TRAJSP: TRAJECTORY-GROUNDED STRUCTURED PROMPTING FOR FEEDBACK-GUIDED TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While structured prompt enhancement enables image generation to benefit from visual knowledge and increased prompter capacity, existing distillation-based approaches require a separate training pipeline for an external LLM enhancer that lacks direct visual access at inference. We study whether unified multimodal models (UMMs) can eliminate the need for this separately trained enhancer by directly observing intermediate images and refining structured conditioning during denoising. Our framework, TRAJSP, first applies a lightweight *noisy understanding adaptation* to enhance the understanding branch's capacity to interpret noisy intermediate images. During sampling, the adapted branch serves as a *trajectory-grounded Polisher*: at an early denoising step, it processes the intermediate visual state jointly with the user prompt to predict a structured prompt comprising global scene fields, per-object elements with normalized bounding boxes, and inter-object relations. By grounding this representation in the evolving visual state, TRAJSP provides targeted conditioning for subsequent denoising steps, with the aim of avoiding redundant prompt expansion. Joint understanding–generation finetuning optimizes shared parameters for both schema prediction from noisy states and schema-conditioned image generation. An optional late-stage Validator is designed to provide targeted feedback for subsequent refinement. Experimental results demonstrate improved semantic alignment and greater fidelity in object-level details, including spatial and action relations, without a standalone prompt enhancer.

1 INTRODUCTION

The quality and faithfulness of text-to-image generation depend strongly on the effectiveness of textual conditioning. Prompt enhancement improves such conditioning by rewriting the user's request before sampling (Hao et al., 2023; Cao et al., 2023), and has become a widely adopted strategy for improving generation quality in modern text-to-image systems. However, a longer description does not necessarily provide more useful conditioning: it may merely restate the original request, introduce redundant descriptions, or hallucinate concepts that fall outside the visual generator's learned distribution, thereby providing little useful guidance or even conflicting with the generation process.

To address the limitations of conventional prompt enhancement, visually distilled structured prompting (VD-SP) distills image-conditioned teacher knowledge into a separate prompt enhancer and organizes descriptions into explicit semantic and geometric fields (Chen et al., 2026b). However, its prompter remains image-free at inference and produces a single-shot description before sampling, preventing it from adapting to the objects, attributes, and relations that actually emerge along a particular generation trajectory. As illustrated in Figure 1, we instead propose TRAJSP, which directly conditions structured prompting on intermediate denoising states, allowing the prompt to adapt to the unfinished scene as generation unfolds.

Unified multimodal models (UMMs) offer a direct way to obtain this feedback: the network that generates an image can also inspect it (Xie et al., 2025a; Deng et al., 2025). In particular, pixel-space UMMs such as SenseNova-U1 (Diao et al., 2026) eliminate the need for a separate visual encoder and integrate near-lossless visual decoding directly into end-to-end DiT training. This de-

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.

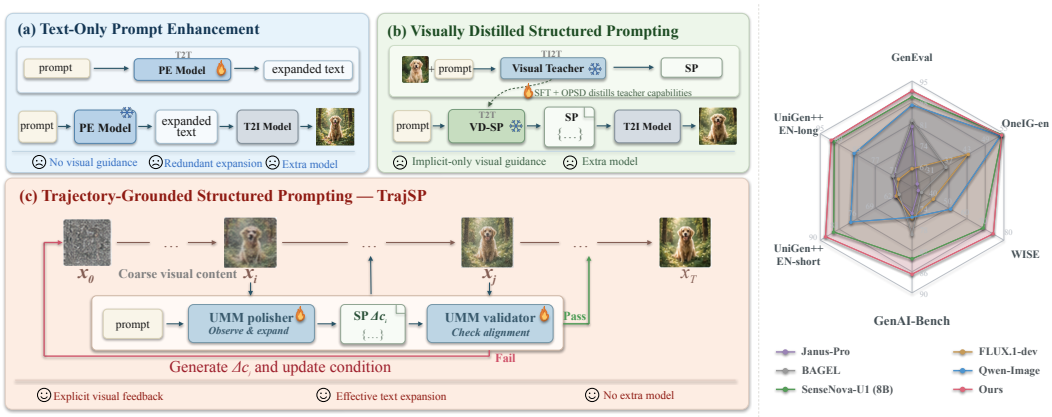

Figure 1: **Prompt-conditioning paradigms and generation overview.** (a) Text-only expansion; (b) visually distilled structured prompting (VD-SP), whose single-shot enhancer is image-free at inference; (c) TRAJSP observes its own trajectory to polish and validate conditioning. The radar summarizes the selected results in Table 1.

sign enables intermediate noisy trajectory states to be decoded directly in pixel space, avoiding the potential distribution mismatch that arises when a VAE decoder trained on clean latents is applied to noisy latent states. On the understanding side, a lightweight noisy-pixel adapter further equips the model to interpret these partially denoised observations, making intermediate trajectory states directly accessible as visual feedback during generation.

Building on this ability to directly inspect and understand intermediate generation states, we propose TRAJSP, which uses visual evidence at two points during sampling. An early Polisher reads the current noisy image together with the original prompt c_0 and predicts a structured prompt schema Δc_i. The schema provides specific conditioning for by describing object identity and attributes, geometry, relations for the remaining steps with less need for ungrounded, repetitive expansion. A late Validator checks semantic alignment and visible quality when the image is easier to recognize. If a detected error cannot be corrected within the remaining trajectory, it returns a revised schema Δc_j and triggers one regeneration. Both roles reuse the UMM; no separate prompt enhancer is served.

Learning this trajectory-grounded feedback mechanism requires supervision for both understanding and generation. We construct GSP-30K, an image–prompt–schema dataset combining synthetic examples and real image–text pairs. A correcting VLM aligns proposed schemas with the actual images. Noisy-understanding warmup first prepares the visual pathway, followed by joint schema prediction and image generation. Prompt masking in the understanding task discourages copying requested content that is unsupported by the visible image. The generator learns from both the original prompt and schema-augmented conditioning.

In summary, our contributions are threefold:

1. **A trajectory-grounded UMM Polisher.** Early noisy visual evidence is converted into a compact semantic and geometric schema, allowing the remaining denoising steps to respond to the sample's emerging composition.

2. **A late UMM Validator.** A shared understanding pathway checks alignment and quality, passes satisfactory samples onward, and returns irrecoverable failures for conditional regeneration with revised conditioning.

3. **Grounded data and joint adaptation.** GSP-30K and two-stage training connect noisy-image understanding with schema-conditioned generation. The timing sweep favors early intervention; component comparisons show gains beyond prompt-only finetuning on the reported metrics.

2 Related Work

Prompt enhancement and structured conditioning. Prompt enhancement has been widely explored as a way to strengthen textual conditioning for text-to-image generation. Promptist and BeautifulPrompt learn to rewrite user prompts into more descriptive forms (Hao et al., 2023; Cao et al., 2023), while DALL-E 3 improves prompt following through highly descriptive training captions (Betker et al., 2023). Beyond free-form rewriting, layout-based methods make objects and spatial constraints explicit to provide more structured control (Feng et al., 2023; Lian et al., 2024; Yang et al., 2024; Feng et al., 2024). More recently, Chen et al. (2026b) aim to make prompt enhancement more informative and structured by transferring image-conditioned supervision into a standalone prompter. As illustrated in Fig. 1(b), they train an image-free prompter from image-conditioned supervision using SFT, cold-start alignment, and verifier-gated on-policy self-distillation. We refer to this approach as visually distilled structured prompting (VD-SP). Although visual information is introduced during training, the resulting prompter still operates from text alone at inference and requires a separate prompt-enhancement model. In contrast, our Polisher directly reads the current noisy image during sampling, obtaining instance-specific evidence from the ongoing trajectory and updating the conditioning within the same generation process.

Unified models and noisy visual understanding. UMMs combine understanding and generation through discrete or continuous visual representations (Chameleon Team, 2024; Zhou et al., 2024; Xie et al., 2025a; Wang et al., 2024; Wu et al., 2024; Deng et al., 2025). SenseNova-U1 directly models pixels and words through a near-lossless interface (Diao et al., 2026), making intermediate states accessible without a VAE decoding stage. Prior studies show that diffusion representations carry useful recognition and correspondence information (Li et al., 2023; Tang et al., 2023; Luo et al., 2023). Our task additionally requires turning noisy visual evidence into a schema that the same model can consume as generative conditioning.

Unified models and noisy visual understanding. UMMs combine understanding and generation through discrete or continuous visual representations (Chameleon Team, 2024; Zhou et al., 2024; Xie et al., 2025a; Wang et al., 2024; Wu et al., 2024; Deng et al., 2025). SenseNova-U1 directly models pixels and words through a near-lossless interface (Diao et al., 2026), making intermediate states accessible without a VAE decoding stage. Prior studies show that diffusion representations carry useful recognition and correspondence information (Li et al., 2023; Tang et al., 2023; Luo et al., 2023). More recently, UniDDT explicitly introduces noisy visual understanding into a unified multimodal model by using a Noisy ViT encoder to extract semantics from noisy latent representations shared by understanding and generation (Wang et al., 2026). Our task goes one step further by turning such noisy visual evidence into a structured schema that the same model can directly consume as generative conditioning.

Visual feedback and self-correction. Recent studies have begun to introduce feedback and iterative reasoning into image generation. Interleaving Reasoning Generation (IRG) alternates textual reasoning with image synthesis, using reflection on an initial generation to guide a subsequent refinement (Huang et al., 2026). Thinking-while-Generating (TwiG) further interleaves textual reasoning throughout generation, allowing the model to reason about synthesized content while guiding subsequent visual generation (Guo et al., 2025). More closely related to trajectory-level correction, PPAD employs an external multimodal LLM as a semantic observer that inspects intermediate diffusion states and converts detected inconsistencies into signals for the remaining denoising steps (Lv et al., 2025). From a control perspective, Zhao et al. (2026) formulate reference-consistent generation as a closed-loop system and iteratively adjust latent control signals according to discrepancies between the generated output and the reference.

Our approach instead integrates prompt enhancement directly into the generation trajectory, using the UMM itself to provide explicit visual feedback for grounded structured prompt enhancement, rather than relying on interleaved reasoning, an external semantic observer, or latent-space control. The Polisher converts an early noisy state into conditioning for subsequent denoising, while the Validator checks late-stage failures and triggers conditional regeneration when needed.

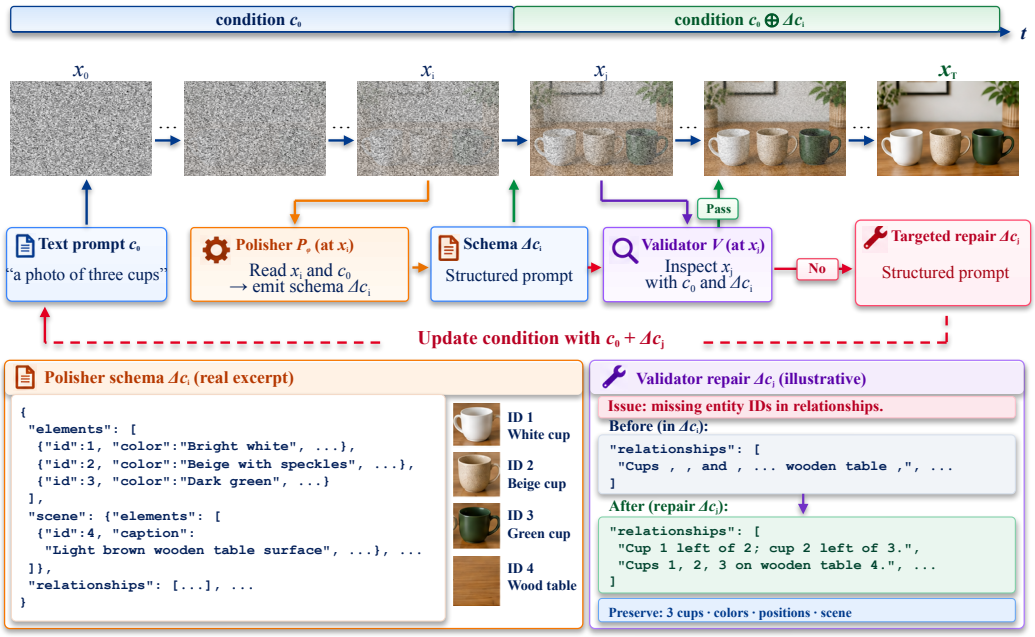

Figure 2: **Overview of the proposed framework.** The Polisher converts an early noisy state x_i and the original prompt c_0 into a structured schema Δc_i, which augments the conditioning for the remaining denoising steps. A late Validator then checks the emerging image against c_0 and Δc_i: passing samples continue to completion, whereas detected failures produce a targeted repair Δc_j and trigger one conditional regeneration. The lower panels show a real Polisher schema and an illustrative Validator repair.

3 METHOD

3.1 FRAMEWORK

As illustrated in Fig. 2, we build TRAJSP on SenseNova-U1-8B-MoT (Diao et al., 2026), using a shared pixel-space backbone for both visual understanding and image generation. Let c_0 denote the original text prompt and Δc a structured prompt schema. We index the sampling trajectory by denoising progress, such that x_0 is the initial noise and x_T is the final image. The framework introduces two trajectory-aware interventions: an early Polisher at step i and a late Validator at step j, with $i < j$. The former converts partially formed visual evidence into grounded generative conditioning, whereas the latter evaluates the more recognizable late-stage image and determines whether targeted regeneration is required.

Early polishing and late validation. At the early intervention step i, the Polisher expert P_ϕ jointly examines the current noisy state x_i and the original prompt c_0, and predicts a structured prompt schema:

$$\Delta c_i = P_\phi(x_i, c_0, t_i), \qquad c_i = c_0 \oplus \texttt{<schema>}\Delta c_i\texttt{</schema>}. \tag{1}$$

Where $\oplus$ denotes concatenation of the serialized schema with the original prompt. As illustrated in Fig. 2, Δc_i organizes the visual evidence available at the current trajectory state into complementary fields, including entities and their attributes, geometric information such as normalized locations and spatial layout, and relations among objects and the scene (Appendix A). Rather than freely elaborating every possible detail in c_0, the Polisher selectively supplements the original request with information supported by the emerging image. In this way, the updated condition preserves the user's intended content while grounding subsequent denoising in the composition, attributes, and relations that have already begun to form. The remaining trajectory is then sampled under c_i.

At the later step j, the Validator expert inspects $(x_j, c_0, \Delta c_i)$, when semantic content and visual quality are more reliably observable. It evaluates prompt–image consistency as well as visible generation defects. If the sample passes validation, the current trajectory continues to completion with-

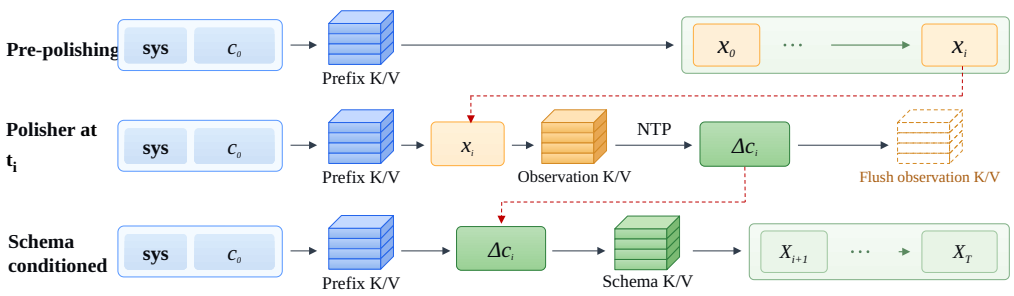

Figure 3: **KV-cache refresh.** Reuse the original prefix, decode a schema from a temporary observation branch, flush its visual cache, and rebuild schema conditioning before continuing generation.

out modification. Otherwise, when the detected error is unlikely to be corrected by the remaining denoising steps, the Validator produces a targeted repair schema Δc_j. As shown in Fig. 2, this repair preserves image-consistent fields while correcting mismatched relations, attributes, or missing content. We then restart sampling once from fresh noise using $c_0 \oplus$ <schema>Δc_j</schema> as the revised condition. Importantly, Δc_j replaces the conflicting fields in Δc_i rather than being blindly appended to it, avoiding contradictory conditioning. The validator is invoked once for every sample and therefore introduces a fixed inference overhead, whereas the additional generation cost is incurred only for samples that trigger regeneration.

Sequence interface and cache refresh. The post-polishing generation sequence is

$$\underbrace{\texttt{[sys]} \oplus c_0}_{\text{fixed prefix}} \oplus \texttt{<schema>}\Delta c_i\texttt{</schema>} \oplus \texttt{<img>}\, x_k \,\texttt{</img>}. \tag{2}$$

Chat delimiters are omitted. Figure 3 illustrates the corresponding KV-cache refresh. Before polishing, the fixed system–prompt prefix is encoded once and its K/V cache is reused during generation. At step i, the Polisher reuses this prefix cache, temporarily appends the current state x_i as an observation, and autoregressively predicts the schema Δc_i. The resulting observation K/V is then discarded, so the noisy visual tokens do not persist as generation context. The accepted schema is instead encoded after the same fixed prefix to form a new schema K/V cache, from which image generation resumes at x_{i+1}. Thus, only the structured visual evidence extracted by the Polisher is committed to subsequent conditioning, while the temporary observation itself is flushed. Invalid schemas fall back to c_0, and the same serialization and cache-refresh procedure is used for revised late-stage conditioning.

3.2 Two-Stage Training

Stage 1: noisy understanding warmup. A timestep-conditioned Transformer adapter maps noisy pixels to the UMM's visual interface, while the UMM remains frozen (Figure 4a). Noisy VQA examples retain the clean image's question and answer while varying the noise strength. This stage trains the adapter to extract effective semantic information across different noise levels.

Stage 2: joint understanding and generation training. For understanding, $(x_t, \widetilde{c}_0, t)$ predicts Δc with token cross-entropy $\mathcal{L}_{\text{und}}$. Masking selected prompt tokens in $\widetilde{c}_0$ discourages the model from reproducing requested content that is unsupported by the visible evidence. For generation, either the original prompt c_0 or the schema-augmented prompt $c_0 \oplus$ <schema>Δc</schema> conditions U1's pixel-space flow-matching loss $\mathcal{L}_{\text{gen}}$. Plain

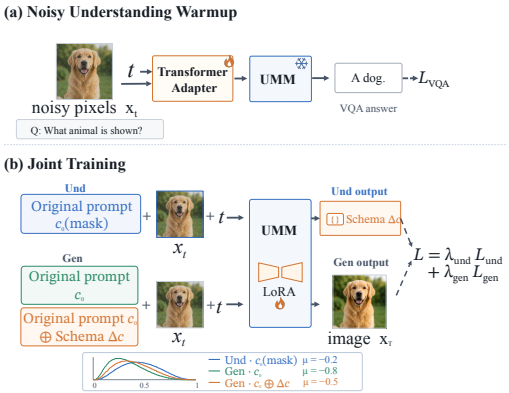

Figure 4: **Two-stage training.** (a) Noisy understanding warmup. (b) Joint schema prediction and image generation. See Appendix B for settings.

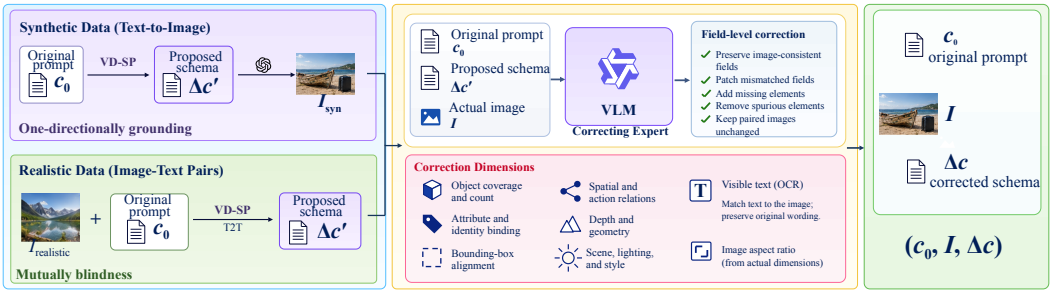

Figure 5: **GSP-30K construction pipeline.** Synthetic and real-data branches first produce structured-prompt proposals $\Delta c'$ with different grounding limitations. A VLM correcting expert aligns each proposal with the actual image through field-level correction of objects, attributes, geometry, relations, visible text, and image format, producing grounded $(c_0, I, \Delta c)$ triplets for training.

and schema-conditioned examples preserve ordinary prompting and teach the model to effectively use structured conditioning, respectively. Shared LoRA parameters and selected visual modules are jointly optimized with

$$\mathcal{L} = \lambda_{\text{und}}\mathcal{L}_{\text{und}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}}. \tag{3}$$

The task-specific timestep sampling for understanding, plain generation, and schema-conditioned generation is illustrated in Figure 4b. Detailed training settings are provided in Appendix B. The Validator operates at low noise and directly reuses the shared understanding pathway without dedicated Validator finetuning.

3.3 GSP-30K: Grounded Structured Prompts

We construct GSP-30K from 15k synthetic images and 15k real image–text pairs, totaling 30k examples (Figure 5). In the synthetic branch, VD-SP expands c_0 into a proposal $\Delta c'$ before image synthesis; the rendered image can still deviate from that proposal, resulting in one-directional grounding. In the real-data branch, the original image is retained while VD-SP expands only its text, so the proposal has no direct access to the paired image, resulting in mutual blindness.

To resolve the grounding mismatch in both branches, we introduce a VLM correcting expert that explicitly aligns each proposed schema with its corresponding image. The correcting expert receives $(c_0, I, \Delta c')$ and produces the image-grounded schema Δc. It preserves supported fields, repairs mismatched attributes, boxes, and relations, adds missed objects, and removes spurious content while keeping I unchanged. OCR fields preserve visible wording, and aspect ratio follows the actual image dimensions. The resulting $(c_0, I, \Delta c)$ triples support both noisy schema prediction and schema-conditioned generation.

4 Experiments

4.1 Settings

We use SenseNova-U1-8B-MoT and the two-stage adaptation described above. Appendix B gives the training and inference settings in a stage-wise format following the U1 technical report (Diao et al., 2026). We choose Polisher step $i = 7$ and Validator step $j = 15$. Generation benchmarks include GenEval (Ghosh et al., 2023), OneIG-Bench (Chang et al., 2025), WISE (Niu et al., 2026), GenAI-Bench (Li et al., 2024), and UniGenBench++ (Wang et al., 2025).

4.2 Results

Noisy Understanding Results. We compare the base and noise-adapted understanding pathways on MME under different signal levels. As shown in Figure 6, noise adaptation substantially improves understanding under moderate and strong corruption, with gains of over 500 points at several intermediate noise levels. The advantage gradually diminishes as the input approaches a clean image, indicating that the adaptation primarily improves robustness to noisy visual states while largely preserving clean-image understanding. Qualitative examples further show that the adapted pathway can

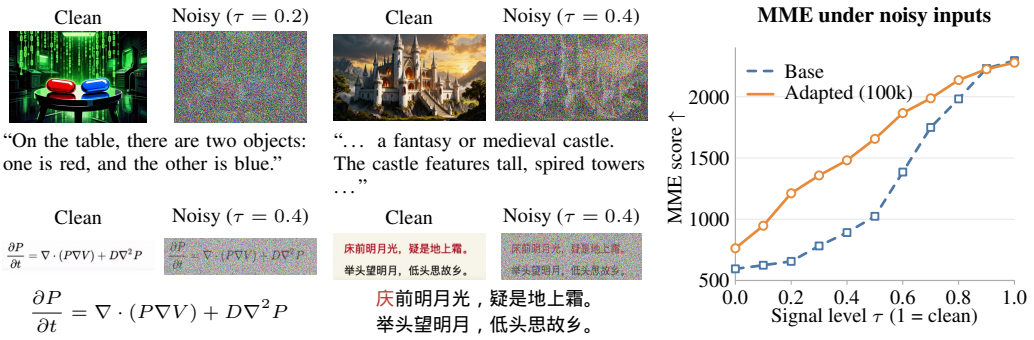

Figure 6: **Understanding noisy visual inputs.** Left: clean/noisy input pairs at $\tau = 0.2$ or 0.4, with actual model responses below; ellipses mark omitted text. Right: MME scores for the base and noise-adapted pathways.

Table 1: Comparison with open-source image generation models.

Model / mode	GenEval	OneIG-en	WISE	GenAI-Bench	UniGenBench++	
					EN-short	EN-long
Janus-Pro (Chen et al., 2025)	80.00	26.70	33.40	75.00	61.36	71.11
FLUX.1-dev (Black Forest Labs, 2024)	66.00	43.40	41.60	76.00	60.97	69.42
Show-o2 (Xie et al., 2025b)	76.00	30.80	61.00*	75.00	61.90	70.33
BAGEL (Deng et al., 2025)	82.00	36.10	35.20	79.00	59.91	71.26
BAGEL w/ self-CoT (Deng et al., 2025)	88.00	32.40	62.80	81.00	–	–
Qwen-Image (Wu et al., 2025)	87.00	53.90	51.00	75.70	78.36	83.94
Z-Image (Z-Image Team et al., 2025)	84.00	54.60	45.30	–	78.10	86.77
Emu3.5 (Cui et al., 2025)	73.00	**56.40**	58.00*	–	–	–
HiDream-O1-Image (Cai et al., 2026)	90.00	–	–	–	–	–
ERNIE-Image (Liu et al., 2026)	89.00	55.40	–	–	–	–
LLaDA-Image (Chen et al., 2026a)	85.00	–	–	–	–	–
SenseNova-U1 (8B) (Diao et al., 2026)	89.69	54.38	68.80	83.56	84.97	90.52
Ours	**91.76**	54.30	**74.20**	**86.50**	**88.00**	**91.45**

Published WISE results follow the updated WISE Verified leaderboard with Qwen3.5-35B-A3B as the judge. $*$: Show-o2 and Emu3.5 use the legacy WISE version.

recover scene identity, appearance, object count and layout, as well as color and shape cues from partially denoised inputs. These results validate that intermediate noisy states contain sufficiently reliable semantic information to support the Polisher.

Generation Results. Table 1 compares our method with open-source image generation models. Relative to the U1 baseline, our method improves GenEval by 2.07, GenAI-Bench by 2.94, and UniGenBench++ EN-short/EN-long by 3.03/0.93 points, while OneIG-en remains nearly unchanged (54.30 versus 54.38). As further shown in Table 2, the gains mainly arise from improvements in spatial reasoning, action understanding, and especially counting.

4.3 Ablation Studies

4.3.1 Polisher and Validator intervention steps

Figure 7 studies the intervention timing for both the Polisher and Validator over denoising steps $k \in \{5, 7, 10, 15, 20\}$. As shown in Figure 7(a), injecting the structured condition at $k = 7$ yields the strongest performance across several semantic dimensions, including spatial relations, actions, and style. In contrast, counting degrades consistently as the intervention is delayed, indicating that some semantic errors become increasingly difficult to correct once the generation trajectory has progressed. We therefore choose $i = 7$ for the Polisher, leaving sufficient denoising steps for the image-grounded schema to affect generation.

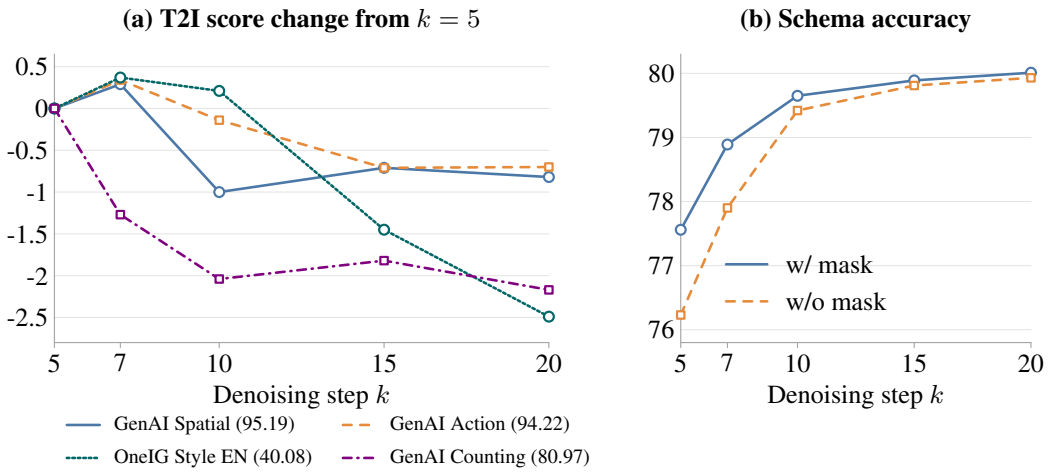

Figure 7: **Polisher timing: generation quality and schema accuracy.** (a) Four T2I metrics on one axis, shown as score differences $s(k) - s(5)$ in percentage points; legend values give the original scores at $k = 5$. (b) Schema accuracy with and without prompt masking.

Table 2: Component and external-prompter comparison. Spatial, Action, and Counting are GenAI-Bench dimensions. Bold denotes the best U1 component row, excluding external VD-SP.

Model	WISE	GenAI	Spatial	Action	Counting
U1	68.80	83.56	95.07	93.80	77.76
+ Finetune	66.80	83.73	95.06	93.81	77.50
+ Polisher	72.90	86.42	95.19	**94.22**	80.97
+ Validator	**74.20**	**86.50**	**96.00**	93.93	**81.98**
VD-SP	80.30	86.43	94.68	93.13	86.14

For the Validator, Figure 7(b) shows that schema accuracy continues to improve at later denoising steps, but with diminishing returns after $k = 15$. We therefore set $j = 15$, where the visual content is already sufficiently recognizable for reliable schema revision while avoiding an unnecessarily late intervention. When the Validator identifies errors that are unlikely to be recovered by the remaining trajectory, such as incorrect object counting, the revised schema is used to trigger conditional regeneration from fresh noise.

4.3.2 Prompt masking for noisy understanding

Figure 7(b) further evaluates prompt masking during noisy-understanding training. Across all evaluated noise levels, masking produces consistently higher structured-schema accuracy than training with the complete prompt, with the largest gains occurring at earlier, noisier states. This supports our motivation that masking discourages the understanding pathway from copying textual cues when they are not yet supported by visible evidence. Importantly, prompt masking is used only as a training strategy; at inference time, both the Polisher and Validator always receive the complete, unmasked original prompt.

4.3.3 Finetuning, polishing, and external VD-SP

Table 2 compares direct U1 generation, prompt-only finetuning, our Polisher/Validator, and external VD-SP. Prompt-only finetuning brings little improvement, whereas the Polisher substantially improves WISE, GenAI-Bench, and counting, indicating that the gains mainly come from trajectory-grounded visual feedback rather than SFT alone. Adding the Validator further improves spatial relation and counting, while slightly reducing action relation.

Compared with external VD-SP, our full method achieves comparable overall GenAI-Bench performance and higher spatial and action scores, but remains lower on WISE and counting. This difference suggests that an external prompter retains advantages from reasoning before generation, whereas our method focuses on conditioning grounded in the ongoing visual trajectory.

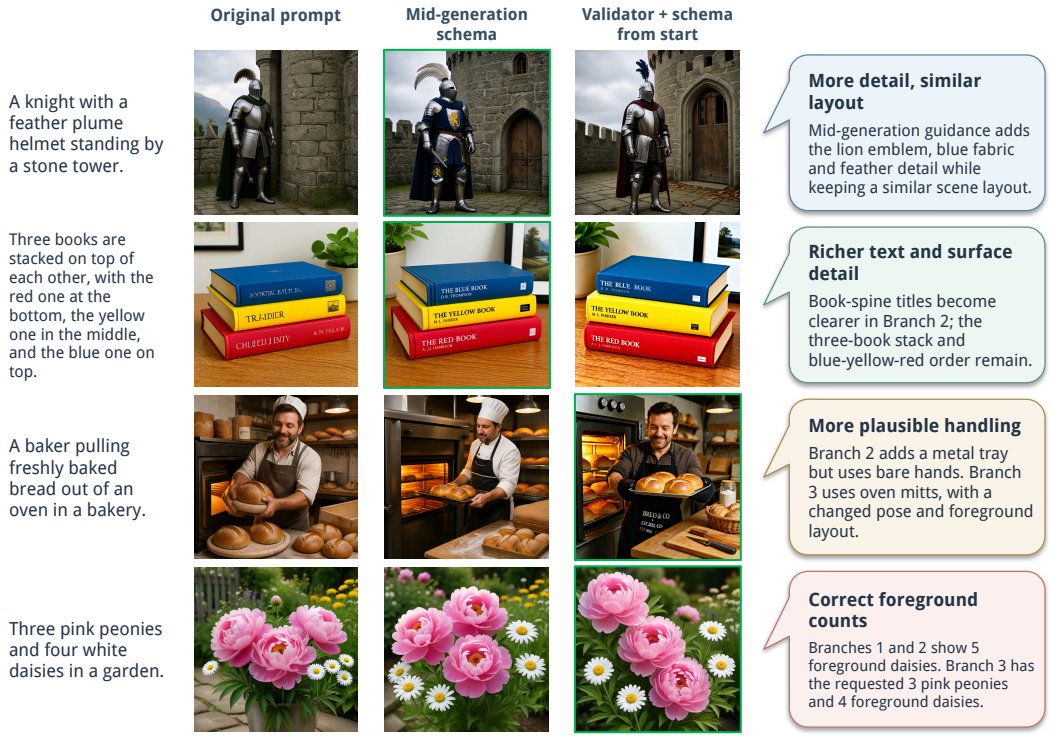

Figure 8: **Qualitative comparison of polishing and validation.** From left to right: original-prompt generation, Polisher-guided continuation, and Validator-guided regeneration. Polishing enriches grounded details, while validation enables stronger semantic correction.

Qualitative results. Figure 8 compares three inference settings to illustrate the complementary effects of the Polisher and Validator. The first column shows standard generation conditioned only on the original prompt c_0. The second column applies the Polisher at an intermediate denoising step and continues the same trajectory with the resulting structured schema, without invoking the Validator. The third column additionally performs a late-stage validation and, for this qualitative comparison, forces one regeneration from fresh noise using the revised schema produced by the Validator.

The comparison highlights two distinct roles. Mid-generation polishing primarily enriches image-grounded attributes, text, and local details while largely preserving the composition already established by the trajectory. Validator-guided regeneration provides a stronger corrective mechanism when the current trajectory contains semantic errors that are difficult to recover through the remaining denoising steps, particularly for action plausibility and object counting.

5 Conclusion

TrajSP enables a pixel-space UMM to leverage visual understanding throughout generation. An early Polisher transforms noisy intermediate evidence into structured conditioning, while a late Validator assesses semantic alignment and visual quality before selectively regenerating failed samples. GSP-30K provides 15k synthetic and 15k real examples with image-corrected schemas for two-stage adaptation and joint training. Experiments demonstrate consistent generation improvements beyond prompt-only finetuning, with intervention at step 7 providing a favorable trade-off between schema reliability and remaining correction capacity. Prompt masking further improves schema accuracy, particularly at early denoising stages. Comparison with VD-SP reveals complementary strengths: trajectory-grounded feedback better supports spatial and action alignment, whereas pre-generation reasoning remains beneficial for knowledge-intensive generation.

AI USE STATEMENT

Generative AI was used to construct GSP-30K: VD-SP expands prompts into structured proposals, image generation produces the synthetic branch, and a VLM correcting expert revises annotations against the paired images (Section 3.3). Annotation quality control includes schema parsing, box-bound and entity-reference checks, and VLM-based checks of image–schema consistency (Appendix A). AI tools also assisted manuscript drafting and language polishing, formatting, plotting, and figure preparation. We did not use generative AI to design the method or experiments, implement the proposed method, interpret experimental results, or translate the manuscript; mathematical proof assistance is not applicable to this work. The authors take responsibility for the final manuscript, including the accuracy and originality of its text, claims, results, and AI-assisted artifacts.

ETHICS STATEMENT

This work studies text-to-image generation using synthetic images, existing image–text pairs, and benchmark evaluations. The pretrained backbone and AI-assisted annotations may inherit or amplify social and demographic biases, and image-grounded correction does not guarantee factual accuracy or fairness. Improved prompt alignment can also facilitate misleading imagery, impersonation, or harmful content. The Validator evaluates semantic alignment and visual quality and should not be treated as a safety filter. Downstream use therefore requires appropriate content safeguards and assessment of bias and misuse. Use or redistribution of data, annotations, and model artifacts must respect source licenses, intellectual-property rights, and applicable privacy requirements, particularly for identifiable people or sensitive text in real images. Any release should be limited to materials whose redistribution is permitted.

REPRODUCIBILITY STATEMENT

Section 3.1 specifies the trajectory interventions, schema conditioning, cache refresh, and invalid-schema fallback, while Section 3.2 describes the two-stage training procedure and objectives. Section 3.3 and Appendix A describe the synthetic and real-data branches, image-grounded annotation correction, schema fields, coordinate conventions, and consistency checks. Appendix B and Table 4 report the trainable modules, optimizer and adaptation settings, task mixture, and noise schedules. Appendix C describes the intervention protocol, sampling implementation, and noisy-understanding evaluation. Section 4.1 identifies the backbone and evaluation benchmarks, and Section 4.3 reports the intervention-step, prompt-masking, and component comparisons. Together, these references document the method and reported experimental setup and provide the basis for reproducing the training and evaluation procedures.

REFERENCES

James Betker, Gabriel Goh, Li Jing, et al. Improving image generation with better captions. Technical report, OpenAI, 2023.

Black Forest Labs. Announcing black forest labs. Official FLUX.1 release announcement, 2024. URL https://bfl.ai/blog/24-08-01-bfl.

Qi Cai, Jingwen Chen, Chengmin Gao, et al. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026. URL https://arxiv.org/abs/2605.11061v1.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): Industry Track*, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Jingjing Chang, Yixiao Fang, Peng Xing, et al. OneIG-Bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025.

Chuyan Chen, Haoxing Chen, Kun Chen, et al. LLaDA-Image: Building strong image generators with fully open training recipes. *arXiv preprint arXiv:2609.03796*, 2026a. URL https://arxiv.org/abs/2609.03796.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, et al. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Zilong Chen, Chaorui Deng, Kunchang Li, Hongyi Yuan, and Haoqi Fan. Scaling properties of text conditioning in visual generation. *arXiv preprint arXiv:2607.29679*, 2026b.

Yufeng Cui, Honghao Chen, Haoge Deng, et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025. URL https://arxiv.org/abs/2510.26583v1.

Chaorui Deng, Deyao Zhu, Kunchang Li, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, et al. SenseNova-U1: Unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026. URL https://arxiv.org/abs/2605.12500.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, et al. LayoutGPT: Compositional visual planning and generation with large language models. In *NeurIPS*, 2023.

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *CVPR*, 2024.

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

Ziyu Guo, Renrui Zhang, Hongyu Li, Manyuan Zhang, Xinyan Chen, Sifan Wang, Yan Feng, Peng Pei, and Pheng-Ann Heng. Thinking-while-generating: Interleaving textual reasoning throughout visual generation. *arXiv:2511.16671*, 2025.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *NeurIPS*, 2023.

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, Junbo Qiao, Yue Guo, Yao Hu, Zhenfei Yin, Philip Torr, Yu Cheng, Wanli Ouyang, and Shaohui Lin. Interleaving reasoning for better text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, et al. GenAI-Bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

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

Jiaxiang Liu, Zhida Feng, Pengyu Zou, et al. ERNIE-Image technical report. *arXiv preprint arXiv:2605.25347*, 2026. URL https://arxiv.org/abs/2605.25347.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2023.

Zheqi Lv, Junhao Chen, Qi Tian, Keting Yin, Shengyu Zhang, and Fei Wu. Multimodal llm-guided semantic correction in text-to-image diffusion, 2025. URL https://arxiv.org/abs/2505.20053.

Yuwei Niu, Munan Ning, Mengren Zheng, et al. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. In *ICML*, 2026.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023.

Shuai Wang, Liang Li, Yang Chen, Ruopeng Gao, Yao Teng, and Limin Wang. Uniddt: Unifying multimodal understanding and generation with decoupled diffusion transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yibin Wang, Zhimin Li, Yuhang Zang, et al. UniGenBench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324v1.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, et al. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025b. URL https://arxiv.org/abs/2506.15564.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *ICML*, 2024.

Z-Image Team, Huanqia Cai, Sihan Cao, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699v1.

Baixuan Zhao, Xinyu Zhang, Huayu Zheng, Shuaicheng Liu, Xiongkuo Min, Guangtao Zhai, and Xiaohong Liu. From open loop to closed loop: A test-time iterative optimization framework for reference-consistent image generation, 2026. URL https://arxiv.org/abs/2607.04691.

Chunting Zhou, Lili Yu, Arun Babu, et al. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A SCHEMA AND DATA DETAILS

Schema fields. A structured prompt Δc is a serialized object with nine top-level fields. Table 3 summarizes their roles. Element attributes may include visible text; OCR content preserves the wording in the image rather than replacing it with text suggested by the original prompt. Entity IDs provide a common reference for element descriptions, boxes, and relationships.

Table 3: Fields of the structured prompt used for image-grounded conditioning.

Field	Content
`ratio`	Aspect ratio of the paired image.
`intent`	Overall scene description and intended content.
`atmosphere`	Scene mood and ambient appearance.
`style`	Visual or artistic style.
`lighting`	Illumination and visible lighting effects.
`elements`	Entity IDs, descriptions, attributes, and spatial boxes.
`relationships`	Spatial, action, and other relations between entities.
`scene`	Scene context and background content.
`photography`	Viewpoint, framing, and photographic properties.

Coordinates and serialization. Boxes use a top-left origin, with x increasing to the right and y increasing downward. Coordinates are normalized to $[0, 999]$ relative to the image width and height. A box is serialized using the backbone's existing box delimiters:

`<|box_start|>(x1,y1),(x2,y2)<|box_end|>`.

The corners must satisfy $0 \leq x_1 < x_2 \leq 999$ and $0 \leq y_1 < y_2 \leq 999$. These coordinates describe image geometry, not pixel indices in a fixed-resolution canvas. The schema is enclosed by the `<schema>` and `</schema>` boundaries when concatenated with c_0, as in Equation 2.

Image-grounded annotation correction. The synthetic and real branches each contribute 15k examples to GSP-30K. For each proposal, the correcting expert receives $(c_0, I, \Delta c')$ and returns Δc while keeping I fixed. It checks object counts, attributes, locations, relationships, background content, and visible text. Supported fields are retained; mismatched fields are repaired, missed objects are added, and unsupported objects are removed. The aspect ratio is set from the paired image. This correction makes the schema a description of the training image, including cases where a synthetic image does not fully realize the original prompt. The same corrected triple $(c_0, I, \Delta c)$ supports noisy schema prediction and schema-conditioned generation.

Structural checks and inference fallback. Checks cover parseability, coordinate bounds and corner order, duplicate entity IDs, dangling entity references, and conflicting relations. These structural checks complement image-based semantic correction; a parseable schema alone does not establish image–text consistency. At inference, an invalid schema falls back to the original condition c_0. An accepted late revision replaces the previous schema span. The temporary observation K/V is discarded, and only the fixed prefix and accepted schema are retained as conditioning, following Section 3.1.

B Training and Implementation Details

Optimization and batching. Figure 4 specifies the two-stage training procedure. Table 4 reports its adaptation settings in the stage-wise format of the U1 report (Diao et al., 2026). Both stages use AdamW with base learning rate 10^{-5}, coefficients $(0.9, 0.95)$, epsilon 10^{-8}, weight decay 0.01, gradient clipping 1.0, and BF16. The base learning rate follows a cosine schedule ending at 10^{-6} after linear warmup. LoRA parameters use the learning-rate multiplier and weight-decay override in the table. Image areas range from 256^2 to 2048^2 pixels; resizing preserves aspect ratio with dimensions aligned to multiples of 32.

Table 4: Stage-specific adaptation settings. Warmup batches count images; joint-training batches count packed sequences. "N/A" indicates a mechanism that is not used in that stage. Hardware and global batch sizes describe the recorded reference allocation.

Setting	Noisy warmup	Joint training
Trainable modules	Pixel adapter	LoRA + Pixel adapter
Steps / warmup steps	100,000 / 2,000	15,000 / 500
Microbatch / accumulation	16 / 2	1 / 1
Reference hardware	96 NPUs	64 NPUs
Reference global batch	3,072 images	64 sequences
Tokens / images per pack	N/A	16,384 / at most 64
LoRA rank / alpha / dropout	N/A	64 / 32 / 0.05
LoRA LR multiplier / decay	N/A	$5\times$ / 0
Understanding prompt mask	N/A	0.3
Understanding : plain : schema	N/A	1 : 1 : 1 by examples
$\lambda_{\text{und}} : \lambda_{\text{gen}}$	N/A	0.2 : 1

Objectives and trainable modules. During warmup, U1 remains frozen and the timestep-conditioned pixel adapter is trained using noisy VQA, as described in Section 3.2 and Figure 4(a). The question and answer are retained from the clean-image example while the visual input is corrupted at different signal levels. Joint training uses token cross-entropy for schema prediction and pixel-space flow matching for image generation, with the weights in Equation 3. LoRA is applied to attention query, key, value, and output projections and to feed-forward projections. The pixel adapter, flow head, timestep and noise-scale embeddings, generation visual embeddings, and selected normalization parameters are also trainable; other parameters remain fixed.

Task mixture and prompt masking. The $1 : 1 : 1$ mixture is defined over examples of understanding, plain text-to-image generation, and schema-conditioned generation, rather than tokens or loss weights. Packing uses a 16,384-token budget and at most 64 images per sequence; the image limit is not the effective batch size. Prompt masking with ratio 0.3 affects only the original prompt in the understanding task. The schema target and the two generation conditions are not masked. Both inference experts receive the full original prompt. The Validator reuses the shared understanding pathway without dedicated Validator finetuning.

Signal levels and noise scale. Let $s \in [0, 1]$ denote signal level, distinct from the integer denoising step k. For a clean RGB image I scaled to $[-1, 1]$ and independent standard Gaussian noise ϵ, the corrupted input is

$$x(s) = sI + (1 - s)\sigma_R\epsilon, \qquad \sigma_R = \min\left(8, \sqrt{N_{\text{img}}/64}\right), \tag{4}$$

where $N_{\text{img}} = (H/32)(W/32)$ is the image-token count after spatial merging. Thus $s = 0$ is pure noise and $s = 1$ is the clean image. For the logistic-normal component, sample $u \sim \mathcal{N}(\mu, 0.8^2)$ and set $s = \text{sigmoid}(u)$. The warmup curriculum keeps $\mu = 0.8$ for the first 30% of training and then decreases it linearly to 0.0. It also retains clean inputs with probability 0.05 and uses a strong-noise component with probability 0.10 and signal levels at most 0.5. In joint training, the task-specific means are -0.2, -0.8, and -0.5 for understanding, plain generation, and schema-conditioned generation, respectively, following Figure 4(b).

C Inference and Evaluation Details

Intervention and restart protocol. Denoising steps count completed updates from the initial noise x_0 toward the final image x_T. The default intervention steps are $i = 7$ for the Polisher and $j = 15$ for the Validator. Polishing continues the same trajectory under the accepted schema; validation either allows completion or triggers at most one restart from fresh noise under the revised schema. The component comparison in Table 2 separately uses a Polisher step of $k = 5$. The third column of Figure 8 forces one Validator-guided regeneration, as described in the main text; it does not measure the frequency of automatic restart decisions.

Sampling implementation and seeds. The archived generation runs use 2048×2048 images and 50 denoising updates. The reference implementation advances the flow using explicit Euler updates on the backbone's transformed timestep grid. Its default classifier-free guidance scale is 4.0, with no guidance normalization, and the default timestep-shift parameter is 3.0. These guidance and shift values are implementation defaults; the archived per-image manifests do not record run-specific overrides. Each manifest records the actual sample seed, benchmark, prompt ID, sample index, image dimensions, schema path, and checkpoint paths. Replaying a sample uses that recorded seed rather than a single shared seed for every image. An intervention step is an index in the sampling grid and should not be identified with signal level $s = k/T$.

Noisy-understanding evaluation. Figure 6 evaluates the base understanding pathway and the adapted pathway at $s \in \{0, 0.1, \ldots, 1\}$. Both receive reference images corrupted according to Equation 4; the adapted pathway is supplied with the known signal level. Images retain their native resolution rounded to multiples of 32, without minimum-size upscaling, with a maximum area of 4096^2 pixels. Each condition contains all 2,374 question records in the evaluated MME set. The plotted values are raw MME scores, not percentages or per-question accuracy. At $s = 0.5$, the score increases from 1,024.18 to 1,656.47 (a gain of 632.29); at $s = 1$, it changes from 2,295.98 to 2,278.89. These results characterize robustness to corrupted reference images. They do not directly measure schema correctness on actual generation trajectories, which is evaluated separately in Figure 7(b).