TRUST-TO-LATENT: ALIGNING LATENT GENERATION WITH MULTIMODAL SEMANTIC JUDGMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The central challenge of long-prompt image generation is to realize entities, attributes, quantities, and relations jointly in a semantically coherent scene. Unified understanding–generation models offer a natural opportunity to address this challenge: their understanding capability can provide a semantic criterion for the content their generation capability seeks to realize. Yet judging completed images does not automatically enable understanding to assess the noisy representations encountered during generation. We introduce **Trust-to-Latent**, which connects these capabilities by extending reward-trained semantic judgment to intermediate generation states. A prompt-conditioned latent evidence interface makes multi-layer generation features accessible to a reader built on a frozen understanding backbone. Image-level preferences, transferred through shared-noise latent pairs, teach the reader to assess evolving visual evidence against the prompt. The complete evaluator then remains fixed while relative preferences among candidates from a common intermediate state guide generator learning. Understanding thus informs how the generator is trained to realize instructions, without reward evaluation or candidate selection at inference. Experiments across five public benchmarks show improved overall instruction following over the base model and reward-optimization baselines, with gains in attribute binding, spatial relations, and joint satisfaction of evaluated requirements. Interface ablations and paired error analyses further support this connection between semantic judgment and compositional generation.

1 INTRODUCTION

A long prompt specifies a scene through interdependent requirements: attributes must belong to the correct entities, objects must appear in prescribed quantities and positions, and later sentences may constrain earlier descriptions. An image can contain every requested object and still violate the instruction through incorrect attribute assignments or reversed relations. Generation must therefore translate dispersed linguistic requirements into visual relationships that hold together. Fine-grained evaluations reveal persistent failures in binding, spatial arrangement, and complex composition, reflecting the difficulty of realizing these requirements jointly (Huang et al., 2023; Wei et al., 2025; Jiao et al., 2025; Yang et al., 2026).

Assessing whether a scene satisfies these requirements calls for connecting visual evidence to the text. Unified understanding–generation models provide a natural basis for bringing this assessment into generative learning (Xie et al., 2025; Wu et al., 2025a): the generation branch constructs visual content, while the understanding branch interprets that content in the context of language. Reward training can equip the latter with a criterion for instruction satisfaction. The coexistence of these capabilities raises a question: how can understanding help shape generation while the requested scene is still taking form?

Reward optimization already allows judgments of completed images to influence preceding generation decisions (Xu et al., 2023; Kirstain et al., 2023; Liu et al., 2025). Directly assessing intermediate alternatives presents a further challenge: recognizing an attribute error in a final image does not automatically confer the ability to assess attribute binding in noisy generation states. Recent latent reward models address this gap through diffusion features and connections to pretrained reward computation (Zhang et al., 2026; Go et al., 2026). We investigate how a unified model's reward-

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.

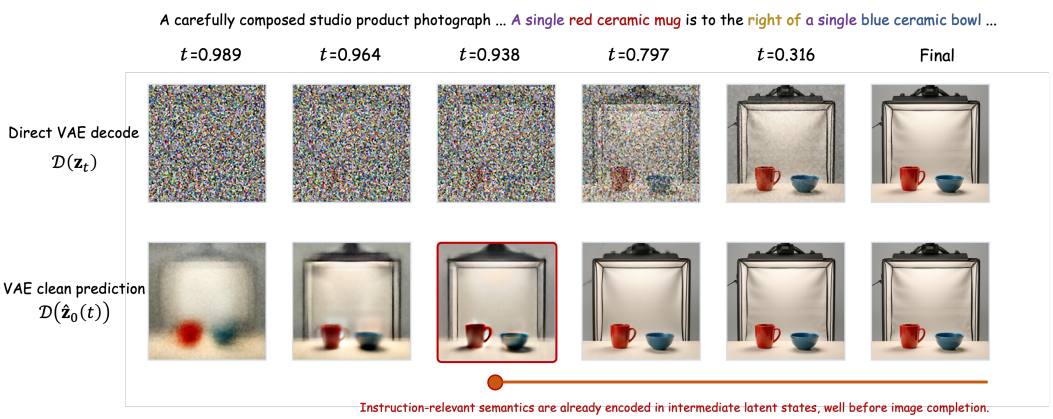

Figure 1: **A generator-assisted readout reveals a spatial violation before image completion.** The top row directly decodes noisy latents; the bottom shows prompt-conditioned clean predictions from the same trajectory. At $t = 0.938$, the clean prediction reveals the incorrect mug–bowl relation that persists to the final image. Each prediction uses the current state and generator without future denoising. Appendix C specifies this visualization protocol.

trained understanding can read evolving generation evidence and use it to support compositional instruction following.

Figure 1 gives this connection a concrete motivation. A prompt-conditioned clean prediction reveals an incorrect spatial relation before image completion, while direct decoding of the same noisy state remains difficult to interpret. The prediction uses the current state and generator without executing future denoising steps. This example suggests that intermediate generation computation can expose semantic evidence before the image is complete. We use an understanding-based reader to turn such evidence into semantic feedback.

We introduce Trust-to-Latent to bring a unified model's understanding capability into generator learning. Our central idea is to use reward-trained understanding to assess the content taking shape within generation. A prompt-conditioned latent evidence interface presents multi-layer generation features to a reader built on a frozen understanding backbone. Learned queries and reading adapters connect textual requirements with evolving visual evidence. Image-level preferences supervise this reader through latent pairs corrupted with a common noise draw and timestep, adapting how evidence is read while reusing the learned understanding computation.

These intermediate judgments must, in turn, become a learning signal for generation. Trust-to-Latent freezes the complete evaluator and updates the policy through relative comparisons among candidates from a common intermediate state. Learning the reader establishes how understanding accesses generation evidence; online optimization applies its judgment to alternatives the current policy produces.

Our contributions connect multimodal understanding with compositional generation:

- **Connecting understanding and generation through semantic feedback.** We develop a framework that makes reward-trained understanding available to intermediate generation and uses its judgments to train the generator, extending their connection beyond assessment of completed images.
- **Learning to read evolving generation evidence.** A prompt-conditioned, multi-layer interface transfers image-level preferences to noisy states through shared-noise supervision. The resulting fixed evaluator supplies local preferences for online generator optimization, with no reward computation or candidate selection at inference.
- **Advancing joint satisfaction of compositional instructions.** Five-benchmark evaluation demonstrates improved instruction following, including attribute binding, spatial relations, and joint satisfaction. Interface ablations and paired error analyses connect these gains to the proposed design and the correction of instruction violations.

2 Related Work

2.1 Compositional Generation and Instruction Following

Compositional evaluation distinguishes the presence of visual concepts from the correct realization of their attributes and relations. GenEval measures object-level properties including co-occurrence, count, color, and position (Ghosh et al., 2023). T2I-CompBench organizes evaluation around attribute binding, object relationships, and complex compositions (Huang et al., 2023), while TIIF-Bench studies fine-grained instruction following under more complex textual requirements (Wei et al., 2025). These perspectives motivate treating long-prompt generation as joint constraint satisfaction rather than text length alone. Our work addresses how semantic judgments of those requirements can inform learning from intermediate generation states.

2.2 Reward Modeling and Generative Alignment

ImageReward learns a visual preference model and uses reward feedback to tune image generation (Xu et al., 2023). Pick-a-Pic supplies preferences for learning PickScore (Kirstain et al., 2023), while HPSv3 develops a vision-language preference evaluator with broader quality coverage (Ma et al., 2025b). Diffusion-DPO adapts direct preference optimization to diffusion models (Wallace et al., 2024). DDPO formulates denoising as a sequential decision process for policy-gradient learning (Black et al., 2024), and Flow-GRPO develops online reinforcement learning for flow matching models (Liu et al., 2025). These methods establish reward judgment as a training signal beyond likelihood-based pretraining. We focus on making reward-trained understanding applicable to evolving generation representations.

2.3 Intermediate Feedback and Latent Reward Models

SPO introduces step-by-step preference optimization for diffusion models (Liang et al., 2025). LPO learns noise-aware latent rewards from diffusion representations for step-level preference optimization (Zhang et al., 2026). Stitched Value Model transfers image-level reward computation to noisy latents by connecting a frozen diffusion backbone to a pretrained reward-model suffix (Go et al., 2026). These methods establish latent reward modeling and cross-representation transfer as useful alignment tools. Our contribution is a text-conditioned evidence reader within a unified understanding–generation model: learned queries jointly access prompt memory and multiple generation-feature depths, while attention adapters expose this evidence to a frozen, reward-trained understanding backbone. Shared-noise pairwise supervision trains this reader for local preference comparisons, and the complete evaluator then remains fixed during generator optimization. This design couples compositional semantic assessment with local generation feedback; the interface ablations evaluate multi-layer evidence, reader adaptation, query memory, and shared corruption.

3 Trust-to-Latent

Trust-to-Latent connects understanding and generation through semantic feedback on intermediate states. We first train the understanding branch to assess instruction satisfaction from image-level evidence, then learn how it reads evolving generation representations through a prompt-conditioned interface. The complete evaluator subsequently remains fixed while the generator learns from preferences among its own latent continuations. These stages make understanding available to generation and turn its judgments into policy updates.

3.1 Learning an Evidence-Grounded Semantic Criterion

A training example contains a prompt c, candidate image x, evidence explanation d, and consensus score y. Prompts specify entities, their attributes and quantities, spatial and non-spatial relations, and dependencies across sentences. Multiple images per prompt expose alternative realizations of the same requirements. Visual judges annotate matches and violations against these requirements and provide a scalar target emphasizing semantic fidelity while retaining visual quality. Repeated textual requirements are merged at annotation time.

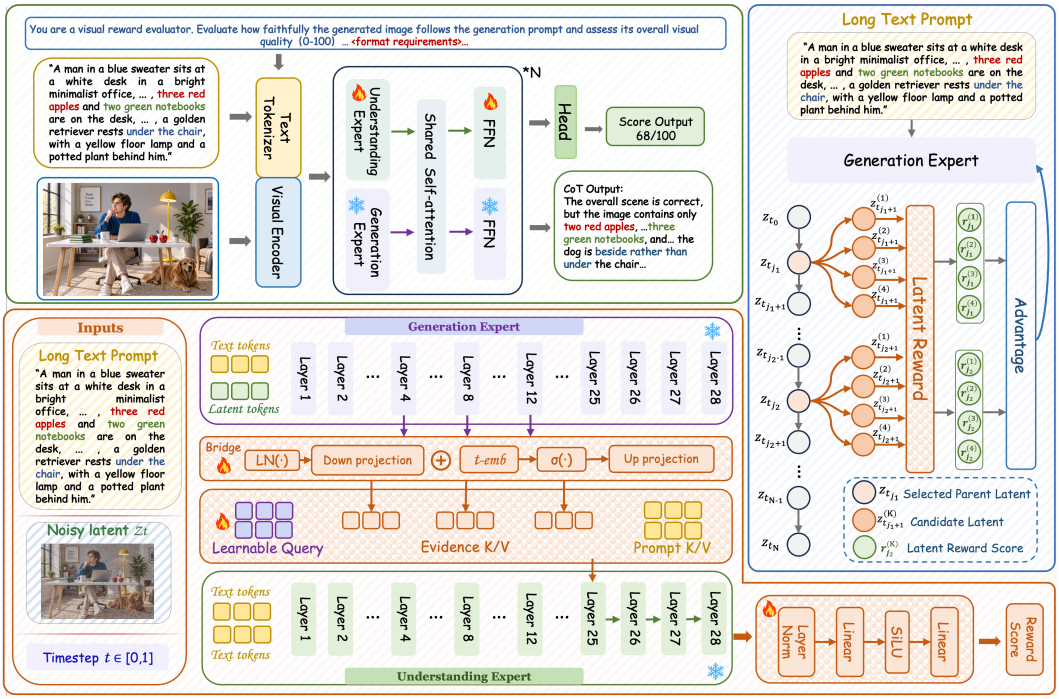

Figure 2: **Overview of Trust-to-Latent.** Image-level evidence and scalar supervision establish an instruction-following criterion in the understanding branch. A trainable interface and reading adapters expose multi-layer generation features to the frozen understanding backbone. During generator optimization, the fixed latent evaluator scores sibling transitions from a common parent state; their relative rewards supervise generation adaptation. Flames mark trainable modules and snowflakes frozen backbone weights; reader QKV adapters are trained with the interface.

The understanding branch is trained to explain visible evidence and to predict a scalar directly. Let $r_\psi(c, x)$ be the output of a small head at a designated score-query position. The scalar task does not condition on a previously generated explanation. We combine the two supervision views as

$$\mathcal{L}_U = \mathbb{E}\Big[\lambda_{\text{ev}}\ell_{\text{CE}}(d \mid c, x) + \lambda_{\text{sc}}\big(r_\psi(c, x) - y\big)^2\Big], \tag{1}$$

Explanation supervision grounds the criterion in image-specific evidence, while regression makes it directly usable as a reward. Prompt-uniform scalar sampling prevents prompts with many candidates from dominating training. We freeze the resulting understanding backbone before learning its latent input interface. The image-level reward also aligns the generator before latent optimization, providing the starting policy and a frozen reference for subsequent updates.

3.2 A Prompt-Conditioned Interface to Latent Evidence

Consider a flow generator (Lipman et al., 2022; Liu et al., 2022) with state z_t, where $t = 1$ denotes noise and $t = 0$ clean content. A frozen generation feature extractor supplies $H_t^\ell = G_{\bar\theta}^\ell(z_t, c, t)$ at selected depths $\ell \in \mathcal{S}$. The interface maps these features into evidence tokens and initializes a bank of queries:

$$\begin{aligned} E_t &= \text{Concat}_{\ell\in\mathcal{S}}\left[B_\phi(\text{LN}(H_t^\ell), e(t)) + b_\ell\right], \\ Q_t^0 &= Q_{\text{learn}} + f_\phi(\bar h_c) + g_\phi(e(t)). \end{aligned} \tag{2}$$

Source embeddings b_ℓ identify feature depths, $e(t)$ encodes noise level, and pooled prompt features $\bar h_c$ condition the queries. Spatial token positions are retained.

Prompt conditioning lets the queries access textual requirements together with spatially indexed candidate features. This is relevant to attribute and relation binding: the same objects can satisfy

one description and violate another through their arrangement. Multiple feature depths supply complementary inputs to the reader, and preference supervision trains the interface to use these inputs for scoring.

The queries read text memory M_c and candidate evidence through understanding blocks with frozen backbone weights and trainable low-rank attention adapters. Interface parameters include the bridge, queries, time conditioning, low-rank attention projections, and scalar head; the pretrained understanding weights remain fixed. Writing their composition as $\mathrm{Reader}_{\psi^\star,\phi}$, the reward is

$$R_\phi(c, z_t, t) = \frac{1}{M} \sum_{m=1}^{M} s_\phi\big(\mathrm{Reader}_{\psi^\star,\phi}(Q_t^0; M_c, E_t)_m\big)\,. \tag{3}$$

Each candidate is scored independently, with queries retaining within-candidate memory. Gradients pass through the understanding blocks to train the interface while the backbone weights remain fixed. Before generator optimization, we also freeze all interface parameters. Together with the frozen feature extractor, this yields a fixed evaluator whose inputs evolve with the policy but whose parameters do not.

3.3 Transferring Image Preferences Across Noise Levels

Within-prompt image preferences provide supervision for this interface. Let z_0^+ and z_0^- encode a preferred and less preferred candidate. We use a common time and Gaussian noise draw:

$$z_t^\pm = (1-t)z_0^\pm + t\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \mathcal{L}_R = \mathbb{E}\ \mathrm{softplus}\big(R_\phi(c, z_t^-, t) - R_\phi(c, z_t^+, t)\big)\,. \tag{4}$$

Shared corruption removes differences in the sampled perturbation: $z_t^+ - z_t^- = (1-t)(z_0^+ - z_0^-)$. Training emphasizes an interior noise range because this difference vanishes as $t \to 1$. Appendix A specifies the preference-training and visual-feature alignment recipe.

The score expresses preference at a given time, not calibrated terminal value. We therefore compare siblings that share the prompt, parent state, and observation time. This local comparison avoids requiring absolute scores to be comparable across different noise levels or scenes. An image-aligned reference policy regularizes the generator updates.

3.4 Online Learning from Local Latent Preferences

We freeze $R_{\phi^\star}$ and collect parent states from a current-policy rollout. From each parent z_t, the stochastic flow sampler draws K independent one-step children at the same next time $s < t$:

$$z_s^{(k)} \sim p_{\theta_{\mathrm{old}}}(\cdot \mid z_t, c), \qquad A_k = \frac{R_{\phi^\star}(c, z_s^{(k)}, s) - \bar{r}}{\mathrm{std}(r_1, \ldots, r_K) + \epsilon}. \tag{5}$$

Here $r_k = R_{\phi^\star}(c, z_s^{(k)}, s)$ and $\bar{r}$ is the sibling mean. Gaussian transitions have model-dependent means and fixed variance. Normalization is local to each parent and time; the parent rollout continues independently of sibling rankings.

For a D-coordinate latent, define the normalized log density $\ell_\theta = D^{-1} \log p_\theta(z_s \mid z_t, c)$ and $\widetilde{\rho}_k = \exp(\ell_\theta - \ell_{\theta_{\mathrm{old}}})$. Using a PPO-style clipped surrogate (Schulman et al., 2017), generator adaptation minimizes

$$\mathcal{L}_G = \mathbb{E}_k\left[-\min(\widetilde{\rho}_k A_k, \mathrm{clip}(\widetilde{\rho}_k, 1-\delta, 1+\delta)A_k) + \frac{\beta}{D} D_{\mathrm{KL}}(p_\theta \| p_{\mathrm{ref}})\right] + \lambda_v \mathcal{L}_{\mathrm{vel}}. \tag{6}$$

Here $\mathcal{L}_{\mathrm{vel}}$ penalizes normalized velocity deviation from a frozen image-aligned reference on fixed probe states. The coordinate-normalized ratio is a local optimization surrogate, not exact joint-density importance sampling. Gradients update generation adapters only; rewards and sampled states are detached.

Adaptation is confined to an early denoising region: the deployed velocity uses the adapted generator for $t > \tau$ and the frozen reference for $t \le \tau$. Collection, replay, and inference share this gate, using the reference model for late denoising. Inference needs no reward evaluation or candidate selection. Appendix A provides the optimization settings.

Table 1: **Performance comparison across five instruction-following benchmarks.** Overall scores and representative category results are reported in percentages (↑). Bold and underlined values indicate the best and second-best results, respectively. Our variants are shaded in light blue.

	DetailMaster			TIIF			LongT2I		UniGen++ (EN-Long)			DPG		
Method / Reward	All	Char.	Spatial	All	Short	Long	All	Rel.	All	Attr.	Comp.	All	Entity	Rel.
Public unified models														
Janus-Pro	59.67	44.42	43.84	69.42	69.18	69.66	79.28	75.03	83.44	88.71	80.47	83.48	90.31	80.18
JanusFlow	52.20	36.78	25.85	57.82	59.55	56.09	74.68	68.50	72.98	79.48	67.06	74.03	82.06	70.54
Show-o	40.67	34.93	26.48	60.00	61.53	58.46	76.27	69.66	78.49	86.50	70.55	81.36	88.76	77.05
Show-o2	52.54	37.85	34.64	63.68	62.57	64.79	79.62	74.21	83.27	89.24	82.80	84.45	91.53	80.64
UniCorn	58.14	43.62	47.91	76.95	77.05	76.86	81.87	78.24	84.92	91.99	83.38	87.58	93.32	83.46
VILA-U	32.42	18.49	8.04	49.82	53.45	46.20	66.44	58.65	57.17	65.52	48.10	64.12	74.44	59.01
BAGEL	57.14	40.61	41.65	72.13	71.39	72.86	79.72	75.18	81.73	89.24	75.22	84.51	91.13	80.79
Reward optimization on BAGEL														
HPSv3	59.51	42.12	47.05	68.66	67.93	69.39	79.91	75.14	73.09	82.61	66.76	84.00	91.42	80.45
PickScore	45.55	36.36	36.18	69.22	68.87	69.57	79.16	74.33	81.20	88.10	76.09	83.62	91.16	80.79
UnifiedReward-2.0	57.10	40.33	41.05	72.11	71.67	72.54	79.52	74.93	81.07	89.47	74.64	84.55	91.24	80.76
VQAScore	57.04	41.00	42.78	72.78	72.30	73.25	80.08	75.88	81.89	90.16	76.68	84.99	91.73	80.64
Trust-to-Latent (UR)	58.78	42.97	49.99	76.75	76.51	76.99	82.66	79.56	85.97	91.69	86.01	87.70	93.58	84.11
Trust-to-Latent (LR)	**59.95**	**45.56**	**55.26**	**79.49**	**79.50**	**79.49**	**83.95**	**81.25**	**88.60**	**94.13**	**88.05**	**89.18**	**94.52**	**85.15**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and reward variants. We use BAGEL (Deng et al., 2025) as the unified backbone for our implementation. **UR** optimizes generation using understanding rewards on completed images; **LR** uses the learned latent evaluator for local feedback. UR and LR start from the same image-aligned generator checkpoint and use the same training budget. Comparisons include public understanding–generation models and the same backbone optimized with HPSv3, PickScore, UnifiedReward-2.0, and VQAScore (Ma et al., 2025b; Kirstain et al., 2023; Wang et al., 2025c; Lin et al., 2024). All reward-optimization comparisons on this backbone use matched training budgets and the same inference settings: 512-pixel images and text guidance 4. Public models retain native generation settings.

Benchmarks and measurement. We evaluate DetailMaster, TIIF, LongT2IBench, UniGenBench++ English Long, and DPG-Bench (Jiao et al., 2025; Wei et al., 2025; Yang et al., 2026; Wang et al., 2025b; Hu et al., 2024), applying the same evaluation protocol to all methods within each benchmark. All scores are evaluated locally and reported as percentages. Generation ablations use a fixed 554-prompt TIIF subset; the full TIIF evaluation contains 5,076 prompts. Table 1 reports overall scores and selected categories; DetailMaster Char. averages three character attributes, and Spatial is its spatial-scene metric. Dataset sizes, scoring rules, and training protocols appear in Appendix A.

4.2 INSTRUCTION FOLLOWING ACROSS FIVE BENCHMARKS

LR achieves the highest overall score among the compared methods on each benchmark in Table 1. Its gains over the base generator are 2.81, 7.37, 4.23, 6.87, and 4.67 points, respectively. Across the six public unified models, Janus-Pro leads on DetailMaster, while UniCorn leads on the other four benchmarks. LR exceeds UniCorn by 2.54 points on TIIF and 3.68 on UniGen++, extending the comparison beyond reward optimization of a shared backbone.

With the backbone held fixed, UR improves all five overall scores, and LR improves further under the same initialization and training budget. This comparison supports the use of latent feedback for instruction following. HPSv3 performs strongly on several DetailMaster scene attributes but falls below the base generator on TIIF and UniGen++, illustrating that strong scene-attribute scores can coexist with weaker instruction following.

Table 2: **Performance comparison on DetailMaster.** BG denotes background.

	Entity and location		Character attributes			Scene attributes			
Method / Reward	Object	Location	Animal	Object	Human	BG	Light	Style	Spatial
Public unified models									
Janus-Pro	42.02	37.56	43.34	**51.70**	**38.22**	95.23	96.62	88.46	43.84
JanusFlow	33.09	24.86	42.80	39.59	27.94	87.98	96.13	91.53	25.85
Show-o	34.29	26.98	36.74	39.23	28.81	53.08	78.63	41.82	26.48
Show-o2	33.68	26.01	38.19	42.20	33.15	91.47	95.74	77.79	34.64
UniCorn	40.97	39.04	46.33	49.66	34.87	96.99	96.98	70.49	47.91
VILA-U	14.67	7.88	22.95	19.57	12.95	48.01	82.11	75.57	8.04
BAGEL	38.90	35.33	42.38	48.13	31.31	94.79	96.20	85.56	41.65
Reward optimization on BAGEL									
HPSv3	39.77	36.97	47.27	47.22	31.86	95.37	95.77	**94.27**	47.05
PickScore	31.36	29.03	43.11	37.14	28.82	93.21	91.88	19.18	36.18
UnifiedReward-2.0	38.24	34.05	43.63	46.48	30.88	94.16	95.72	89.65	41.05
VQAScore	38.69	35.27	45.34	47.20	30.47	94.86	96.42	82.34	42.78
Trust-to-Latent (UR)	40.41	39.07	44.80	48.55	35.57	98.18	97.12	75.30	49.99
Trust-to-Latent (LR)	**42.72**	**41.67**	**47.67**	51.31	37.71	**98.74**	**98.09**	66.42	**55.26**

Table 3: **Performance comparison on TIIF.** Results cover five individual and five compound requirements, aggregating short and long prompts within each category. Col., Cnt., Cmp., Diff., and Tex. denote color, counting, comparison, differentiation, and texture.

	Individual requirements					Compound requirements				
Method / Reward	2D	3D	Count	Compare	Text	Col.+3D	Cnt.+Col.	Cnt.+Tex.	Cmp.+Col.	Diff.+Tex.
Public unified models										
Janus-Pro	87.95	75.78	64.69	62.56	58.19	78.48	75.19	64.05	71.73	59.85
JanusFlow	80.80	67.33	57.59	52.42	34.08	66.67	61.56	52.06	57.71	41.13
Show-o	89.29	74.00	65.86	58.15	19.61	72.42	67.66	59.41	65.89	48.52
Show-o2	84.82	76.89	68.87	63.88	23.52	78.79	74.16	67.27	72.66	64.53
UniCorn	88.62	78.67	78.02	68.94	73.52	**80.61**	79.35	72.81	76.40	67.24
VILA-U	71.21	63.33	41.34	48.02	27.95	66.36	47.27	39.69	47.20	32.27
BAGEL	86.61	78.89	73.74	61.45	66.00	78.79	75.58	65.98	68.69	58.87
Reward optimization on BAGEL										
HPSv3	88.62	77.78	71.98	62.33	55.96	76.97	71.69	65.21	64.02	54.68
PickScore	81.70	75.56	70.62	61.23	61.28	74.55	72.08	67.01	66.12	55.17
UnifiedReward-2.0	85.94	78.44	72.37	60.57	67.19	77.27	74.42	67.40	67.29	60.34
VQAScore	86.38	77.11	74.81	63.00	66.81	78.18	74.94	66.49	67.52	60.10
Trust-to-Latent (UR)	89.51	80.67	75.88	65.42	77.54	79.09	80.00	72.55	72.66	63.05
Trust-to-Latent (LR)	**89.73**	**83.11**	**80.06**	**70.48**	**84.12**	**80.61**	**80.26**	**75.90**	**77.34**	**70.69**

4.3 Fine-Grained Semantic Alignment

Spatial organization shows a larger gain than object presence in DetailMaster. Table 2 reports improvements from 35.33 to 41.67 in location and from 41.65 to 55.26 in spatial relations, compared with 38.90 to 42.72 in object presence. All three character-attribute categories also improve. These results indicate progress in arranging entities and assigning their properties as well as generating the entities themselves. The TIIF results in Table 3 extend this pattern to counting, comparison, and compound requirements; differentiation with texture rises from 58.87 to 70.69.

The gains vary by requirement: LR improves 35 of 39 TIIF categories, with four negation-related regressions, while DetailMaster style declines from 85.56 to 66.42. Table 2 and Appendix E report the category-level results; the overall metrics include all categories.

Joint Constraint Satisfaction. On TIIF, the fraction of images satisfying all assessed requirements rises from 35.54% for the base generator to 42.93% for LR. This 7.39-point increase shows that more images meet their evaluated requirements jointly. Paired analysis further shows that LR corrects

11.53% of all atomic answers while introducing errors on 4.16%, yielding the observed 7.37-point net gain. Gains also persist across all four LongT2I constraint-count groups, at 3.57, 4.31, 4.10, and 4.96 points from the least to the most constrained group. These results connect the overall gains to both joint satisfaction and individual-answer accuracy; Appendix D provides the breakdowns.

4.4 Reward and Feedback Design

Table 4: **Semantic score readout.** PairAcc@5 (%); Δ relative to explanation-first scoring.

Readout	PairAcc@5 ↑	Δ
Explanation then digits	70.03	+0.00
Direct digit generation	76.81	+6.78
Discrete score token	76.68	+6.65
Scalar regression head	**78.02**	+7.99

Table 5: **Latent evidence design on TIIF.** Short and Long denote the two prompt-length groups.

Interface	Short	Long	TIIF
Single-layer evidence	77.45	76.21	76.83
w/o reader QKV adapters	76.02	74.18	75.10
w/o query memory	77.01	74.79	75.90
Independent pair noise	78.36	77.14	77.75
Full interface	**80.08**	**79.81**	**79.94**

Semantic Score Readout. Table 4 compares scoring protocols on 2,664 images from 148 held-out prompts under a common training budget. PairAcc@5 measures within-prompt ordering for reference scores at least five points apart. The scalar head reaches 78.02%, exceeding explanation-first scoring by 7.99 points and direct digit generation by 1.21. The larger separation is between explanation-first and direct readouts; regression gives a further improvement among the direct variants. All direct variants retain explanation supervision as a separate training view. The result therefore favors extracting the score directly under the tested output representations and losses. Appendix B specifies the protocols.

Latent Evidence Interface. The full interface reaches 79.94 on TIIF and leads both prompt-length groups in Table 5. Removing reader QKV adapters or query memory produces the largest overall reductions, 4.84 and 4.04 points, exceeding those from single-layer evidence or independent pair noise. Thus, how the reader processes evidence matters alongside which features it receives. Every ablation has a larger effect on long prompts; removing QKV adaptation costs 5.63 points on long prompts versus 4.06 on short prompts. These results support the adapted reader and its query memory as central components of the interface.

Learning from Latent Feedback. Table 6 compares four update rules under a shared design-study protocol, with a common initialization, frozen latent evaluator, and matched training budgets. One-step and window group-relative updates improve TIIF by 3.42 and 3.80 points over the 67.50 base score. Reward-gradient distillation achieves the highest score, 72.03, a gain of 4.53 points, while pairwise preference optimization decreases performance by 3.04 points. The 7.57-point spread highlights the importance of how feedback is used. Improvements under both group-relative learning and gradient distillation show that the latent reward supports multiple optimization approaches. Appendix B gives the protocols.

Table 6: **Latent-feedback optimization study.** TIIF scores on 554 prompts under a shared study configuration, distinct from final LR training. Budgets are matched; Δ is relative to the base generator.

Optimization	TIIF ↑	Δ
Base generator	67.50	–
One-step group-relative	70.92	+3.42
Window group-relative	71.30	+3.80
Pairwise preference	64.45	-3.04
Reward-gradient distillation	**72.03**	+4.53

4.5 Qualitative Comparison

Figure 3 illustrates distinctions beyond object presence. Our outputs place the yellow sphere between the other objects and closer to the cube, depict the robot holding the rose beside a green plant, and retain the black cube outside the glass sphere with the red pyramid inside. Several alternatives misplace objects, duplicate the rose on the plant, or confuse shapes. These examples complement the quantitative evidence for attribute and relation binding.

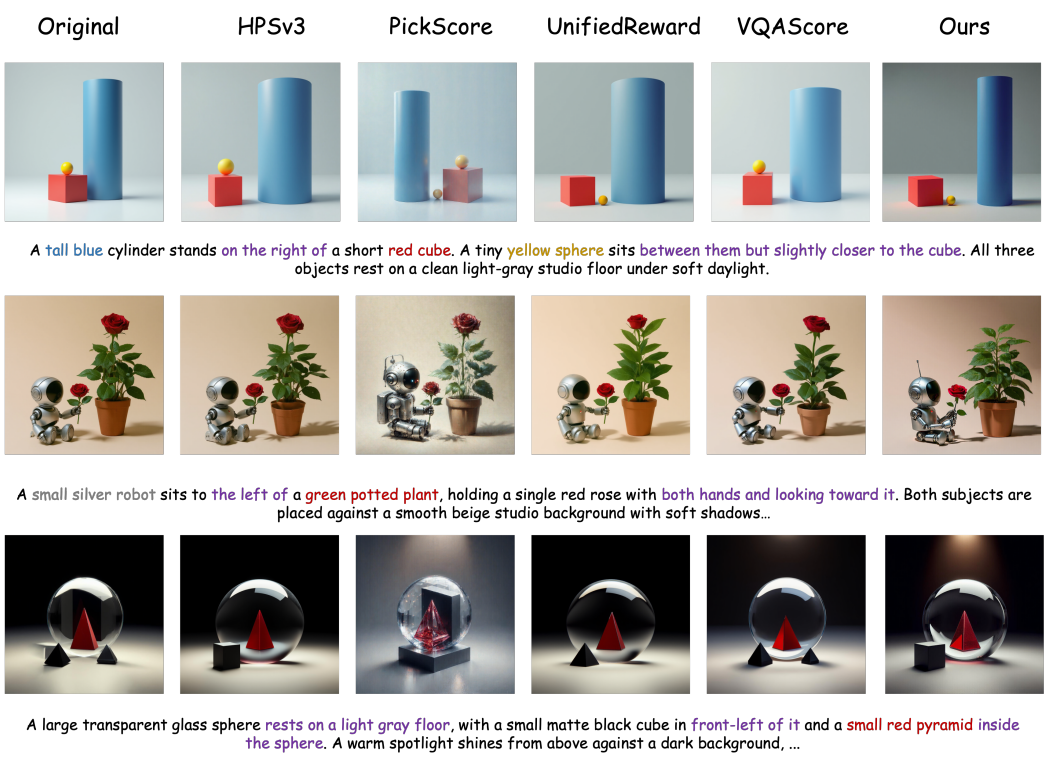

Figure 3: **Compositional instruction following.** Columns compare the base generator, four reward-optimized variants, and Trust-to-Latent. Rows illustrate relative position and proximity, object–action binding, and containment with shape constraints. Colored phrases identify the requested attributes and relations.

4.6 Discussion

Using the base generator as a common reference, UR repairs 9.78% of atomic answers and introduces errors on 5.15%; LR repairs 11.53% and introduces errors on 4.16%. Its improvement therefore combines more corrections with fewer regressions relative to the base generator. Joint satisfaction also rises from 39.07% for UR to 42.93% for LR. The two measurements capture complementary outcomes: improved accuracy across individual requirements and a larger fraction of images meeting all assessed requirements together. Appendix D details these comparisons.

The results link latent evidence modeling to improved compositional instruction following: the full interface outperforms its ablations, several update rules benefit from its feedback, and more generated images satisfy their assessed constraints jointly. The category-level results further distinguish improvements in attribute and relation binding from changes in other requirements.

5 Conclusion

Trust-to-Latent connects the understanding and generation capabilities of a unified model by making reward-trained semantic judgment available to intermediate generation. A learned evidence interface enables understanding to assess evolving visual content, and the resulting fixed evaluator turns these judgments into local feedback for generator learning. Across five benchmarks, this approach improves overall instruction following, including spatial relations, attribute binding, and joint satisfaction of evaluated constraints. The results support using understanding to help generation realize compositional instructions, without reward evaluation or candidate selection at inference.

AI USE STATEMENT

The use of generative AI tools was limited to language editing, including correcting grammatical errors and refining the manuscript's clarity, fluency, and academic style. The authors independently developed the research methodology, conducted the experiments, and performed the scientific analysis.

ETHICS STATEMENT

This work studies semantic instruction following in image generation through evaluation on public benchmarks. We support responsible research practices and transparent reporting of methods and evaluation results.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document data construction, training configurations, evaluation protocols, and design studies in Appendices A and B of the separate supplementary material. Section 3 defines the learning objectives, and Appendix C describes the motivation-study protocol. These materials support independent assessment and reproduction of the reported results.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Zhen Fang, Ruiyan Han, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, et al. Unicorn: Towards self-improving unified multimodal models through self-generated supervision. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13638–13669, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Hyojun Go, Hyungjin Chung, Prune Truong, Goutam Bhat, Li Mi, Zhaochong An, Zixiang Zhao, Dominik Narnhofer, Serge Belongie, Federico Tombari, et al. Stitched value model for diffusion alignment. *arXiv preprint arXiv:2605.19804*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Qirui Jiao, Daoyuan Chen, Yilun Huang, Xika Lin, Ying Shen, and Yaliang Li. Detailmaster: Can your text-to-image model handle long prompts? *arXiv preprint arXiv:2505.16915*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13199–13208. IEEE, 2025.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. *URL https://arxiv. org/abs/2505.05470*, 2:5, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025a.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anythi. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24591–24602. IEEE, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025b.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, Bairui Li, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025a.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Zhichao Yang, Tianjiao Gu, Jianjie Wang, Feiyu Lin, Xiangfei Sheng, Pengfei Chen, and Leida Li. Longt2ibench: A benchmark for evaluating long text-to-image generation with graph-structured annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11820–11828, 2026.

Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, Chunhong Pan, et al. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *Advances in Neural Information Processing Systems*, 38:106202–106235, 2026.