flipping the source appearance (Ge et al., 2026; Pharr et al., 2016). Generating realistic mirror reflections therefore requires inferring both what should appear in the mirror and where it should appear from the available scene evidence, rather than merely synthesizing plausible-looking content.

To improve reflection consistency, recent studies commonly formulate mirror-reflection generation as an image inpainting task. MirrorFusion and MirrorVerse train on large-scale synthetic data and use depth conditioning to improve reflection realism (Dhiman et al., 2025b;a), while PhysReflect

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

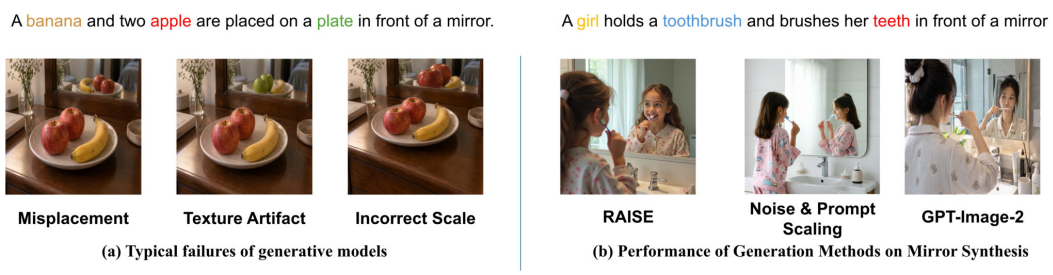

Figure 1: **Semantic plausibility does not guarantee reflection consistency.** Current T2I models can generate visually plausible scenes while producing reflections inconsistent with their source objects, including mismatched object appearance (a) and fine-grained human attributes (b). Such failures motivate reasoning about both what should appear in the mirror and where it should appear.

introduces geometric and perceptual supervision during diffusion training (Ge et al., 2026). Fill My Mirror and PhysMirror derive geometric conditions from reconstructed or constructed three-dimensional scenes to guide mirror-reflection synthesis (Basson et al., 2026; Mai et al., 2026). Methods that train diffusion components require backbone-specific adaptation, while explicit geometric guidance relies on compatible conditioning interfaces, which can complicate direct reuse across generators (Dhiman et al., 2025b; Basson et al., 2026). Accurate geometric guidance can also be difficult to obtain from a single image in complex real-world scenes, where depth, mirror pose, and occlusion relationships cannot always be reliably recovered. Estimation errors can propagate into reflected-position and structural deviations, limiting applicability in these settings (Ge et al., 2026; Basson et al., 2026).

Agent-based reasoning and visual feedback offer a complementary route to improving image synthesis through iterative prompt revision and image editing (Khan et al., 2025; Wu et al., 2026; 2024; Jiang et al., 2026; Kovalev et al., 2025). These methods primarily target semantic alignment, such as recovering missing objects and correcting attributes and counts (Wu et al., 2026; 2024; Kovalev et al., 2025). However, mirror-reflection errors are inherently relational: reflected content must be consistent with the corresponding source objects, mirror configuration, and viewpoint, rather than merely appear plausible in isolation. Consequently, improving text–image alignment alone does not resolve what should appear in the mirror or where it should appear. This motivates combining scene-based content inference with reflection-specific spatial guidance for error diagnosis and repair.

To address these relational constraints, we propose MirrorRefine, a multi-agent framework for mirror-reflection generation through multi-action evolutionary refinement (Figure 2). Rather than treating the existing mirror content as a reliable target, MirrorRefine reasons from the textual request and visible source scene to infer what should be reflected and uses reflection-layout priors to help determine where the reflected content should appear. These cues are integrated with multidimensional quality evaluation and multi-action refinement to guide successive corrections. Its two core components are the Reflection Optimization Controller and a Quality Evaluator guided by reflection spatial geometry. The controller structures the textual request and visible scene information outside the mirror, including source-object attributes, mirror boundaries, and cues about viewpoint and illumination, to infer the intended reflected content. These inferences serve as contextual constraints for the mirror region. For each selected action, the controller constructs synthesis prompts and instructions for generation (Action G), inpainting (Action I), or editing (Action E).

To provide complementary spatial guidance, the Quality Evaluator incorporates an offline-trained Reflection Spatial Geometry (RSG) module, which predicts coarse reflection bounding-box regions within the mirror from an RGB image with its mirror interior erased and a binary mirror mask. The evaluator supplies these priors and textual conditions to a vision–language model (VLM) to assess reflection geometry, appearance, and illumination. Based on this assessment, it selects the next refinement action and returns structured revision suggestions to the controller. The controller updates the synthesis conditions and executes the selected operation, after which the resulting image is reassessed to guide subsequent refinement. All component parameters remain fixed during this process.

Our main contributions are as follows.

- We propose a **multi-agent framework for mirror-reflection generation through multi-action evolutionary refinement**, improving the realism of synthesized mirror reflections through multiple refinement rounds.
- We develop **scene-based reflection content inference** that infers the intended reflected content from textual descriptions and visible scene information outside the mirror, providing reflection-specific content constraints for synthesis.
- We introduce **Reflection Spatial Geometry (RSG)**, which predicts a coarse reflection-layout prior from a mirror-interior-erased image and a mirror mask to guide error diagnosis and repair localization.
- We evaluate MirrorRefine through **benchmark comparisons, a user study, and controlled ablations** assessing completion quality, reflection consistency, human preference, component contributions, and synthesis-attempt counts.

2 Related Work

2.1 Mirror-reflection generation

MirrorFusion learns depth-conditioned completion from synthetic data by training a conditioning branch with a frozen generation U-Net (Dhiman et al., 2025b); MirrorVerse expands scene diversity and introduces a curriculum spanning single-object, multi-object, and real-image settings (Dhiman et al., 2025a). PhysReflect combines geometric and perceptual supervision during diffusion training, including projection, semantic, and illumination constraints (Ge et al., 2026). Explicit geometry offers another route: Fill My Mirror projects estimated scene content into the mirror and completes it through two-mask diffusion (Basson et al., 2026), while PhysMirror constructs object–mirror arrangements and renders depth conditions from text (Mai et al., 2026). MirrorRefine instead combines source-scene content inference with a learned coarse layout prior, without requiring complete three-dimensional reconstruction at inference.

2.2 Agent-based generation and iterative refinement

TIR uses multimodal critique and refinement history to revise prompts (Khan et al., 2025); VisualPrompter expands missing semantic concepts (Wu et al., 2026); CRAFT uses dependency-structured visual questions for targeted prompt updates (Kovalev et al., 2025). Image-level methods turn feedback into editing: SLD combines detected layouts with LLM corrections to add, remove, relocate, or modify objects (Wu et al., 2024), and GenArtist coordinates tools through a planning tree, intermediate verification, and backtracking (Wang et al., 2024). RAISE integrates requirement analysis, adaptive actions, and tool-supported verification (Jiang et al., 2026). M3 assigns refinement roles to specialized agents and compares edited candidates with retained images (Yang et al., 2026). These systems address user-specified objects, attributes, and relations. Mirror generation additionally requires inferring a reflected view consistent with the visible source scene. Our agents use reflection-specific content and spatial evidence to diagnose inconsistencies, select mirror-region operations, and evaluate their outcomes.

2.3 Spatial guidance for image generation and repair

ControlNet learns conditioning branches for depth, edges, and segmentation while freezing the diffusion backbone (Zhang et al., 2023). LLM-grounded Diffusion derives captioned layouts from text and guides pretrained diffusion without parameter updates (Lian et al., 2024). BrushNet encodes masked-image features in a separate branch for inpainting (Ju et al., 2024). RSG supplies a mirror-specific coarse layout inferred from the visible source scene, guiding error diagnosis and repair localization rather than requiring an exact reflected contour.

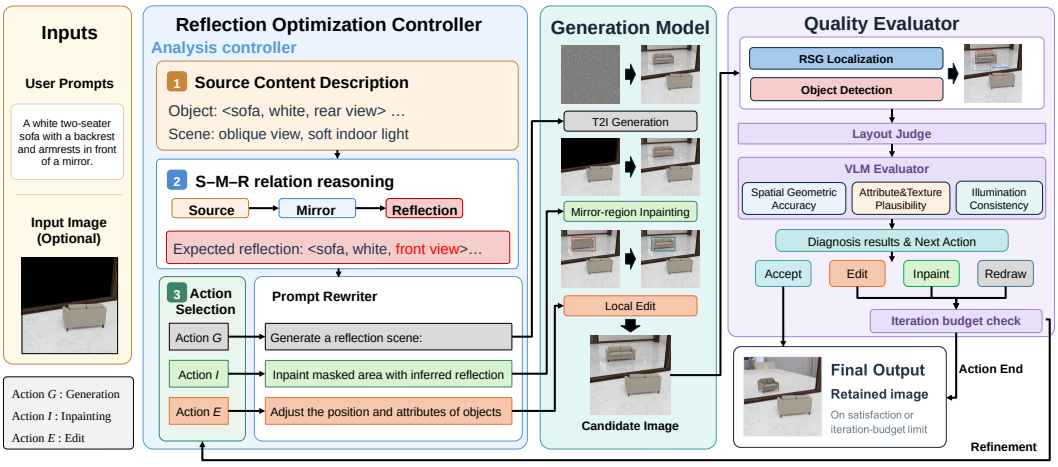

Figure 2: MirrorRefine overview. Source-scene analysis and S–M–R reasoning provide reflection content; RSG and VLM assessment guide action selection and correction. Only one synthesis action executes per round. Verified candidates update the retained image; Accept and budget exhaustion return it. The optional input is not ground truth, and examples illustrate the workflow.

3 METHOD

3.1 OVERVIEW

MirrorRefine coordinates a Reflection Optimization Controller R_ϕ, a Generation Model G_θ, and a Quality Evaluator J_ψ (Figure 2). Given a prompt p_0 and an optional image x^{in}, the controller infers reflected content from the requirements and visible source scene, then constructs synthesis conditions. The evaluator combines coarse spatial evidence with VLM diagnosis to select the next action and request corrections. Each candidate is verified before replacing the retained image. The three actions are full-image generation (Action G), mirror-region inpainting (Action I), and targeted editing (Action E). All component parameters, including the offline-trained RSG, remain fixed during refinement.

3.2 SCENE-BASED REFLECTION INFERENCE

The controller's Analysis Controller A_ϕ first describes source objects, their attributes, scene relationships, viewpoint, and illumination. For image inputs, it uses visible content outside the mirror rather than treating the existing reflection as a reliable target. Reflection Content Inference then establishes Source–Mirror–Reflection (S–M–R) relations to infer reflected identities, visible surfaces, orientations, and spatial relationships. A compatible reflection may show a different surface of the same object: the sofa in Figure 2 presents its rear directly but may reveal its seating side in the mirror. Unobservable details remain uncertain, and original requirements are distinguished from observations and model assumptions.

The Prompt Rewriter W_ϕ combines this reflection context with the selected action and revision suggestions. It constructs a scene prompt p_{t+1} and, for Action I or Action E, an instruction e_{t+1} specifying the correction and content to preserve. Original requirements remain fixed across rounds. For fixed-scene completion, the supplied image remains the reference for preserving content outside the mirror. Initialization and action-specific templates appear in Appendix A.1.

3.3 SPATIALLY GUIDED DIAGNOSIS

Without a spatial reference, VLMs can struggle to judge whether reflected objects occupy physically plausible positions. RSG predicts coarse reflection bounding-box regions from an RGB image with its mirror interior erased and a binary mirror mask. These regions supply expected locations for assessing the observed reflection without requiring exact contours or a complete three-dimensional

reconstruction. We train SegFormer-B0 (Xie et al., 2021) offline on Mirror-Reflect-Inference, comprising 300,000 Blender-rendered images; training details appear in Appendix A.2.

Object Detection locates visible content in the candidate. Layout Judge and the VLM associate detections with RSG regions using object identity and scene context; the coarse regions themselves do not identify objects. The Quality Evaluator checks *geometry* through positions, scale, orientation, and occlusion; *appearance* through identity, attributes, and texture across direct and reflected views; and *illumination* through observable lighting and shadow relationships. Ambiguous associations and insufficient evidence remain unknown. These judgments guide refinement without certifying physical correctness.

The evaluator selects an allowed synthesis action and supplies repair targets, requested changes, and preservation requirements. Feedback links each issue to an object or relation, an affected region, and supporting observations, so the next instruction addresses the diagnosed inconsistency. Action I reconstructs a local or complete mirror region, whereas Action E targets a localized change. Fixed-scene completion permits these two actions; text-only generation also permits Action G. Accept terminates refinement when all necessary checks pass.

3.4 Synthesis and Candidate Retention

Text-only input initializes through Action G; image-and-text input initializes through Action I using x^{in}. After evaluating this initial candidate, the system maintains a retained image–prompt pair $(x_t^\star, p_t^\star)$ and its diagnosis. Subsequent candidates use the selected action a_t:

$$\widetilde{x}_{t+1} = \begin{cases} G_\theta(p_{t+1}; z_{t+1}), & a_t = \mathsf{G}, \\ G_\theta(x_t^\star, p_{t+1}, e_{t+1}, M_t^{\text{inpaint}}; z_{t+1}), & a_t = \mathsf{I}, \\ G_\theta(x_t^\star, p_{t+1}, e_{t+1}, M_t^{\text{edit}}; z_{t+1}), & a_t = \mathsf{E}, \end{cases} \tag{1}$$

where z_{t+1} denotes generation randomness and M_t^{inpaint} and M_t^{edit} specify the intended repair regions. Mask-capable tools receive binary support; instruction-based tools express regional conditions through instructions and visual references. Spatial-support construction and backend mappings are detailed in Appendices A.2 and A.4.

A candidate replaces the retained pair only when verification supports the targeted repairs and preservation requirements without detecting a new defect. Otherwise, the preceding image remains the editing base and feedback revises the next instruction. Retention may resolve only part of the diagnosis; terminal Accept requires all necessary checks to pass. The rule depends on evaluator judgments and does not guarantee monotonic physical improvement. Appendix A.3 specifies the full update rule.

Both configurations use $B = 8$ single-image synthesis attempts per sample and run. Initialization, failures, and retries consume this budget when dispatched; non-synthesis calls do not. If the budget expires before Accept, the system returns the retained image with unresolved issues. Algorithm 1 gives the complete procedure.

4 Experiments

4.1 Experimental Setup

We evaluate fixed-scene completion on 500 MirrorBenchV2 samples (250 single-object and 250 multi-object) (Dhiman et al., 2025a), 200 MSD samples (Yang et al., 2019), and 200 CC-Mirror image–caption pairs curated from MS COCO (Lin et al., 2014). PSNR, SSIM, and LPIPS assess reference-image quality; reflection-mask IoU and BF1 assess geometry where object annotations are available. GLAM-Bench evaluates geometry, lighting, and appearance using offline GPT-5.6 Sol scoring, separately from online refinement feedback. Data selection, metric protocols, and baseline configurations are detailed in Appendices B.1 and B.2.

The local system combines Qwen3.8-27B with FLUX generation, inpainting, and editing tools; the online system uses GPT-5.6 Sol with GPT Image 2. Both use $B = 8$, counting initialization, failures, and retries. RSG remains frozen. Exact tool mappings and training settings appear in Appendix A.2 and Section A.4.

Table 1: MirrorBenchV2 completion. B is the synthesis-attempt cap, including initialization and retries. PSNR is in dB; GLAM-Bench reports geometry, lighting, and appearance on a 0–100 scale. **Bold**/underlined scores are best/second-best within each group. Protocols and additional geometry metrics appear in Appendices B.2 and B.3.

		Visual quality			Alignment	Geometry		GLAM-Bench		
Method	B	PSNR↑	SSIM↑	LPIPS↓	CLIPscore↑	IoU↑	BF1↑	Geo.↑	Light↑	App.↑
I. Base generators										
FLUX.1-dev	1	12.4881	0.7653	0.2357	26.3824	0.4003	**0.1458**	54.24	70.27	**68.77**
Show-o	1	13.3620	0.7513	0.2345	27.3191	0.3780	0.1288	56.27	**71.56**	68.44
SANA-1.0 1.6B	1	12.8555	0.7389	0.2437	27.1641	0.3958	0.1426	48.72	66.54	64.57
SD3.5 Large	1	13.2918	0.7486	0.2224	27.6105	0.3930	0.1325	52.30	69.92	67.61
Janus-Pro-7B	1	14.2220	**0.7766**	0.2293	25.7203	0.3978	0.1206	43.79	62.09	58.60
SANA-1.5 4.8B	1	13.9403	0.7463	0.2273	27.6779	0.4003	0.1394	52.29	67.64	63.03
Lumina-Image 2.0	1	**15.2823**	0.7720	**0.2202**	26.3583	**0.4088**	0.1408	52.28	68.47	63.03
BAGEL	1	14.3693	0.7724	0.2250	**28.3460**	0.3980	0.1209	45.93	65.29	61.51
HiDream-I1-Full	1	13.9558	0.7599	0.2268	27.3411	0.3912	0.1360	51.75	68.62	64.00
Qwen-Image	1	13.7078	0.7625	0.2281	27.6197	0.3978	0.1308	**57.31**	69.80	66.81
II. Mirror-specific methods										
MirrorFusion	4	17.7785	0.7498	0.1331	26.5270	0.4795	0.2132	65.00	71.20	75.00
MirrorVerse	4	18.8104	0.7632	0.1097	26.4394	0.5863	0.3011	67.30	72.60	75.10
PhysReflect	4	**19.1194**	**0.7731**	**0.1071**	**26.5741**	**0.6932**	**0.4228**	**68.10**	**74.30**	**75.90**
III. Inference-Time Scaling										
Noise Scaling	8	12.2226	0.7363	0.2407	26.4848	0.3981	0.1352	53.93	68.81	65.48
Reflect-DiT	8	12.2824	0.7294	0.2474	26.5260	0.3947	0.1346	45.13	64.57	58.87
Noise & Prompt Scaling	32	12.3200	0.7454	0.2328	26.4564	0.3992	0.1366	54.47	70.24	66.81
TIR	4	13.0027	0.7476	0.2367	26.4295	0.4048	0.1365	51.60	69.29	65.04
T2I-Copilot	4	12.7556	0.7318	0.2282	26.0419	0.3896	0.1317	60.07	71.30	69.68
RAISE	32	12.8981	0.7410	0.2334	26.1252	0.3740	0.1313	57.31	71.94	68.07
MirrorRefine-local (Ours)	8	16.5389	**0.7538**	**0.1641**	26.1756	**0.4273**	**0.1673**	69.54	72.73	75.66
MirrorRefine-Online (Ours)	8	**16.6513**	0.7459	0.1660	**26.7914**	0.4069	0.1477	**74.43**	**74.84**	**78.74**

4.2 Quantitative and qualitative comparisons

MirrorBenchV2. Table 1 shows complementary quality and consistency trends. Relative to FLUX.1-dev, the local configuration improves PSNR by 4.0508 dB and LPIPS by 0.0716, with higher IoU but lower SSIM. Online leads the reported GLAM-Bench geometry, lighting, and appearance scores (74.43/74.84/78.74); Local ranks second within inference-time scaling and has higher IoU/BF1 than Online. Both remain below PhysReflect in PSNR and IoU. These are system-level comparisons: differing backbones, budgets, and selectors prevent attributing gains solely to the controller.

Real-world and open-domain completion. On MSD and CC-Mirror, Local achieves the highest PSNR among the listed methods (18.02 and 20.35 dB), while Online leads all three GLAM-Bench dimensions. Local also leads SSIM and LPIPS on CC-Mirror; on MSD, fine-tuned PhysReflect retains the best SSIM and LPIPS. Thus, neither deployment uniformly leads reference-image and evaluator-based metrics. Full results and dataset-specific comparisons appear in Tables 6 and 7.

User study. Five volunteers compared six methods on 200 samples, selecting the best output for image quality, image–text consistency, and image plausibility (Appendix B.4). Local leads all three criteria in Table 2, with an overall selection rate of 42.77% versus PhysReflect's 31.37%. Its largest margin is in image–text consistency (12.70 percentage points). These pooled preferences complement the objective scores without establishing physical correctness.

Qualitative comparison. Figures 3 and 4 illustrate more coherent reflected appearance, scale, and scene content. MirrorBenchV2 examples are comparable to PhysReflect, with clearer advantages in the illustrated MSD and CC-Mirror scenes. Additional examples appear in Appendix B.6.

Table 2: Best-image selection rates (%) in the six-method user study. Overall is the three-dimension mean. **Bold**/underlined denote best/second-best.

Method	Image quality↑	Image–text consistency↑	Image plausibility↑	Overall↑
MirrorRefine-local (Ours)	**41.90**	**42.10**	**44.30**	**42.77**
PhysReflect	30.40	29.40	34.30	31.37
T2I-Copilot	8.90	9.70	7.80	8.80
RAISE	8.60	8.10	7.10	7.93
Noise Scaling	8.20	8.30	5.30	7.27
TIR	2.00	2.40	1.20	1.87

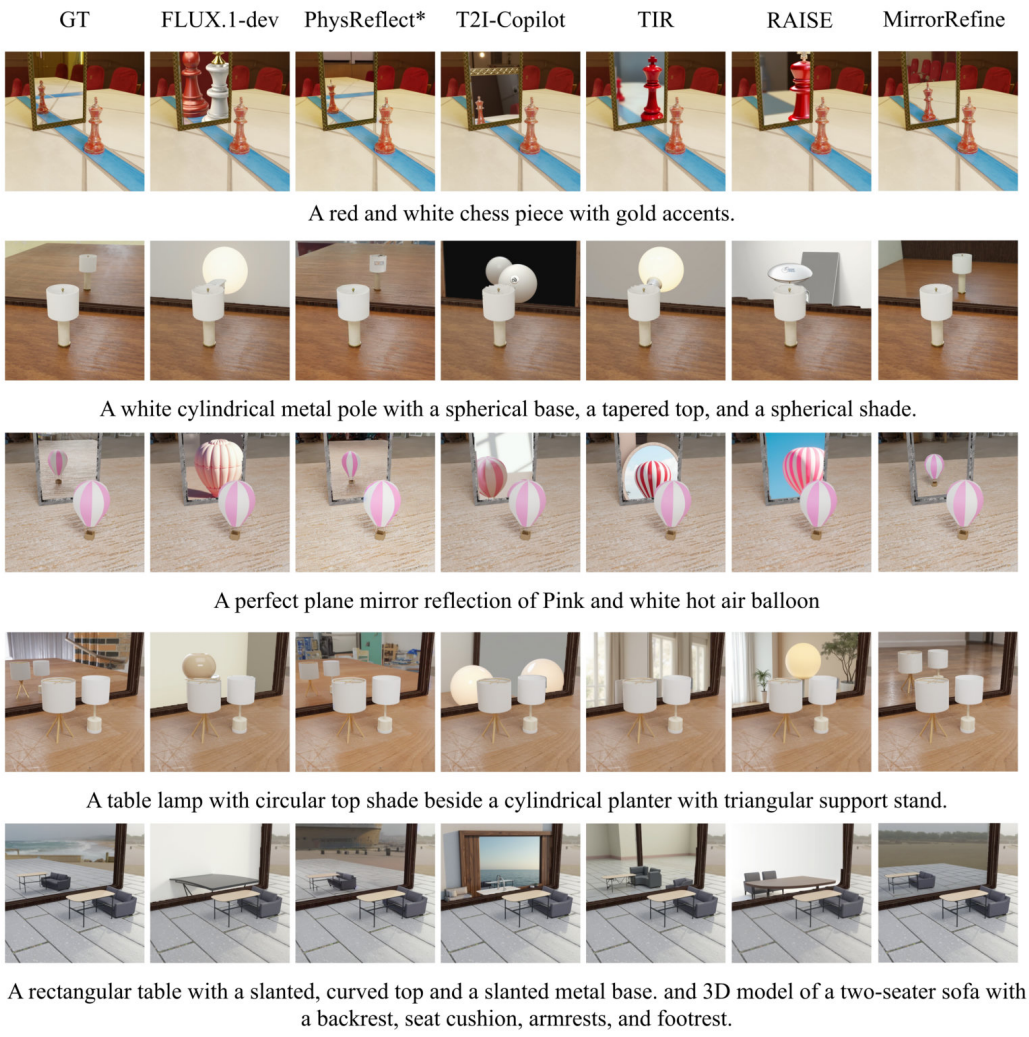

Figure 3: Qualitative comparison on MirrorBench V2. GT denotes the ground-truth reference. An asterisk (*) denotes methods trained or fine-tuned on this dataset.

4.3 Component and model-configuration ablations

Reflection context and spatial guidance. Table 3 ablates the local pipeline (Appendix B.5). Removing context affects appearance and lighting more than removing RSG, whereas RSG removal causes larger geometry, IoU, and BF1 drops. Without spatial guidance, VLMs struggle to judge re-

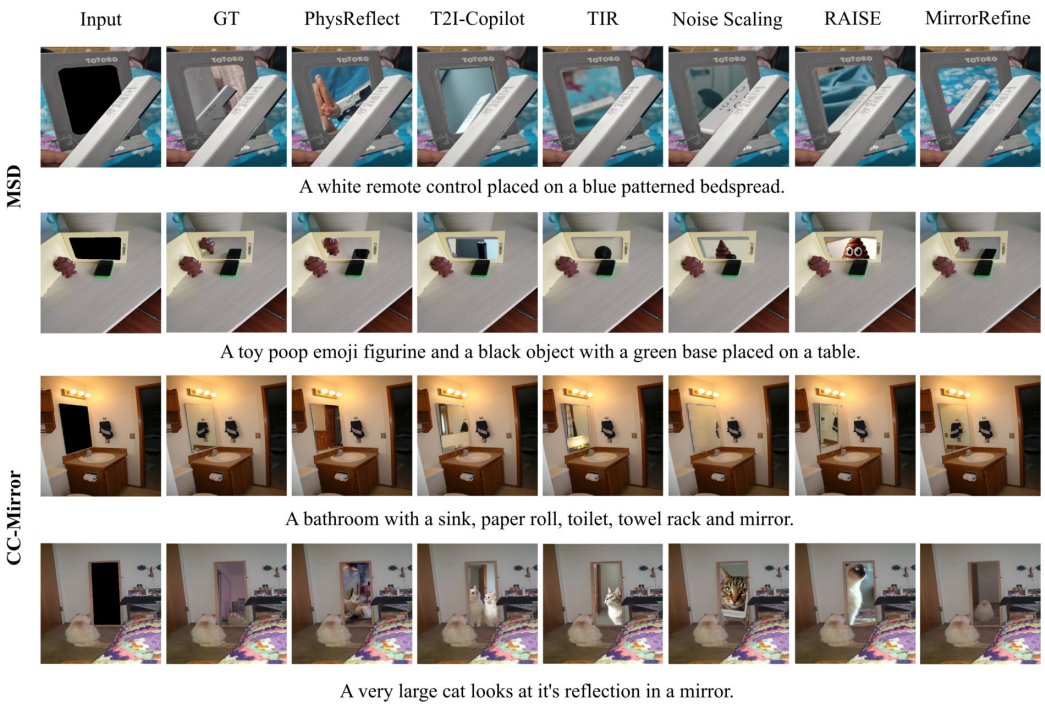

Figure 4: Qualitative mirror-completion comparisons on MSD (top two rows) and CC-Mirror (bottom two rows). Input shows the source image with the mirror region masked; GT denotes the ground-truth reference.

Table 3: Local component ablations on MirrorBenchV2 at $B = 8$. IoU is reflection-mask IoU; Rounds is mean synthesis attempts including initialization, failures, and retries. Quality scores are ranked within the table (**best**/second-best). Settings: Appendix B.5.

	Visual quality				Geometry		GLAM-Bench			
Variant	PSNR↑	SSIM↑	LPIPS↓	CLIPscore↑	IoU↑	BF1↑	Geo.↑	Light↑	App.↑	Rounds
Full pipeline	**16.5389**	**0.7538**	0.1641	26.1756	**0.4273**	**0.1673**	**69.54**	**72.73**	**75.66**	4.1
Without reflection context	14.6038	0.7505	0.1868	26.1022	0.4153	0.1452	68.01	70.39	71.94	4.3
Without RSG	14.4587	0.7526	0.1884	26.2397	0.3967	0.1388	63.37	71.51	73.26	4.7
Without both	13.6707	0.7416	0.1944	**26.3824**	0.3895	0.1365	58.99	69.08	68.52	5.4
Fill for all actions	16.2365	0.7508	**0.1632**	26.1810	0.4269	0.1621	69.08	71.83	72.30	4.2

flected positions from appearance alone. Restoring even coarse RSG layouts raises geometry from 63.37 to 69.54 and IoU from 0.3967 to 0.4273, while reducing mean rounds from 4.7 to 4.1. These final-image gains support spatially guided assessment and correction. Removing both cues gives the weakest reference-image quality and requires 5.4 rounds.

Synthesis tools and online evaluation. Using Fill for every action slightly improves LPIPS (0.1632 versus 0.1641), but lowers PSNR and all GLAM-Bench scores, with 4.2 versus 4.1 rounds. Table 4 fixes the 27B controller and varies the synthesis backend and online evaluator (Appendix B.5). Replacing the 27B evaluator with GPT-5.6 Sol improves geometry by 7.31 points with hosted synthesis, but reduces it by 0.70 with FLUX. Both GPT-evaluator configurations use 3.8 rounds. Evaluator benefits therefore depend on the backend, and fewer rounds do not establish lower latency. These hybrid configurations differ from the fully online system, which also uses GPT-5.6 Sol for control.

Table 4: Backend / online evaluator with a fixed Qwen3.8-27B controller at $B = 8$. 27B: Qwen3.8-27B; GPT: GPT-5.6 Sol; Online: GPT Image 2. FLUX/27B is the full local baseline. Quality rankings are within this table; metrics follow Table 3.

	Visual quality				Geometry		GLAM-Bench			
Backend / evaluator	PSNR↑	SSIM↑	LPIPS↓	CLIPscore↑	IoU↑	BF1↑	Geo.↑	Light↑	App.↑	Rounds
FLUX / 27B	16.5389	**0.7538**	**0.1641**	26.1756	**0.4273**	**0.1673**	**69.54**	72.73	75.66	4.1
FLUX / GPT	**16.6231**	0.7471	0.1659	26.4805	0.4272	0.1658	68.84	72.33	74.97	3.8
Online / 27B	16.5383	0.7443	0.1651	26.3722	0.4223	0.1436	62.11	73.68	75.67	4.1
Online / GPT	16.5359	0.7481	0.1664	**26.5380**	0.4239	0.1492	69.42	**74.13**	**76.23**	3.8

4.4 Synthesis workflow example

Figure 5 shows a refinement sequence from a supplied image with the mirror region masked. Round 1 uses mirror-region inpainting (Action I) to establish the reflected cabinet and surrounding scene. Rounds 2–4 use targeted editing (Action E) to adjust the reflected cabinet's placement and appearance, while the visible source cabinet and room remain recognizable. The sequence illustrates the transition from initial completion to localized revision. The final Output panel displays the returned image, rather than an additional synthesis round.

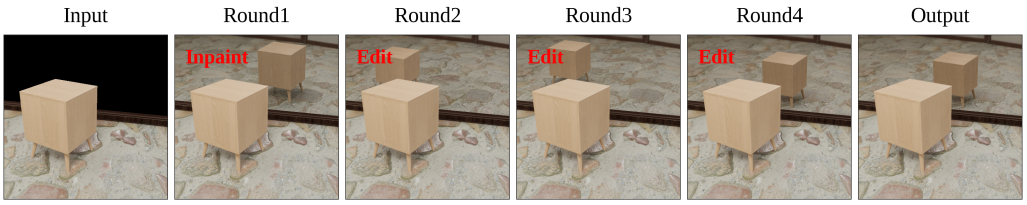

Figure 5: Example synthesis workflow. From left to right: masked input, Round 1 with Action I, Rounds 2–4 with Action E, and the returned output. The intermediate panels show successive revisions of the reflection.

5 Limitations and Discussion

Single-image ambiguity and VLM errors limit geometric verification; coarse layouts and local edits can also fail (Appendix B.7). Hosted backends limit reproducibility. Evaluation covers static fixed-scene completion, not temporal world modeling, the text-only interface, or direct diagnostic accuracy. Human preferences come from five volunteers, and synthesis counts do not measure wall-clock speed. GLAM-Bench and preference scores do not certify physical correctness.

6 Conclusion

MirrorRefine combines source-driven reflection inference, coarse spatial guidance, and verified multi-action refinement. Completion benchmarks, human preferences, and ablations support this combination while revealing trade-offs between reference-image quality and evaluator-rated consistency. The framework retains unresolved evidence rather than treating plausible appearance as verified geometry.

AI use statement

In this work, we used generative AI tools to polish the language of the manuscript and improve its readability. We did not use generative AI tools to generate research ideas, methods, experimental results, analyses or figures. We reviewed all AI-assisted edits to ensure technical accuracy and preserve the intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ofek Basson, Shimon Vainer, Yacov Hel-Or, and Ohad Fried. Fill my mirror: Geometry-constrained mirror inpainting. *arXiv preprint arXiv:2609.03740*, 2026. URL https://arxiv.org/abs/2609.03740v1.

Black Forest Labs. FLUX.1-dev: Model card. Hugging Face model card, 2024a. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Accessed 2026-09-24.

Black Forest Labs. FLUX.1-Fill-dev: Model card. Hugging Face model card, 2024b. URL https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev. Accessed 2026-09-24.

Black Forest Labs. FLUX.1-Kontext-dev: Model card. Hugging Face model card, 2025. URL https://huggingface.co/black-forest-labs/FLUX.1-Kontext-dev. Accessed 2026-09-24.

ByteDance Seed. BAGEL-7B-MoT: Model card. Hugging Face model card, 2025. URL https://huggingface.co/ByteDance-Seed/BAGEL-7B-MoT. Accessed 2026-09-24.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Caesar_COCO-Stuff_Thing_and_CVPR_2018_paper.html.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zijian Gong, Jianzhuang Pan, Yi Peng, Rui Tian, Siyu Wang, Bo Zhao, Ting Yao, and Tao Mei. HiDream-I1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. URL https://arxiv.org/abs/2505.22705.

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2I-Copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19396–19405, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chen_T2I-Copilot_A_Training-Free_Multi-Agent_Text-to-Image_System_for_Enhanced_Prompt_Interpretation_ICCV_2025_paper.html.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Cheng_Boundary_IoU_Improving_Object-Centric_Image_Segmentation_Evaluation_CVPR_2021_paper.html.

DeepSeek. Janus-Pro-7B: Model card. Hugging Face model card, 2025. URL https://huggingface.co/deepseek-ai/Janus-Pro-7B. Accessed 2026-09-24.

Ankit Dhiman, Manan Shah, and R Venkatesh Babu. MirrorVerse: Pushing diffusion models to realistically reflect the world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11239–11249, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dhiman_MirrorVerse_Pushing_Diffusion_Models_to_Realistically_Reflect_the_World_CVPR_2025_paper.html.

Ankit Dhiman, Manan Shah, Rishubh Parihar, Yash Bhalgat, Lokesh R Boregowda, and R Venkatesh Babu. Reflecting reality: Enabling diffusion models to produce faithful mirror reflections. In *International Conference on 3D Vision*, 2025b. doi: 10.1109/3DV66043.2025.00081. URL https://arxiv.org/abs/2409.14677v2.

S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King,

Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. doi: 10.1126/science.aar6170. URL https://doi.org/10.1126/science.aar6170.

Shuheng Ge, Hongwei Ren, Li Zhang, and Xiangqian Wu. PhysReflect: Geometry and perception guided diffusion for physically-plausible mirror reflections. *arXiv preprint arXiv:2609.23442*, 2026. URL https://arxiv.org/abs/2609.23442v1.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. doi: 10.48550/arXiv.1803.10122. URL https://arxiv.org/abs/1803.10122.

Liyao Jiang, Ruichen Chen, Chao Gao, and Di Niu. RAISE: Requirement-adaptive evolutionary refinement for training-free text-to-image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22038–22048, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jiang_RAISE_Requirement-Adaptive_Evolutionary_Refinement_for_Training-Free_Text-to-Image_Alignment_CVPR_2026_paper.html.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. BrushNet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.06976.

Mohammed Abdul Hafeez Khan, Yash Jain, Siddhartha Bhattacharyya, and Vibhav Vineet. Test-time prompt refinement for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6565–6575, 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/MARS2/html/Khan_Test-time_Prompt_Refinement_for_Text-to-Image_Models_ICCVW_2025_paper.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

V. Kovalev, A. Kuvshinov, A. Buzovkin, D. Pokidov, and D. Timonin. CRAFT: Continuous reasoning and agentic feedback tuning for multimodal text-to-image generation. *arXiv preprint arXiv:2512.20362*, 2025. URL https://arxiv.org/abs/2512.20362v2. Version 2, revised January 2026.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-DiT: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15657–15668, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_Reflect-DiT_Inference-Time_Scaling_for_Text-to-Image_Diffusion_Transformers_via_In-Context_Reflection_ICCV_2025_paper.html.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=hFALpTb4fR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao

Lu. NVILA: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4122–4134, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_NVILA_Efficient_Frontier_Visual_Language_Models_CVPR_2025_paper.html.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. URL https://arxiv.org/abs/2501.09732.

Xuan-Bach Mai, Duy-Phuc Nguyen, Quoc-Van Le, Tam V. Nguyen, Thanh-Toan Do, Huu Le, Duong-Van Nguyen, Minh-Triet Tran, and Trung-Nghia Le. PhysMirror: Physics-aware mirror object generation. *arXiv preprint arXiv:2607.03470*, 2026. URL https://arxiv.org/abs/2607.03470v1. IROS 2026; cited arXiv version 1.

OpenAI. GPT-5.6 Sol: Model documentation. OpenAI developer documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed 2026-09-24.

OpenAI. GPT Image 2: Model documentation. OpenAI developer documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-image-2. Accessed 2026-09-24.

Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann, 3 edition, 2016. URL https://www.pbr-book.org/3ed-2018/Reflection_Models/Specular_Reflection_and_Transmission.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Conghui He, Bo Zhang, Xiaohong Liu, Hongsheng Li, Yu Qiao, Chang Xu, and Peng Gao. Lumina-Image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025. URL https://arxiv.org/abs/2503.21758.

Qwen. Qwen-Image: Model card. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen-Image. Accessed 2026-09-24.

Qwen Team. Qwen3.8-27B: Model card. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed 2026-09-24.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Show Lab. Show-o: Model card. Hugging Face model card, 2024. URL https://huggingface.co/showlab/show-o. Accessed 2026-09-24.

Stability AI. Stable Diffusion 3.5 Large: Model card. Hugging Face model card, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-large. Accessed 2026-09-24.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. GenArtist: Multimodal LLM as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/e7c786024ca718f2487712bfe9f51030-Abstract-Conference.html.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ieeexplore.ieee.org/document/1284395/.

Shiyu Wu, Mingzhen Sun, Weining Wang, Yequan Wang, and Jing Liu. VisualPrompter: Semantic-aware prompt optimization with visual feedback for text-to-image synthesis. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hIwVFRLaFy.

Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting LLM-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6327–6336, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Self-correcting_LLM-controlled_Diffusion_Models_CVPR_2024_paper.html.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/34345e243156da67605d4b63d71c8d98-Abstract-Conference.html.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68578–68598. PMLR, 2025b. URL https://proceedings.mlr.press/v267/xie25b.html.

Bangji Yang, Ruihan Guo, Jiajun Fan, Chaoran Cheng, and Ge Liu. M3: High-fidelity text-to-image generation via multi-modal, multi-agent and multi-round visual reasoning. *arXiv preprint arXiv:2602.06166*, 2026. URL https://arxiv.org/abs/2602.06166v1.

Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W. H. Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8809–8818, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yang_Where_Is_My_Mirror_ICCV_2019_paper.html.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15329–15339, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhuo_From_Reflection_to_Perfection_Scaling_Inference-Time_Optimization_for_Text-to-Image_Diffusion_ICCV_2025_paper.html.

A REFINEMENT PROTOCOL AND IMPLEMENTATION

A.1 CONTROLLER INPUTS AND ACTION-SPECIFIC PROMPTS

A run fixes the original prompt p_0, the optional input image x^{in}, and the required scene and reflection properties. Source-scene observations and inferred reflection content inform repair instructions without changing these requirements. For fixed-scene completion, x^{in} supplies the exterior preservation reference; its mirror contents are not assumed correct. Ground-truth RGB and evaluation-only geometry do not enter the refinement loop.

Text-only input initializes with Action G, whereas image-and-text input initializes with Action I. The controller constructs the initial scene prompt p_{init}; Action I additionally uses the input image, inferred mirror content, preservation instructions, and valid mirror-region support. Both routes synthesize and evaluate an initial candidate before establishing the retained image–prompt pair $(x_0^\star, p_0^\star)$ and diagnosis $F_0^\star$. Loading an input image alone does not generate a candidate.

Original requirements, image observations, and model assumptions are recorded separately. Before text-only initialization, unspecified scene details remain generation assumptions. During refinement, the Analysis Controller combines the original requirements, source-object descriptions, Source–Mirror–Reflection relations, the retained diagnosis, and the latest candidate feedback. Reflection Content Inference specifies the expected identity, visible surfaces, orientation, and relationships from source-scene and viewpoint evidence. Unobservable details remain uncertain. Action Selection receives the action specified by the input path or the Quality Evaluator, together with revision suggestions, and organizes the next instruction.

Action-conditioned rewriting. The Prompt Rewriter selects the corresponding action template and combines reflection context with the revision suggestions. It constructs p_{t+1} and, for Action I or Action E, an instruction e_{t+1}:

- Action G: describe the requested scene, mirror, reflected content, and correction objective for full-image generation.
- Action I: specify the local or full mirror region to reconstruct, the intended reflected content, and the exterior and unaffected mirror content to preserve.
- Action E: identify the target object and localized defect, the directed change, and the content to preserve.

Only the selected action is executed. After initialization, Action I and Action E use the retained image as their editing base. Fixed-scene completion restricts synthesis to these two actions; Action G is available in the text-to-image track. After full-image generation, source regions and viewpoint-dependent hypotheses are reassessed for the new scene.

A.2 RSG TRAINING AND SPATIAL SUPPORT

Architecture and inputs. RSG uses SegFormer-B0 (Xie et al., 2021) with an ImageNet-pretrained MiT-B0 encoder and an MLP prediction head. Its four-channel input concatenates RGB with the mirror interior erased and a binary mirror mask. Mirror pixels are replaced by the ImageNet mean before resizing. Inference uses 512×512 inputs and maps the predicted soft layout back to the image coordinates. A fixed threshold of 0.5 converts the soft layout into a binary mask, from which connected-component bounding boxes are extracted. RSG receives neither reflected RGB nor depth, camera calibration, or source-object masks at inference. Its parameters remain frozen during refinement.

Mirror-Reflect-Inference and training. We construct Mirror-Reflect-Inference in Blender using randomly sampled object assets and mirror arrangements. A fixed set of 16 camera viewpoints uniformly distributed over a sphere produces 300,000 synthetic samples. Geometric supervision from rendering provides coarse reflection layouts through reflection and projection of visible source surfaces. Depth and camera annotations are used only for offline supervision. SegFormer-B0 is trained on a single NVIDIA H100 with 512×512 inputs, batch size 64, and 10 epochs.

Figure 6 reports mean union-box IoU on 256 test samples across the ten epochs. It rises from 63.88% at epoch 1 to 71.37% at epoch 10. This measures coarse layout agreement rather than exact reflection segmentation.

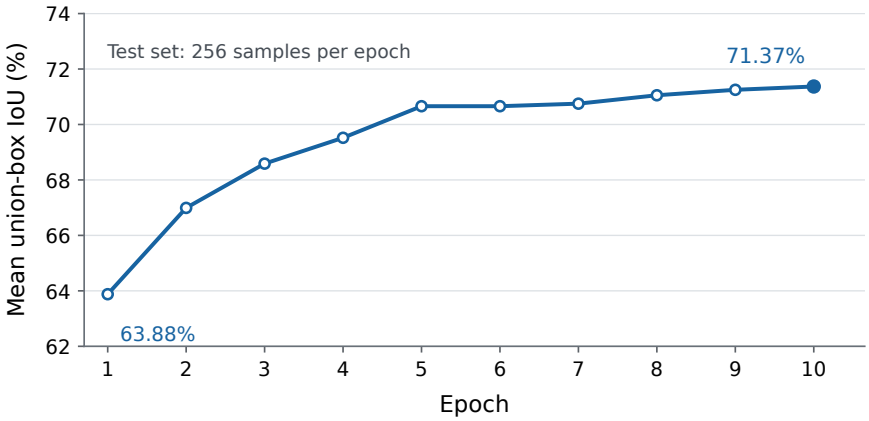

Figure 6: RSG localization on 256 test samples during ten training epochs. The vertical axis reports mean union-box IoU (%); the final value is 71.37%.

Object association and editable regions. Object Detection identifies content visible in the candidate, while RSG supplies possible reflection locations. Layout Judge and the VLM associate these regions using source-object identity, appearance, scene relationships, and mirror context. RSG connected regions do not provide object identities: multiple objects can merge into one region. Ambiguous associations remain uncertain, and an empty prediction alone does not establish a missing reflection.

Action I supports local or full-mirror reconstruction. Local inpainting handles larger corrections that are difficult to perform with Action E. Let M_t^{local} be the diagnosis-supported local repair region and M_t^{mirror} the valid mirror region. The controller constructs

$$M_t^{\text{inpaint}} = \begin{cases} M_t^{\text{local}} \cap M_t^{\text{mirror}}, & \text{local mirror inpainting,} \\ M_t^{\text{mirror}}, & \text{full-mirror inpainting.} \end{cases} \tag{2}$$

For a confirmed displacement, $M_t^{\text{local}} = M_t^{\text{err}} \cup \widehat{M}_t^{\text{target}}$ combines the erroneous region and the coarse destination associated by the diagnosis in Section 3.3. The instruction requests removal at the incorrect location and reconstruction at the destination to avoid leaving a duplicate. Other local corrections use their affected regions; missing content or widespread inconsistencies, such as mirror-wide illumination errors, can require full-mirror inpainting. Action E uses the localized support $M_t^{\text{edit}} \subseteq M_t^{\text{mirror}}$. All requested editing supports are restricted to the permitted mirror region.

A.3 Diagnosis and Candidate Retention

The evaluator combines the full image, associated object regions, and RSG evidence to assess spatial geometry, attributes and texture, and illumination. Each diagnosed issue identifies the object or relation, the affected region, supporting observations, and a suggested correction. Original requirements and previously supported content are checked again after synthesis. Different visible surfaces are allowed when consistent with the viewpoint; insufficient evidence remains unknown. These VLM judgments guide refinement and do not certify exact physical correctness.

Retention rule. Before synthesis, the controller fixes the repair targets and preservation requirements. A candidate replaces the retained image only when complete feedback confirms that the selected problems are repaired, previously supported requirements remain supported, and no new defect is detected. Any previously known failure becoming unknown causes rejection, including

Algorithm 1 Reflection refinement with retained candidates

Input: original prompt, optional input image, fixed requirements, and synthesis-attempt budget $B \leq 8$.
Output: retained image–prompt pair (if available), outcome, and unresolved issues.

1. Initialize the attempt counter to zero. Use Action G for text-only input or Action I for image and text, checking the remaining budget before each attempt and incrementing the counter when it is dispatched. Evaluate the candidate and establish the retained pair and diagnosis. If initialization terminates without an evaluable candidate, return failure without a retained pair.
2. If all necessary checks pass, return the retained pair through Accept. If the attempt counter reaches B, return it with unresolved issues.
3. The Quality Evaluator selects a permitted action and repair targets from the retained diagnosis. The controller constructs the action-specific prompt and editing support. Check the remaining budget, dispatch one synthesis attempt, and increment the counter immediately. Obtain feedback. Every failed attempt and retry consumes this same budget.
4. Apply the retention rule and update the retained pair and diagnosis when it passes. Otherwise preserve the retained pair and use the feedback to revise the next instruction. Return to Step 2.

failures outside the selected targets. Retention need not resolve every remaining issue: a successful position correction can be retained while a separately diagnosed illumination problem remains.

Let Retain_t denote this decision. The image–prompt pair is updated as

$$(x^\star_{t+1}, p^\star_{t+1}) = \begin{cases} (\widetilde{x}_{t+1}, p_{t+1}), & \mathrm{Retain}_t = 1, \\ (x^\star_t, p^\star_t), & \mathrm{Retain}_t = 0. \end{cases} \tag{3}$$

The retained diagnosis is updated with an admitted candidate. Rejected feedback informs the next instruction while the retained image remains available. Retention is distinct from terminal acceptance: Accept requires all necessary checks on the retained image to pass. The rule depends on evaluator judgments and does not guarantee monotonically improving physical image quality.

A.4 Backends, budgets, and termination

The local configuration uses Qwen3.8-27B for visual understanding and reasoning. FLUX.1-dev, FLUX.1-Fill-dev, and FLUX.1-Kontext-dev implement Action G, Action I, and Action E, respectively. The online configuration uses GPT Image 2 for synthesis and GPT-5.6 Sol for reasoning. All component parameters remain fixed during refinement.

Mask-capable interfaces receive the source image, prompt, and binary editing support. Instruction-based interfaces receive the source image and instructions describing the target region, with available visual references. A spatial reference is not equivalent to a submitted binary mask. Preservation is checked after synthesis because regional control depends on the backend. The generation conditions follow Eq. (1) in the main text.

Budget convention. B limits single-image synthesis attempts per input and per run. Both local and online configurations use $B = 8$, with one candidate per synthesis round, so initialization and all subsequent generation, inpainting, editing, and retry attempts total at most eight rounds. Each dispatched attempt consumes one unit regardless of failure, candidate rejection, or terminal acceptance. A request for n images consumes n units. Prompt rewriting, evaluation, and auxiliary calls that do not request image synthesis do not consume this budget.

The controller checks the remaining synthesis-attempt budget before every request or retry. Accept returns the retained image when all necessary checks pass. Otherwise, budget exhaustion returns the retained image with unresolved issues and outcome `BUDGET_EXHAUSTED`. Incomplete feedback cannot replace the retained image. An unrecoverable execution failure returns a failure outcome and the retained result, if available.

B ADDITIONAL EXPERIMENTS

B.1 EVALUATION SETUP

Tasks and input permissions. MirrorBenchV2 contains 500 completion samples: 250 single-object and 250 multi-object test images selected after removing visually similar and low-quality images. MSD contains 200 real-image samples selected by the same procedure. Completion methods receive the source scene and mirror mask; reference reflections and reflected-object annotations are withheld from generation and online diagnosis. Initial Action I uses the supplied image. The tasks concern static opaque objects, one planar mirror, and a single reflection bounce.

CC-Mirror curation and annotation. CC-Mirror contains 200 image–caption pairs curated from MS COCO (Lin et al., 2014). Caption matching and semantic exclusions yield 2,338 candidates, followed by manual image selection. Candidate masks draw on SAM/SAM 2 (Kirillov et al., 2023; Ravi et al., 2025) outputs and COCO-Stuff annotations (Caesar et al., 2018). All 200 final mirror-region masks were manually verified: 100 were redrawn and 100 approved from existing masks. These masks specify completion regions. IoU and BF1 additionally require aligned reflected-object annotations, which are not available for the comparisons in Tables 6 and 7.

Metric applicability. Reference-image and reference-mask metrics apply to fixed-scene completion. SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) measure structural and perceptual similarity, respectively; CLIP-based image–text alignment (Radford et al., 2021) is reported as CLIPscore. Boundary IoU (Cheng et al., 2021) complements mask IoU and boundary F1 in the extended geometry evaluation. Geometry metrics require aligned reflected-object annotations; mirror-region masks or RGB references alone do not provide these targets. For the current-batch MSD and CC-Mirror outputs in Tables 6 and 7, PSNR, SSIM, and LPIPS follow the PhysReflect reference functions on the full 512×512 canvas after setting pixels outside the mirror to zero in both reference and generated images. They are not normalized over mirror pixels alone. PSNR and SSIM retain the reference functions' dynamic data range, and LPIPS uses SqueezeNet with inputs scaled to $[-1, 1]$. The starred MSD variants reuse archived final outputs selected from four candidates using exterior-region SSIM. Their AI evaluation retains each image's native resolution, and their reference-image metrics follow the archived per-image evaluations rather than the current-batch protocol.

GLAM-Bench evaluation. GLAM-Bench (GLAM: A Benchmark for Geometry, Lighting, and Appearance in Mirror Reflection Generation) evaluates final-image reflection consistency. For MirrorBenchV2, it uses GPT-5.6 Sol (OpenAI, 2026a) in separate Geometry, Lighting, and Appearance contexts. Each receives the final image and original request, together with the source scene and mirror region. Method names, online judgments, RSG overlays, and selection scores are hidden. Scores use a 0–100 scale. Offline evaluation and online refinement use separate scoring contexts, although both can use GPT-5.6 Sol; this separation does not eliminate shared model-specific biases.

The real-image comparisons in Tables 6 and 7 use GPT-5.6 Sol for AI geometry, lighting, and appearance scoring. The evaluator receives the final image, mirror-erased input, mirror region, and original prompt, without the reference reflection. Each metric is averaged over its valid values associated with the final outputs; indeterminate AI judgments are excluded rather than assigned zero.

Table 1 uses one final output per MirrorBenchV2 scene. The real-image comparisons in Tables 6 and 7 aggregate the final outputs for their respective datasets. The user-study sample composition and selection protocol are described in Appendix B.4.

B.2 IMPLEMENTATION AND BASELINE CONFIGURATIONS

The default local system uses Qwen3.8-27B for visual understanding and reasoning, with FLUX.1-dev, FLUX.1-Fill-dev, and FLUX.1-Kontext-dev for Action G, Action I, and Action E, respectively (Qwen Team, 2026; Black Forest Labs, 2024a;b; 2025). The online system combines GPT-5.6 Sol reasoning with GPT Image 2 synthesis (OpenAI, 2026a;b). Parameters remain frozen during re-

finement. Object detection and SAM 2 (Ravi et al., 2025) provide observations; RSG predicts coarse layouts from mirror-erased RGB. RSG training is described in Appendix A.2.

System and controller comparisons. The base generators are FLUX.1-dev (Black Forest Labs, 2024a), SD3.5 Large (Stability AI, 2024), SANA-1.0 and SANA-1.5 (Xie et al., 2025a;b), Lumina-Image 2.0 (Qin et al., 2025), HiDream-I1-Full (Cai et al., 2025), Show-o (Show Lab, 2024), Janus-Pro-7B (DeepSeek, 2025), BAGEL (ByteDance Seed, 2025), and Qwen-Image (Qwen, 2025). These generators use source-scene and mirror-region adapters and produce one candidate without external search. Janus-Pro-7B uses a completion adapter; Qwen-Image retains its generation checkpoint. FLUX.1-Fill-dev is included as an additional inpainting baseline in Table 5.

The MirrorVerse baseline denotes its MirrorFusion 2.0 model. In the MirrorBenchV2 comparison, MirrorFusion, MirrorVerse, and PhysReflect retain mirror-specific conditioning and select among four candidates using exterior-region SSIM (Dhiman et al., 2025b;a; Ge et al., 2026). Our Noise Scaling implementation adapts inference-time noise search (Ma et al., 2025), using fixed prompts and an NVILA selector (Liu et al., 2025). Noise & Prompt Scaling additionally rewrites prompts, adapting the noise- and prompt-level mechanisms discussed by Zhuo et al. (2025); it is not the full reflection-tuned ReflectionFlow system. Reflect-DiT retains its SANA generator and feedback context (Li et al., 2025). TIR, RAISE, and T2I-Copilot use completion adapters; T2I-Copilot runs without human intervention (Khan et al., 2025; Jiang et al., 2026; Chen et al., 2025). MirrorBenchV2 outputs are evaluated at 512×512, including TIR outputs generated at mixed native resolutions. Evaluation adds no exterior compositing beyond each method's own output procedure. Backbone, prompt, and selector differences make the main table a system comparison.

Image-generation budgets. Throughout the experiments, B caps single-image synthesis attempts per input and per run, including initialization, failures, and retries. A request for k candidates consumes k units regardless of request batching. In Table 1, base generators use $B = 1$ and mirror-specific methods use $B = 4$. Noise Scaling and Reflect-DiT use $B = 8$; RAISE uses up to four batches of eight ($B = 32$), and Noise & Prompt Scaling uses up to 16 batches of two ($B = 32$). TIR and T2I-Copilot each allow one initial image and three corrections ($B = 4$). Both MirrorRefine configurations use at most eight single-candidate synthesis rounds in total ($B = 8$), including initialization and any retries.

B.3 Additional quantitative comparisons

The local method's boundary IoU improves over FLUX.1-dev and FLUX.1-Fill-dev, while its per-object BF1 remains below Fill (0.1532 versus 0.1616); aggregate overlap gains therefore do not imply uniformly better object boundaries.

On MSD (Table 6), MirrorRefine-local achieves the highest PSNR among the listed methods at 18.02 dB, exceeding the strongest non-MirrorRefine baseline, BAGEL, by 0.83 dB. PhysReflect* attains the best SSIM and LPIPS (0.8659 and 0.1130), compared with 0.8614 and 0.1147 for the local configuration. The online configuration instead leads all three AI dimensions, with geometry, lighting, and appearance scores of 81.90, 75.87, and 88.08. The local configuration ranks second in geometry and appearance, but its lighting score of 74.02 remains below both fine-tuned mirror-specific methods.

On CC-Mirror (Table 7), MirrorRefine-local attains the best PSNR, SSIM, and LPIPS among the listed methods (20.35 dB, 0.8328, and 0.1009), with the online configuration second in all three. The online configuration leads AI geometry, lighting, and appearance at 78.16, 75.26, and 83.07. The local configuration ranks second in geometry and appearance, while its lighting score of 74.08 is lower than T2I-Copilot's 74.93. Across both datasets, the local configuration has higher PSNR and lower LPIPS than the online configuration, whereas the online configuration receives higher AI scores in all three dimensions. SSIM favors the online configuration on MSD and the local configuration on CC-Mirror, so neither configuration uniformly leads across metrics. Dataset and protocol differences preclude interpreting cross-table score gaps as controlled domain-transfer effects.

Table 5: Complementary geometry evaluation of methods in Table 1, with FLUX.1-Fill-dev as an additional inpainting baseline. BIoU denotes boundary IoU; IoU_{obj} and BF1_{obj} are per-object means on the multi-object subset. All scores use a 0–1 scale. **Bold** and underlined values indicate the best and second-best reported scores within each group.

Method	BIoU↑	IoU_{obj} ↑	BF1_{obj} ↑
I. Base generators			
FLUX.1-dev	0.2011	0.3580	0.1376
Show-o	0.1777	0.3588	0.1282
SANA-1.0 1.6B	0.1982	0.3536	0.1356
SD3.5 Large	0.1878	0.3645	0.1332
FLUX.1-Fill-dev	**0.2226**	0.3713	**0.1616**
Janus-Pro-7B	0.1666	0.3700	0.1195
SANA-1.5 4.8B	0.1995	0.3614	0.1293
Lumina-Image 2.0	0.1971	**0.3741**	0.1318
BAGEL	0.1696	0.3620	0.1219
HiDream-I1-Full	0.1859	0.3569	0.1261
Qwen-Image	0.1822	0.3688	0.1278
II. Mirror-specific methods			
MirrorFusion	0.2930	0.4559	0.2114
MirrorVerse	0.5119	0.6286	0.4127
PhysReflect	**0.5287**	**0.6397**	**0.4228**
III. Inference-Time Scaling			
Noise Scaling	0.1968	0.3637	0.1331
Reflect-DiT	0.1831	**0.3827**	**0.1559**
TIR	0.1909	0.3661	0.1311
T2I-Copilot	0.1863	0.3639	0.1288
RAISE	0.1783	0.3336	0.1315
MirrorRefine-local (Ours)	**0.2382**	0.3749	0.1532

B.4 User-study protocol

Five volunteers aged 20–30 evaluated 200 samples: 50 single-object and 50 multi-object samples from MirrorBenchV2, 50 from MSD, and 50 from CC-Mirror. Outputs from six methods—MirrorRefine-local, PhysReflect, T2I-Copilot, RAISE, Noise Scaling, and TIR—were displayed simultaneously at randomized positions. Volunteers selected the best output separately for image quality, image–text consistency, and image plausibility. For each dimension, the selection rate is the proportion of choices selecting a method as best. The overall selection rate is the equally weighted arithmetic mean of the three dimensional rates, not a fourth independent rating. Rounding makes the reported overall rates sum to 100.01%.

Table 2 reports the aggregate selection rates. MirrorRefine-local leads all three dimensions and the overall rate in this six-method comparison. These results describe comparative preferences within the displayed set, not pairwise win rates or absolute reflection-success rates. Image plausibility measures perceived reasonableness and does not establish physical correctness.

B.5 Ablation settings

Reflection context, spatial guidance, and synthesis tools. Table 3 evaluates the default local system on MirrorBenchV2 completion; its full-pipeline row corresponds to MirrorRefine-local in Table 1. Reflection context denotes the reflection-specific conditioning formed from the source-scene description and reflected-content inference. The context-off variant omits this conditioning; RSG-off omits the RSG prior. The all-FLUX.1-Fill variant assigns every synthesis action in this image-conditioned completion task to FLUX.1-Fill. Mean generation rounds count single-image synthesis attempts, including initialization, failures, and retries, under the same $B = 8$ accounting. They measure image usage rather than elapsed time or success.

Table 6: MSD mirror completion. PSNR is in dB; AI geometry (Geo.), lighting (Light), and appearance (App.) scores use a 0–100 scale. * denotes fine-tuning on MSD. **Bold** and underlined values indicate the best and second-best scores among all methods listed in this table.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIPscore↑	Geo.↑	Light↑	App.↑
I. Base generators							
FLUX.1-dev	16.36	0.8629	0.1356	29.63	59.06	68.02	70.83
Show-o	15.65	0.8482	0.1376	28.44	62.77	70.84	80.30
SANA-1.0 1.6B	15.15	0.8377	0.1482	29.41	55.06	66.48	70.37
SD3.5 Large	16.09	0.8510	0.1402	30.09	64.97	66.78	75.97
Janus-Pro-7B	15.92	0.8553	0.1388	27.50	52.23	63.43	65.38
SANA-1.5 4.8B	15.50	0.8355	0.1472	29.76	55.50	66.65	71.11
Lumina-Image 2.0	17.00	0.8611	0.1320	28.69	63.94	69.12	73.83
BAGEL	17.19	0.8590	0.1236	**30.25**	67.26	70.91	77.89
HiDream-I1-Full	15.61	0.8629	0.1395	28.69	55.17	65.25	67.04
Qwen-Image	16.90	0.8630	0.1305	30.15	67.57	69.21	78.62
II. Mirror-specific methods							
MirrorFusion	15.88	0.8365	0.1374	28.13	64.01	67.71	70.62
MirrorVerse	16.48	0.8333	0.1342	28.16	64.53	71.57	71.61
MirrorVerse*	16.54	0.8561	0.1195	28.69	69.50	74.59	75.16
PhysReflect	16.72	0.8416	0.1313	28.01	67.26	72.24	73.76
PhysReflect*	16.85	**0.8659**	**0.1130**	28.52	72.01	75.11	77.59
III. Inference-Time Scaling							
Noise Scaling	16.41	0.8624	0.1346	29.78	65.67	68.27	74.80
Reflect-DiT	15.11	0.8337	0.1487	28.79	41.33	54.95	45.46
Noise & Prompt Scaling	16.08	0.8563	0.1386	29.08	59.84	65.45	69.28
TIR	16.21	0.8608	0.1362	29.54	63.43	70.66	77.02
T2I-Copilot	16.15	0.8589	0.1359	28.82	60.55	70.36	71.59
RAISE	16.59	0.8315	0.1385	28.81	66.25	65.22	71.50
MirrorRefine-online (Ours)	17.59	0.8621	0.1211	29.48	**81.90**	**75.87**	**88.08**
MirrorRefine-local (Ours)	**18.02**	0.8614	0.1147	28.96	77.94	74.02	81.79

Table 7: Mirror completion on CC-Mirror. PSNR is in dB; AI geometry (Geo.), lighting (Light), and appearance (App.) scores use a 0–100 scale. **Bold** and underlined values indicate the best and second-best scores among all methods listed in this table.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIPscore↑	Geo.↑	Light↑	App.↑
II. Mirror-specific methods							
MirrorFusion	17.30	0.7955	0.1472	25.79	62.44	67.45	73.26
MirrorVerse	17.12	0.7924	0.1493	25.42	62.59	69.30	73.95
PhysReflect	17.21	0.7982	0.1477	25.63	64.28	69.93	74.31
III. Inference-Time Scaling							
Noise Scaling	15.93	0.8067	0.1541	26.56	66.86	72.31	75.49
Reflect-DiT	15.18	0.7839	0.1682	**27.38**	55.91	66.59	62.01
Noise & Prompt Scaling	15.88	0.8025	0.1667	26.33	65.39	73.14	74.23
TIR	15.75	0.8051	0.1585	26.54	67.76	71.59	75.24
T2I-Copilot	15.86	0.8039	0.1535	25.88	67.57	74.93	72.51
RAISE	16.37	0.8043	0.1642	26.43	68.75	73.50	76.23
MirrorRefine-online (Ours)	18.66	0.8199	0.1197	26.36	**78.16**	**75.26**	**83.07**
MirrorRefine-local (Ours)	**20.35**	**0.8328**	**0.1009**	25.98	71.91	74.08	81.02

Generation backend and online evaluator. Table 4 fixes the Qwen3.8-27B controller for MirrorBenchV2 completion and crosses two synthesis backends (the local FLUX tools or hosted GPT Image 2) with two online Quality Evaluators (Qwen3.8-27B or GPT-5.6 Sol). The FLUX/Qwen3.8-27B combination is the full-pipeline baseline in Table 3. All four configurations retain reflection context and RSG and follow the same $B = 8$ synthesis accounting. Evaluator replacement changes the Quality Evaluator used during refinement. GLAM-Bench scores describe the final outputs and are reported separately from this online feedback. These configurations retain the 27B controller, whereas the default online system in Table 1 uses GPT-5.6 Sol reasoning. Their results therefore describe different system configurations.

B.6 Additional qualitative examples

Additional mirror-completion comparisons. Figure 7 complements the MSD and CC-Mirror comparisons in Figure 4. The water-bottle and plush-toy scenes illustrate preservation of reflected object appearance, while the puppy and living-room scenes extend the comparison to animals and complex indoor environments. These examples further illustrate the visual consistency of MirrorRefine across diverse scene content.

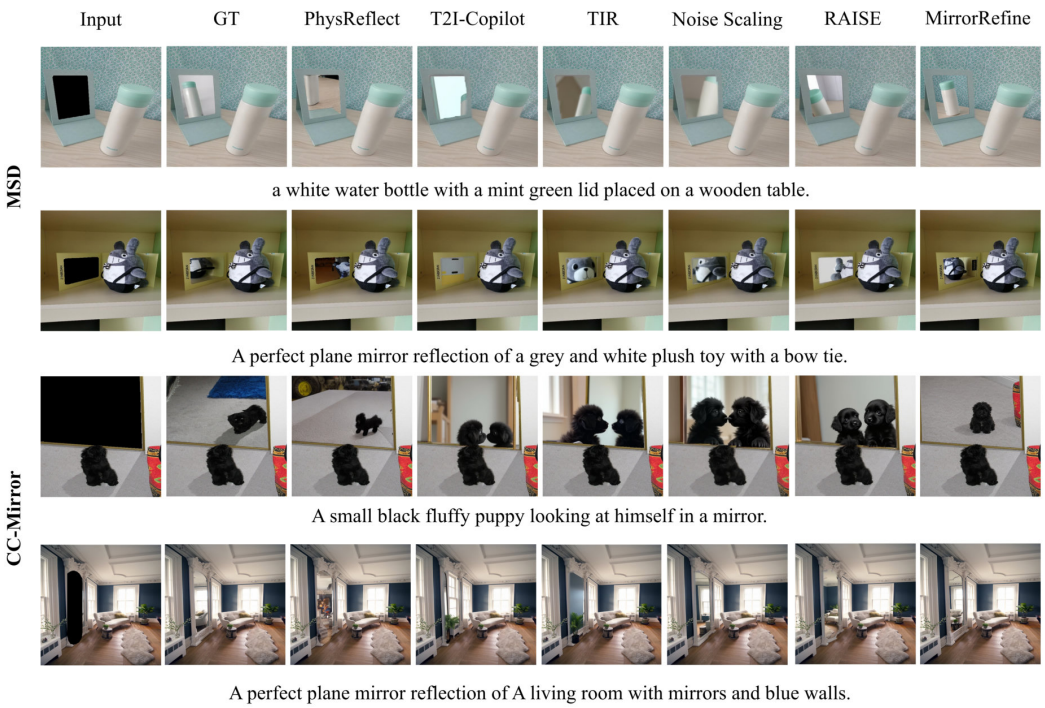

Figure 7: Additional qualitative mirror-completion comparisons on MSD (top two rows) and CC-Mirror (bottom two rows), covering a water bottle, a plush toy, a puppy, and a living room. Input shows the source image with the mirror region masked; GT denotes the ground-truth reference.

Exploratory reflection-content edit. Figure 8 shows an exploratory image edit: the initial sofa reflection repeats its rear surface, whereas the edited candidate exposes the seating side. The edit was produced without RSG, and its geometry remained unresolved during verification. It illustrates a visible content change, rather than an accepted output of the full MirrorRefine pipeline or evidence of correct reflection geometry.

B.7 Failure case analysis

Figure 9 illustrates two representative failure modes concerning reflection content and the assumed mirror geometry. In the bathroom example, the synthesized reflection repeats elements resembling

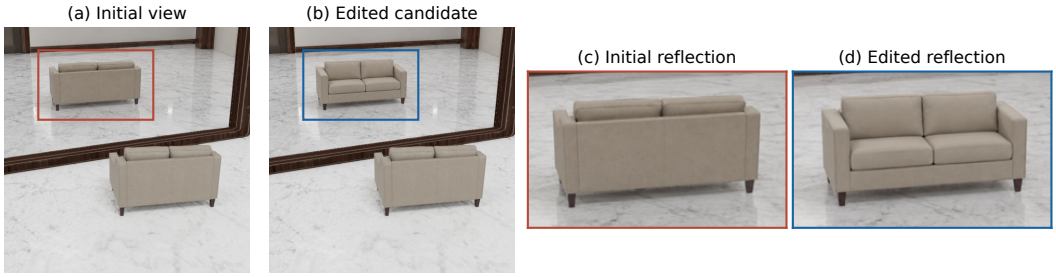

Figure 8: Exploratory reflection-content edit without RSG. Full images and paired mirror crops show a change from a repeated rear surface to a seating-side view. The candidate's reflection geometry remains unverified.

the bathtub and towels visible outside the mirror, whereas the reference reflection contains framed wall art. When the synthesis conditions do not explicitly constrain the reflected content, context inference can overemphasize visible surroundings and transfer them into the mirror region. This restricts the synthesis to an extrapolation of observed content instead of allowing for scene content visible only in the reflected view.

The vehicle example illustrates a limitation with nonstandard reflective configurations. The reference shows the dog's face in the side mirror, but the generated reflection depicts the back of its head and a person holding a camera. Interpreting an atypical reflection using planar-mirror assumptions can lead to an incorrect reflected viewpoint and content. These cases motivate more explicit treatment of uncertainty in reflected content and of configurations beyond the planar-mirror setting considered by the method.



Figure 9: Two representative failure cases. Left: the generated bathroom reflection reproduces visible surrounding content instead of the framed wall art in the reference. Right: the vehicle-mirror output depicts the back of the dog's head rather than its reflected face, illustrating an incorrect reflected viewpoint. GT denotes the ground-truth reference.