CODA: CONTRASTIVE DECISION ADVANTAGES FOR UNIFIED TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified models integrate language understanding and image synthesis within a single model, and reinforcement learning further improves their generation capability through outcome-level rewards on completed images. However, relying on a single final-image reward to jointly optimize semantic reasoning and visual generation leaves two difficulties: ***ambiguous semantic attribution*** and ***sparse visual supervision***. The former mixes high-level semantic reasoning with its low-level visual realization, while the latter leaves a long chain of visual decisions without feedback before the image is completed. To address them, we propose **COntrastive Decision Advantag (CODA)**, a framework that derives ***contrastive decision advantages*** by comparing each decision with a matched alternative using shared randomness, which contains Semantic Reasoning Guidance and Masked Trajectory Guidance. Specifically, Semantic Reasoning Guidance isolates the effect of prompt optimization by comparing images generated from the optimized and original prompts. Masked Trajectory Guidance contrasts each probed commitment with equally sized alternatives and weights its outcome gain by the induced distribution shift, promoting earlier commitments that are both correct and structurally influential. A trajectory-level outcome advantage is further incorporated to provide a global signal. Grounded in the same constraint-anchored evaluation from the original instruction, the semantic advantage supervises the text tokens forming the optimized prompt, while the process advantage supervises each probed commitment. Extensive experiments on GenEval, T2I-CompBench, and WISE demonstrate that CODA achieves outstanding performance across tasks involving complex scene composition and knowledge-intensive image generation.

1 INTRODUCTION

Text-to-image generation has advanced rapidly (Podell et al., 2023; Black Forest Labs, 2024), and a prominent trend integrates visual understanding and image synthesis into a single unified model (Chen et al., 2025; Guo et al., 2026; Ye et al., 2026). By coupling the two abilities, such models can leverage their linguistic knowledge to interpret compositional and knowledge-intensive instructions before visual generation begins. Building on this capability, reinforcement learning (RL) has further strengthened these models, yielding substantial gains in visual quality and instruction following (Jiang et al., 2025; Guo et al., 2025; Wu et al., 2025; Wang et al., 2025).

Existing methods typically sample a full trajectory of semantic reasoning followed by visual generation, score the completed image, and broadcast this single outcome reward across the whole trajectory to optimize both the reasoning and the visual synthesis (Jiang et al., 2025; Zhang et al., 2025; Wu et al., 2025; Luo et al., 2025). This design, however, leaves two difficulties: *ambiguous semantic attribution* and *sparse visual feedback*.

The former arises because semantic reasoning and visual generation operate at different granularities yet share one terminal reward. Semantic reasoning makes a high-level, global plan of the intended scene, whereas visual generation realizes it through a long series of local decisions. Scoring both by the same completed image entangles their contributions, since a better description does not guarantee a faithful image, nor does a high-quality image confirm that the description helped. An accurate semantic decision may thus be penalized when the subsequent visual sampling fails, while an in-

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.

accurate one may be rewarded when visual sampling happens to yield a successful image, thereby making estimates of the causal contribution of semantic reasoning unreliable.

The latter stems from the lack of feedback before image completion, leaving a single terminal reward to supervise a long sequence of dependent visual decisions (Shao et al., 2024). Assigning the same trajectory-level advantage to every step fails to distinguish beneficial commitments from harmful ones within the same trajectory (Ouyang et al., 2026). Beyond this temporal sparsity, visual commitments also differ substantially in influence. Some establish the global structure, whereas others merely refine local details (Tong et al., 2026). Uniform supervision ignores this variation, overlooks pivotal commitments, and slows optimization.

These difficulties raise a natural question: ***how can we design accurate, fine-grained advantages that enable decision-level supervision for both semantic reasoning and the masked generation process?***

To address these challenges, we propose **COntrastive Decision Advantage (CODA)**, which supervises each decision by contrasting its outcome with that of a matched alternative under shared randomness, so that their score difference isolates the decision's own contribution. Specifically, CODA comprises Semantic Reasoning Guidance and Masked Trajectory Guidance. Through decision-level comparisons, the former enables accurate semantic attribution and the latter provides dense visual feedback, jointly improving image quality and instruction following.

Semantic Reasoning Guidance treats the optimized prompt as the semantic decision and the original prompt as its control, estimating from paired generations whether the optimization better satisfies the original requirements. Because both branches are scored under matched conditions and differ only in the prompt, the resulting advantage reflects the reasoning's own effect and supervises the tokens of the optimized prompt, giving accurate semantic attribution.

Masked Trajectory Guidance supervises local visual decisions on the masked decoder. At selected steps, it contrasts the actual commitment with equally sized alternatives to measure whether the commitment improves the final image. Moreover, we quantify the distribution shift induced by each commitment using KL divergence and use it to weight the resulting commitment-level contrastive reward, encouraging the model to make correct and structurally influential commitments earlier. Together with a trajectory-level outcome advantage as global guidance, this yields a fine-grained process reward that supplies dense feedback throughout rather than a single terminal signal.

Our contributions are as follows:

- We formulate accurate and fine-grained supervision for unified text-to-image generation via a contrastive decision mechanism, which alleviates ambiguous semantic attribution and sparse visual feedback problems to some extent.
- We design Semantic Reasoning Guidance and Masked Trajectory Guidance modules, which respectively score the semantic plan against its original-prompt control and turn matched commitment comparisons into a fine-grained visual process advantage, ultimately improving final image quality and instruction following.
- Extensive experiments on GenEval, T2I-CompBench, and WISE show that CODA outperforms existing methods, especially on compositional and knowledge-intensive prompts that require reasoning, achieving 84.25% on GenEval and improving baseline overall WISE score from 0.35% to 0.50%.

2 METHOD

2.1 PRELIMINARIES

Masked Visual Generation Masked generative models represent an image as a sequence of discrete visual tokens and predict masked positions in parallel through bidirectional attention (Chang et al., 2022; Xie et al., 2024). Let $\mathbf{T} = (T_1, \ldots, T_S)$ be the text condition and $\mathbf{Y} = (Y_1, \ldots, Y_N)$ the visual token sequence. For a mask set $\mathcal{M} \subseteq \{1, \ldots, N\}$ and its complement $\overline{\mathcal{M}}$, the optimization

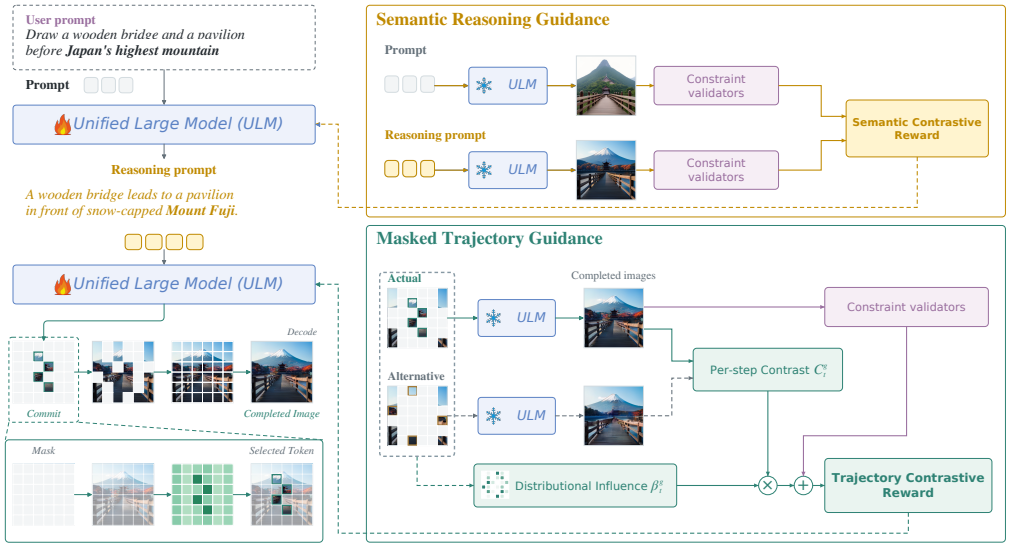


Figure 1: An overview of the CODA framework, which comprises two complementary components: Semantic Reasoning Guidance and Masked Trajectory Guidance. The former compares paired generations from the optimized and original prompts to enable accurate semantic attribution, while the latter scores the visual token commitments to provide fine-grained supervision throughout generation. Through a contrastive decision mechanism, the two components jointly improve image quality and instruction following.

objective is:

$$\mathcal{L}_{\text{MTP}} = -\mathbb{E}_{\mathcal{M}}\left[\sum_{i\in\mathcal{M}} \log p_\theta(Y_i \mid \mathbf{T}, \mathbf{Y}_{\overline{\mathcal{M}}})\right]. \tag{1}$$

Inference starts from a fully masked visual sequence and alternates parallel prediction with selective commitment. Given step t, the state $s_t = (\mathbf{T}, \mathbf{Y}^{(t)}, \mathcal{M}_t)$ consists of the text condition, the partial visual sequence $\mathbf{Y}^{(t)}$, and the set of remaining masked positions $\mathcal{M}_t$. At each $i \in \mathcal{M}_t$, the model samples a candidate token $\hat{Y}_{t,i}$ and computes its confidence:

$$q_{t,i} = \log p_\theta(\hat{Y}_{t,i} \mid s_t) + \tau_t g_{t,i}, \qquad \hat{Y}_{t,i} \sim p_\theta(\cdot \mid s_t). \tag{2}$$

Here $g_{t,i} \sim \text{Gumbel}(0,1)$ and τ_t decreases during generation. The sampler selects a commitment set $\mathcal{U}_t \subseteq \mathcal{M}_t$ according to these confidences, and the corresponding candidates $\hat{Y}_{t,i}$ are committed into $\mathbf{Y}^{(t)}$ to form the next state; the other positions remain masked. Thus, the newly committed tokens advance the visual state, rather than all candidates predicted in parallel.

Reinforcement Learning for Discrete Understanding and Generation Given an input x, a unified discrete model generates a trajectory $\mathbf{a} = (a_1, \ldots, a_L)$, where each action a_ℓ is either a text or a visual token. Let h_ℓ denote the context preceding a_ℓ. The trajectory log-probability factorizes autoregressively as:

$$\log p_\theta(\mathbf{a} \mid x) = \sum_{\ell=1}^{L} \log p_\theta(a_\ell \mid h_\ell). \tag{3}$$

An external evaluator assigns a scalar reward $\mathcal{R}(\mathbf{a}, x)$ to the final output. Conventional outcome-based optimization samples a group of trajectories for the same input and normalizes their rewards into a group-relative advantage shared by all actions in each trajectory (Shao et al., 2024). Such a uniform assignment, however, disregards the unequal contributions of individual actions. For a group of G trajectories generated from the same input x, let $r_i = \mathcal{R}(\mathbf{a}_i, x)$. The clipped objective and standard group-relative advantage are:

$$\mathcal{L}(A, \rho_\theta) = -\mathbb{E}\left[\min\left(\rho_\theta A, \text{clip}(\rho_\theta, 1-\epsilon, 1+\epsilon)A\right)\right], \qquad A_i = \text{norm}_i(r_i) = \frac{r_i - \mu}{\sigma + \delta}. \tag{4}$$

Here $\rho_\theta = \frac{p_\theta(a|s)}{p_{\theta_{\text{old}}}(a|s)}$ for an action a at state s, where $p_{\theta_{\text{old}}}$ is the sampling policy and ϵ is the clipping range; the expectation averages the supervised decisions in sampled trajectories. Throughout, norm_i denotes group z-score normalization over the indicated sample index i, holding other indices fixed; μ and σ are the group mean and standard deviation, and $\delta > 0$ ensures numerical stability.

2.2 Method Overview

2.2.1 Contrastive Decision Advantage

A unified model p_θ couples multimodal understanding and generation in two stages: given a prompt P, it first performs semantic reasoning to produce a condition P' better suited to synthesis and then generates a visual trajectory $\tau = (a_0, \ldots, a_{T-1})$ and an image I conditioned on P'. Since the final score cannot separate the contributions of individual decisions, CODA estimates the effect of each decision by comparing it against a matched alternative under shared randomness, and optimizes both stages accordingly. At state s, a decision d and its alternative $\bar{d}$ produce I_d^ξ and $I_{\bar{d}}^\xi$ under shared randomness ξ. Their contrastive decision advantage is

$$A_{\text{cons}}(s, d) = \mathbb{E}_\xi \left[\mathcal{O}(I_d^\xi) - \mathcal{O}(I_{\bar{d}}^\xi) \right], \tag{5}$$

where $\mathcal{O}$ is the unified outcome evaluation. As the branches differ only in the examined decision, their gap in $\mathcal{O}$ isolates its causal effect. We instantiate this principle as two forms of guidance driven by the same evaluation. Semantic reasoning guidance contrasts images from the original and optimized prompts, whereas masked trajectory guidance contrasts actual visual commitments against alternatives. Note that Eq. 5 states the general form of the proposed method, and its concrete realizations are detailed in Sec. 2.3 and 2.4.

2.2.2 Unified Outcome Evaluation

To compare decisions consistently, all advantage estimates share a single constraint-anchored evaluation $\mathcal{O}$. From the original prompt P, we extract a verifiable constraint set $\mathcal{C}(P) = \{c_1, \ldots, c_M\}$ covering the requested objects, attributes, counts, and spatial relations, which stays fixed throughout training. Each constraint has a type $r \in \mathcal{T} = \{s, n, c, e\}$ (spatial relation, number, color, and object existence). The evaluation of a generated image I is then:

$$\mathcal{O}(I) = \sum_{r \in \mathcal{T}(P)} \mathcal{R}_r(I) + \lambda_h \mathcal{R}_h(I, P), \tag{6}$$

where $\mathcal{R}_r$ is the weighted contribution of type r, and $\mathcal{R}_h$ is a holistic preference term for overall visual quality, balanced by λ_h. All validators build on an open-vocabulary detector $\mathcal{D}$ for object localization (Liu et al., 2023) and a holistic preference model for $\mathcal{R}_h$ (Wu et al., 2023). Each type contribution weights the average per-instance match over its constrained objects:

$$\mathcal{R}_r(I) = \frac{w_r}{N_r} \sum_{i=1}^{N_r} m_r(\mathcal{D}(I),\, y_{r,i}), \qquad r \in \{s, n, c, e\}, \tag{7}$$

where $y_{r,i}$ is the target of the i-th constraint of type r, and $m_r \in [0, 1]$ is a type-specific match. For existence, spatial relations, and color, m_r is an exact-match indicator on the detections, namely object presence, the box-based relation of an object pair, and the region color predicted in a joint vision-language embedding space, respectively. For counting, $m_n = \exp(-|\hat{n}_i - y_{n,i}|/\tau_n)$ softly penalizes the deviation between the detected count $\hat{n}_i$ and its target, with tolerance τ_n. Note that a validator is active only when its type appears in $\mathcal{T}(P)$, and the active scores combine into $\mathcal{O}$. Detailed Evaluator settings are given in Appendix A.

2.3 Semantic Reasoning Guidance

Semantic reasoning guidance measures how the optimized prompt affects the final generation. Given the original prompt P, the model produces an optimized prompt P' suited to visual synthesis; we treat P' as the semantic decision and P as its matched control. To isolate the effect of the reasoning process from visual sampling noise, we generate K image pairs under shared seeds:

$$I_k = \mathcal{M}_{\text{t2i}}(P; \xi_k), \qquad I'_k = \mathcal{M}_{\text{t2i}}(P'; \xi_k), \tag{8}$$

where K is the number of shared seeds, ξ_k is the k-th seed, and I_k, I'_k are generated from P and P'. The semantic advantage averages the evaluation difference over these pairs:

$$A_{\text{sem}} = \frac{1}{K} \sum_{k=1}^{K} \left[\mathcal{O}(I'_k) - \mathcal{O}(I_k) \right]. \tag{9}$$

Each image is scored against constraints extracted from the original prompt, so the optimized prompt cannot raise its advantage by dropping or altering a requirement.

2.4 Masked Trajectory Guidance

Masked generation commits visual tokens step by step, yet a single trajectory-level reward cannot distinguish the contribution of individual commitments. Masked trajectory guidance therefore scores each commitment by contrasting it against matched alternatives under shared randomness, isolating its causal contribution to the final image. For the same optimized prompt P', we sample G original trajectories. For trajectory $g \in \{1, \dots, G\}$, the rollout at step t takes the joint visual action $a_t^g = (\hat{Y}_{t,i}^g)_{i \in \mathcal{U}_t^g}$ by committing the candidate tokens at $\mathcal{U}_t^g \subseteq \mathcal{M}_t^g$, and reaches the successor state s_{t+1}^g. To form a contrastive decision, we hold out $\mathcal{U}_t^g$ and let the sampler commit an equal number of positions from the remaining candidates; repeating this J times yields alternative commitments $\bar{\mathcal{U}}_t^{g,j}$ and states $\bar{s}_{t+1}^{g,j}$. Starting from their respective successor states, the actual and j-th alternative branches generate the final images I^g and $\bar{I}_t^{g,j}$ with the same subsequent randomness ξ:

$$I^g \sim p_{\theta_{\text{old}}}(\cdot \mid s_{t+1}^g; \xi), \qquad \bar{I}_t^{g,j} \sim p_{\theta_{\text{old}}}(\cdot \mid \bar{s}_{t+1}^{g,j}; \xi), \quad j = 1, \dots, J. \tag{10}$$

As the two branches commit equally number of tokens and differ only in the step-t selection, their outcome gain compares the actual final image with the average over its alternatives:

$$C_t^g = \mathcal{O}(I^g) - \frac{1}{J} \sum_{j=1}^{J} \mathcal{O}(\bar{I}_t^{g,j}). \tag{11}$$

Its sign indicates whether this commitment helps or harms the final image, with the shared randomness ξ isolating the effect from sampling noise. The alternative branches only estimate C_t^g and are excluded from the outcome group.

Despite indicating the optimization direction, outcome gain does not capture how strongly the commitment influences subsequent generation. To quantify this, we compare the predictions over the positions that remain masked before and after the actual commitment, i.e., over $\mathcal{Q}_t^g = \mathcal{M}_t^g \setminus \mathcal{U}_t^g$. We measure the influence of a commitment by the KL divergence between the token predictive distributions before and after it over these positions; a commitment that imposes structural changes on the image typically shifts these predictions more:

$$\beta_t^g = \frac{1}{|\mathcal{Q}_t^g|} \sum_{i \in \mathcal{Q}_t^g} D_{\text{KL}}\big(p_{\theta_{\text{old}},i}(\cdot \mid s_{t+1}^g) \,\big\|\, p_{\theta_{\text{old}},i}(\cdot \mid s_t^g)\big), \tag{12}$$

where $p_{\theta,i}(\cdot \mid s)$ is the visual token distribution at position i in state s, and s_t^g, s_{t+1}^g are the states before and after committing $\mathcal{U}_t^g$. Furthermore, by combining β_t^g and C_t^g, we encourage commitments that are both structurally influential and correct to be made earlier. The group-standardized process advantage is

$$A_t^{\text{proc},g} = \text{norm}_g(\beta_t^g C_t^g), \tag{13}$$

Here, both C_t^g and β_t^g are computed with stopped gradients under the frozen rollout policy.

2.5 Joint Policy Optimization

We apply the common objective in Eq. 4 to jointly optimize semantic reasoning and masked trajectory generation. For semantic reasoning, the advantage A_{sem} in Eq. 9 supervises the text-token actions that form P'. For masked trajectory generation, in addition to the process advantage $A_t^{\text{proc},g}$ in Eq. 13, we optimize the final outcome using A_{out}^g obtained from Eq. 4 with reward $\mathcal{O}(I^g)$. This trajectory-level advantage is shared by all steps of trajectory g.

Table 1: Comparison to the state-of-the-art diffusion models and autoregressive models on GenEval.

Method	Params.	Single obj.	Two obj.	Counting	Colors	Position	Color attr.	Overall
Diffusion Models								
SD-XL-base-1.0 Podell et al. (2023)	2.6B	98.34	74.14	39.21	85.36	15.09	23.11	55.87
PixArt-α Chen et al. (2023)	0.6B	98.05	50.31	44.67	80.12	8.55	7.21	48.15
SD-3.5-L Stability AI (2024)	8B	98.34	89.76	74.04	82.69	35.46	46.01	71.05
FLUX.1-schnell Black Forest Labs (2024)	12B	–	–	–	–	–	–	71.38
Autoregressive Models								
SimpleAR Wang et al. (2025)	1.5B	–	90.00	–	–	28.00	45.00	63.00
MaskGRPO Luo et al. (2025)	1.3B	99.00	90.00	69.00	85.00	35.00	59.00	73.00
PARM Guo et al. (2025)	7B	*99.21*	85.64	67.15	84.02	65.83	64.04	77.64
Janus-Pro Chen et al. (2025)	7B	99.03	88.98	58.75	**90.02**	*78.89*	*66.12*	*80.29*
MetaQuery-XL Pan et al. (2025)	7B	–	–	–	–	–	–	80.00
Show-o Xie et al. (2024)	1.3B	98.11	80.04	66.25	84.17	31.03	50.14	68.29
CODA (ours)	1.3B	**99.71**	*91.25*	**82.04**	85.21	**80.44**	**66.82**	**84.25**

Following Mask-GRPO (Luo et al., 2025), both visual terms optimize actual commitments using the committed-token likelihood ratio. The text and visual ratios are

$$\rho_\ell^{\mathrm{sem}}(\theta) = \frac{p_\theta(a_\ell \mid h_\ell)}{p_{\theta_{\mathrm{old}}}(a_\ell \mid h_\ell)}, \qquad \rho_t^g(\theta) = \frac{p_\theta(a_t^g \mid s_t^g)}{p_{\theta_{\mathrm{old}}}(a_t^g \mid s_t^g)} = \prod_{i \in \mathcal{U}_t^g} \frac{p_\theta(\hat{Y}_{t,i}^g \mid s_t^g)}{p_{\theta_{\mathrm{old}}}(\hat{Y}_{t,i}^g \mid s_t^g)}. \tag{14}$$

Both ratios are evaluated by the same unified model p_θ. The semantic ratio covers one autoregressive text-token action, whereas the visual ratio is the joint likelihood ratio of the tokens committed at one masked-generation step. Thus, outcome and process supervision share the same ratio but enter separate clipped losses:

$$\mathcal{L}_{\mathrm{total}} = \underbrace{\mathbb{E}_\ell[\mathcal{L}(A_{\mathrm{sem}}, \rho_\ell^{\mathrm{sem}})]}_{\mathcal{L}_{\mathrm{sem}}} + \underbrace{\lambda_v \, \mathbb{E}_{g,t}[\mathcal{L}(A_{\mathrm{out}}^g, \rho_t^g)]}_{\mathcal{L}_{\mathrm{out}}} + \underbrace{\lambda_p \, \mathbb{E}_{g,\, t \in \mathcal{P}_g}[\mathcal{L}(A_t^{\mathrm{proc},g}, \rho_t^g)]}_{\mathcal{L}_{\mathrm{proc}}} + \lambda_{\mathrm{KL}} \mathcal{K}. \tag{15}$$

Here $\mathcal{P}_g$ denotes the set of probed steps within trajectory g, and λ_v, λ_p, and λ_{KL} balance their corresponding terms. And $\mathcal{K} = D_{\mathrm{KL}}(p_\theta \,\|\, p_{\mathrm{ref}})$ is the standard KL-divergence penalty. Accordingly, CODA jointly optimizes the tokens of P' and the committed visual tokens with shared parameters and their respective advantages, while preserving the original masked generation procedure.

3 Experiments

3.1 Experimental Setup

Training Settings. Our training dataset consists of text prompts generated using Qwen3.5 (Yang et al., 2025) based on several templates covering the 80 COCO object classes (Lin et al., 2014) (see Appendix A for details). These prompts are further processed by Qwen3.5 to extract objects and their associated attributes, such as color, position, and counting, for reward computation. A total of 15K prompts are generated in this manner for training. We use Show-o (Xie et al., 2024) as our base model and train it with a learning rate of 10^{-6}. During training, each prompt uses $K = 2$ shared-seed image pairs and $G = 4$ visual trajectories, with $J = 4$ alternative commitments drawn at each probed step. The probed step set $\mathcal{P}_g$ consists of two steps uniformly sampled from the full $T = 50$ decoding steps. For outcome evaluation, we adopt HPS (Wu et al., 2023) as the human preference model and GroundingDINO (Liu et al., 2023) as the object detector. The proposed model is implemented using PyTorch and trained on NVIDIA H20 GPUs.

Benchmarks. We evaluate our method on GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023), and WISE (Niu et al., 2025). GenEval and T2I-CompBench are widely used to assess attribute-level text-image alignment, covering aspects such as object presence, counting, colors, spatial relations, and finer attribute binding. WISE evaluates world-knowledge understanding in text-to-image generation, covering cultural common sense, spatio-temporal reasoning, and natural science, thereby probing whether models can reason about knowledge-intensive prompts. We follow the recommended evaluation protocol for each benchmark.

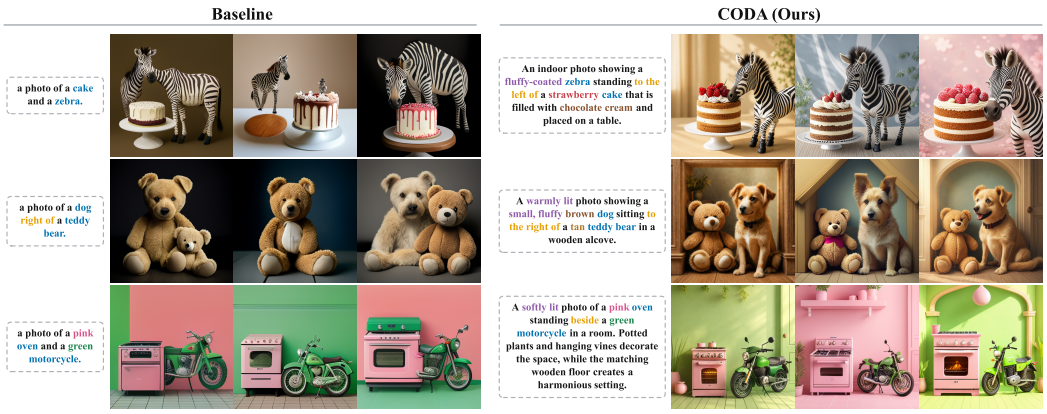


Figure 2: Qualitative comparisons on GenEval. Original prompts and baseline generations are shown on the left, with CODA-optimized prompts and corresponding generations on the right.

3.2 Main Results

Tables 1 and 2 present a comprehensive comparison between the proposed method and other state-of-the-art models across the visual generation benchmarks GenEval and T2I-CompBench, covering both diffusion-based methods such as SD-3.5-L (Stability AI, 2024), FLUX.1-schnell (Black Forest Labs, 2024) and autoregressive-based models such as MetaQuery (Pan et al., 2025), Janus-Pro (Chen et al., 2025). Notably, our method achieves substantial improvements over the baseline model Show-o (Xie et al., 2024), with an average enhancement of 15.96% across all metrics on GenEval. It consistently outperforms both diffusion-based and autoregressive-based approaches.

Table 2: Comparison to the state-of-the-art diffusion models and autoregressive models on T2I-CompBench.

Method	Color	Shape	Texture	Spatial
Diffusion Models				
SD-XL-base-1.0 Podell et al. (2023)	58.79	46.87	52.99	21.31
FLUX.1-schnell Black Forest Labs (2024)	66.90	49.27	64.77	20.64
Autoregressive Models				
PARM Guo et al. (2025)	75.21	56.32	66.04	29.12
Janus-Pro Chen et al. (2025)	63.59	35.28	49.36	20.61
T2I-R1 Jiang et al. (2025)	*81.30*	*58.52*	**72.43**	33.78
Show-o (baseline) Xie et al. (2024)	68.45	49.61	65.34	*38.12*
CODA (ours)	**83.97**	**62.54**	*69.71*	**45.22**

Similarly, on T2I-CompBench, our method achieves strong performance across various tasks, obtaining the highest scores of 83.97% in Color and 45.22% in Spatial. These improvements are partly attributed to the semantic reasoning stage, which enables global-level planning and reasoning, while the subsequent trajectory guidance ensures faithful execution of the initial design intent with fine-grained refinement, thereby yielding accurate and coherent final results. Moreover, by leveraging its rich world knowledge for upfront planning, the proposed model surpasses the diffusion-based FLUX.1-schnell by 17.07% and 24.58% in Color and Spatial, respectively, despite its relatively small size (i.e., 1.3B).

The qualitative results shown in Figure 2 further confirm CODA's improved instruction following. Compared with the baseline, CODA benefits from two complementary components. Semantic Reasoning Guidance produces richer and more contextually appropriate descriptions, while Masked Trajectory Guidance steers visual generation toward higher-quality images that better follow the original instructions.

To further assess the reasoning capability required for complex semantics and world knowledge, we evaluate the proposed model on WISE. As reported in Table 3, it improves the overall score by 15% over the baseline and ranks second only to the considerably larger BAGEL (Deng et al., 2025) (i.e., 14B-A7B). The improvements are consistent across all dimensions; for instance, our model surpasses the baseline by 17% and 16% on Time and Space, respectively. These results indicate that semantic reasoning guidance enables the model to achieve high-level understanding and reasoning for complex real-world scenarios.

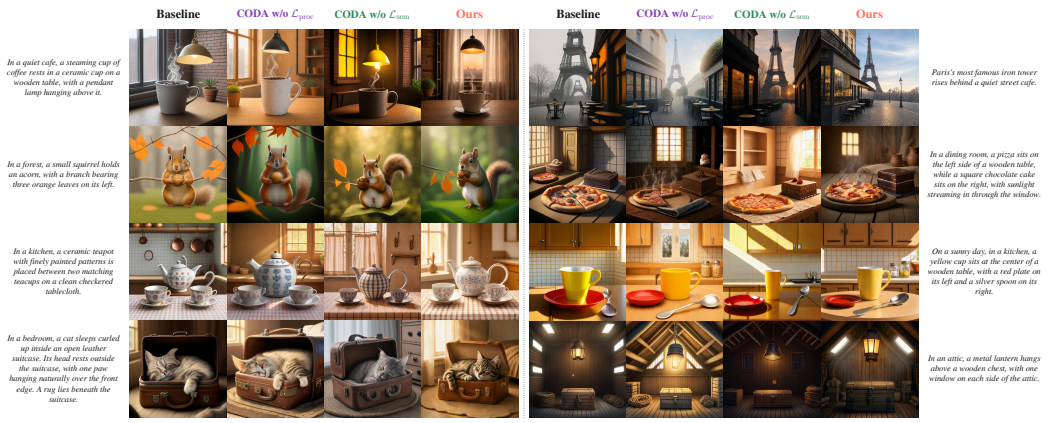


Figure 3: Image generation results for the same prompts under four settings: Baseline, CODA w/o $\mathcal{L}_{\text{proc}}$, CODA w/o $\mathcal{L}_{\text{sem}}$, and full CODA.

Table 3: Comparison to the state-of-the-art diffusion models and autoregressive models on WISE.

Method	Cultural↑	Spatio-Temporal		Natural Science			Overall↑
		Time↑	Space↑	Biology↑	Physics↑	Chemistry↑	
Diffusion Models							
Playground-v2.5 Li et al. (2024)	**49**	**58**	55	43	48	33	49
SD-XL-base-1.0 Podell et al. (2023)	43	48	47	*44*	45	27	43
FLUX.1-dev Black Forest Labs (2024)	*48*	**58**	62	42	51	35	*50*
Autoregressive Models							
Emu3 Wang et al. (2024)	34	45	48	41	45	27	39
Manzano-3B Zhou et al. (2025)	42	51	59	**45**	51	32	46
Bagel-14B-A7B Deng et al. (2025)	44	55	**68**	*44*	**60**	**39**	**52**
Janus-Pro Chen et al. (2025)	30	37	49	36	42	26	35
Show-o (baseline)Xie et al. (2024)	28	40	48	30	46	30	35
CODA (ours)	40	*57*	*64*	*44*	*57*	*36*	*50*

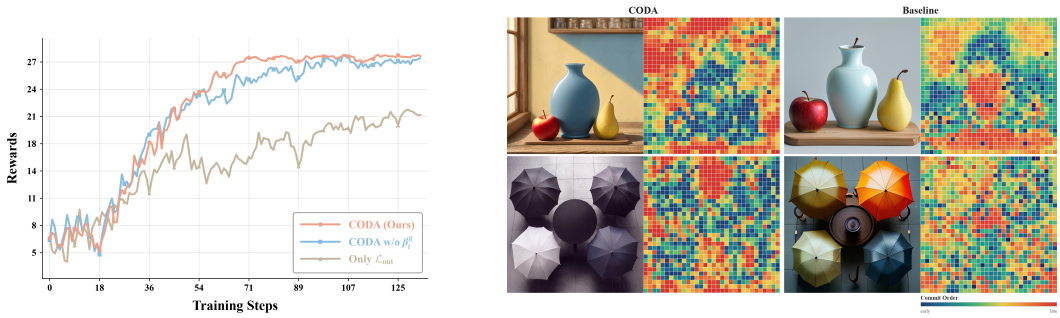
3.3 Ablation Studies

To isolate the contributions of CODA's outcome, semantic, and process signals, we conduct ablation studies on GenEval and WISE. As shown in Table 4, introducing the global outcome loss $\mathcal{L}_{\text{out}}$ over the baseline improves GenEval by 3.86% and WISE by 3%. Building on $\mathcal{L}_{\text{out}}$, adding $\mathcal{L}_{\text{sem}}$ yields further gains of 4.73% and 5%, confirming the benefit of high-level semantic planning, whereas adding $\mathcal{L}_{\text{proc}}$ instead raises GenEval by 7.44% and WISE by 3%, showing that commitment-level feedback is particularly effective for compositional alignment. Removing the influence weight β_t^g from the full objective drops GenEval from 84.25% to 80.25% and WISE from 50% to 47%, verifying that emphasizing structurally consequential commitments is critical. Combining $\mathcal{L}_{\text{sem}}$, $\mathcal{L}_{\text{out}}$, and $\mathcal{L}_{\text{proc}}$ achieves 84.25% on GenEval and 50% on WISE. Notably, the larger gain of $\mathcal{L}_{\text{sem}}$ over $\mathcal{L}_{\text{proc}}$ on WISE is consistent with WISE's emphasis on high-level semantics and world knowledge.

Table 4: Ablation of CODA's core components on GenEval and WISE.

Method	GenEval	WISE
Baseline	68.29	35
+ $\mathcal{L}_{\text{out}}$	72.15	38
+ $\mathcal{L}_{\text{out}}$ +$\mathcal{L}_{\text{sem}}$	76.88	43
+ $\mathcal{L}_{\text{out}}$ +$\mathcal{L}_{\text{proc}}$	79.59	41
+ $\mathcal{L}_{\text{out}}$ +$\mathcal{L}_{\text{sem}}$ +$\mathcal{L}_{\text{proc}}$ w/o β_t^g	*80.25*	*47*
+ $\mathcal{L}_{\text{out}}$ +$\mathcal{L}_{\text{sem}}$ +$\mathcal{L}_{\text{proc}}$	**84.25**	**50**

Figure 3 visualizes these effects under four settings: Baseline, CODA w/o $\mathcal{L}_{\text{proc}}$, CODA w/o $\mathcal{L}_{\text{sem}}$, and Ours. The baseline frequently violates global structure, for instance placing the red plate beneath the yellow cup rather than to its left. Removing $\mathcal{L}_{\text{proc}}$ weakens dense supervision over visual commitments, resulting in less coherent local structures and less reliable attribute binding. In contrast, removing $\mathcal{L}_{\text{sem}}$ reduces the effectiveness of semantic reasoning guidance, making spatial relations, object counts, and knowledge-driven references harder to preserve. With both losses, CODA pro-



(a) (b)

Figure 4: (a) Reward curves for Only $\mathcal{L}_{\text{out}}$, CODA w/o β_t^g, and full CODA; β_t^g speeds up convergence and dampens late-stage fluctuations. (b) Generated examples with token commitment-order maps, where blue and red denote earlier and later commitments.

duces images that are globally faithful and locally coherent, while further handling references such as Paris's "most famous iron tower."

3.4 Further Analysis

3.4.1 Analysis of the CODA Process

Figure 4 shows the training reward curves under three settings: Only $\mathcal{L}_{\text{out}}$, CODA w/o β_t^g, and full CODA. The Only $\mathcal{L}_{\text{out}}$ baseline relies solely on a trajectory-level outcome reward and assigns the same advantage to every action, providing a sparse and ambiguous signal that cannot separate a useful commitment from an error within an otherwise successful trajectory. In contrast, CODA converts the shared final-image objective into decision-specific supervision through Semantic Reasoning Guidance and Masked Trajectory Guidance, attributing reward to concrete decisions such as semantic refinement and visual token commitments. As a result, both CODA variants achieve clearly higher rewards than the outcome-only baseline and reach a substantially higher plateau. Moreover, compared with CODA w/o β_t^g, full CODA converges more quickly and exhibits smaller late-stage fluctuations, suggesting that β_t^g up-weights the commitments with the greatest effect on subsequent generation, stabilizing structural decisions earlier and yielding a more consistent layout.

3.4.2 Visualization of Visual Commitment Order

Figure 4 visualizes the commitment order of visual tokens for the baseline and CODA, where cooler cells indicate earlier commitments and warmer cells later ones. Under CODA, object contours such as the boundaries of the apple, ceramic jar, and umbrellas are committed early, whereas homogeneous background regions are resolved later. In contrast, the baseline shows no comparable priority and fixes object boundaries and background in a less structured order. This offers a spatially grounded view of the influence measure, since the positions that most constrain the remaining generation, namely object boundaries, are committed before the smoother regions that depend on them, yielding an interpretable coarse-to-fine commitment pattern.

4 Discussion and Conclusion

We introduced CODA, a reinforcement learning framework that provides decision-level supervision for unified text-to-image generation. Existing methods broadcast a single final-image reward, leaving two difficulties. First, *ambiguous semantic attribution* entangles high-level planning with low-level visual realization. Second, *sparse visual supervision* leaves dependent decisions without intermediate feedback and overlooks their unequal structural influence. CODA resolves both through contrastive decision advantages. Semantic Reasoning Guidance isolates the effect of prompt optimization, while Masked Trajectory Guidance scores visual commitments and weights them by the induced KL divergence to emphasize structurally influential decisions. Experiments based on the 1.3B Show-o demonstrate outstanding performance on GenEval, T2I-CompBench, and WISE.

AI USE STATEMENT

Qwen3.5 was used to generate 15K training prompts covering the 80 COCO object classes and to extract object and attribute annotations for reward computation. Generative AI tools also assisted with language polishing and manuscript organization. The authors are responsible for the accuracy and integrity of the final manuscript and accompanying materials.

ETHICS STATEMENT

This work uses synthetic prompts and pretrained text-to-image and evaluation models. As a result, the reported reward signals may reflect errors or biases in the underlying detector and preference model, particularly for object recognition and attribute binding. No human-subject data or personally identifiable information are used. We will follow the licenses and usage terms of all released models, datasets, and assets.

REPRODUCIBILITY STATEMENT

The paper describes CODA's constraint-anchored outcome evaluation, Semantic Reasoning Guidance, Masked Trajectory Guidance, and joint optimization objective. The Appendix expands the prompt construction, matched counterfactual sampling, influence calculation, advantage normalization, and implementation details.

REFERENCES

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

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

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.

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

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with CoT? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26295–26305, 2026.

Kai Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level CoT. In *Advances in Neural Information Processing Systems*, 2025.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

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 *European Conference on Computer Vision*, pp. 740–755, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Yifu Luo, Xinhao Hu, Keyu Fan, Haoyuan Sun, Zeyu Chen, Bo Xia, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement learning meets masked generative models: Mask-GRPO for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.13418.

Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. URL https://arxiv.org/abs/2503.07265v1.

Yidong Ouyang, Zhe Wang, Sourav Bhabesh, and Dmitriy Bespalov. Reinforcing the generation order of multimodal masked diffusion models. *arXiv preprint arXiv:2607.08056*, 2026.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Stability AI. Stable diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, 2024. Model release; method based on Esser et al. (2024), arXiv:2403.03206.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. MetaMorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Yunze Tong, Mushui Liu, Canyu Zhao, Didi Zhu, Wanggui He, Shiyi Zhang, Hongwei Zhang, Peng Zhang, Jinlong Liu, and Hao Jiang. Alleviating sparse rewards by modeling step-wise and long-term sampling effects in flow-based GRPO. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06422.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. SimpleAR: Pushing the frontier of autoregressive visual generation through pretraining, SFT, and RL. *arXiv preprint arXiv:2504.11455*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

M. Wu, L. Wang, P. Zhao, F. Yang, J. Zhang, J. Liu, Y. Zhan, W. Han, H. Sun, J. Ji, et al. Re-Prompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. *arXiv preprint arXiv:2505.17540*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

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. *arXiv preprint arXiv:2408.12528*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zixuan Ye, Quande Liu, Cong Wei, Yuanxing Zhang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhan Luo. Visual-aware CoT: Achieving high-fidelity visual consistency in unified models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9116–9126, 2026.

Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqin Yang, Qi Dai, Jianming Bao, Dongdong Chen, Chong Luo, and Lili Qiu. ReasonGen-R1: CoT for autoregressive image generation through SFT and RL. *arXiv preprint arXiv:2505.24875*, 2025.

Yanghao Zhou et al. Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer. *arXiv preprint arXiv:2509.16197*, 2025.

A Appendix

A.1 Overview

This appendix provides additional implementation details for CODA, including a discussion of related work, training-prompt construction, the constraint-based outcome evaluator, matched counterfactual rollouts, process-level supervision, optimization, and evaluation. We follow the notation introduced in Sections 2 and 3.1. Section A.2 discusses related work on CoT/RL for text-to-image generation and unified multimodal models. Section A.3 describes the training data, Sections A.4–A.6 give the main training details, Section A.7 describes the evaluation protocol, and Section A.9 presents additional qualitative results.

A.2 Related Work

CoT and RL for text-to-image generation. Recent works have explored chain-of-thought (CoT) and reinforcement learning (RL) to enhance text-to-image (T2I) generation. T2I-R1 (Jiang et al., 2025) formulates bi-level CoT that jointly optimizes semantic planning and visual synthesis, and RePrompt (Wu et al., 2025) instead targets reasoning-augmented prompt reformulation before generation. PARM (Guo et al., 2025) introduces stepwise verification and preference alignment to reinforce autoregressive image generation, while Thinking-while-Generating (Guo et al., 2026) interleaves textual reflection with visual synthesis, and Visual-Aware CoT (Ye et al., 2026) performs iterative visual planning and correction to preserve reference consistency. On the RL side, SimpleAR (Wang et al., 2025) and ReasonGen-R1 (Zhang et al., 2025) adopt the Group Relative Policy Optimization (GRPO) paradigm (Shao et al., 2024) for autoregressive image generation using CLIP-style or preference-based scalar rewards, whereas Mask-GRPO (Luo et al., 2025) extends GRPO to masked visual generation with a committed-token likelihood surrogate. TP-GRPO (Tong et al., 2026) derives incremental rewards from completed flow trajectories at selected turning points, and Ouyang et al. (2026) additionally train a controller to reorder unmasking positions. However, these methods primarily rely on a single terminal reward broadcast across the whole trajectory, or introduce heuristic step-level signals that still cannot separate the contribution of high-level semantic planning from that of low-level visual realization. In contrast, our CODA supervises each decision through contrastive advantages computed against matched alternatives under shared randomness, providing accurate semantic attribution and dense visual feedback within a single unified objective.

Unified multimodal models. Unified multimodal models integrate visual understanding and generation within a shared backbone, but differ in their visual representations and decoding mechanisms. One line of methods employs continuous embeddings and relies on an external diffusion decoder to complete synthesis: MetaMorph (Tong et al., 2024) predicts continuous visual features rendered by a diffusion module, and MetaQuery (Pan et al., 2025) bridges frozen multimodal encoders with pretrained diffusion generators through learned queries. Another line unifies text and image modalities in a discrete token space, ensuring architectural simplicity. Show-o (Xie et al., 2024) couples autoregressive text modeling with discrete masked image generation, while EMU3 (Wang et al., 2024) tokenizes images, text, and video into a shared vocabulary and trains a single transformer from scratch. Recent efforts further scale this recipe: Janus-Pro (Chen et al., 2025) decouples visual encoders for understanding and generation, BAGEL (Deng et al., 2025) pushes unified modeling to a 14B mixture-of-experts backbone, and Manzano (Zhou et al., 2025) balances the two abilities through a hybrid tokenizer design. These architectural advances make language-based planning available to visual generation but leave open how to translate the resulting semantic reasoning into faithful low-level realization. Our CODA is orthogonal to these architectural choices. Building on the 1.3B Show-o backbone, it introduces decision-level supervision through Semantic Reasoning Guidance and Masked Trajectory Guidance, and can in principle be applied to any unified model that exposes an explicit reasoning stage and a stepwise visual decoding process.

A.3 Training Prompts and Constraint Records

Data construction. We use Qwen3.5 to construct 15,000 training prompts from templates covering the 80 COCO object classes. Qwen3.5 is also used to extract object and attribute information from each prompt. For every example, we store the original prompt P together with its structured constraint record.

The constraint record is extracted from P before policy optimization and remains fixed during training. Therefore, the evaluation target always follows the original instruction. For example, if an optimized prompt P' changes a color, count, or spatial relation, the generated image is still evaluated against the constraints in P.

Template families. We construct prompts from the five template families shown in Table 5. A prompt may contain more than one type of constraint. Qwen3.5 expands the templates into more natural descriptions while keeping the required objects, attributes, counts, and relations unchanged. We remove exact duplicates and discard prompts with contradictory attributes or invalid relations. The final training set contains 15,000 prompts.

Family	Template example
Existence	A photo of a {object}.
Counting	A photo of {number} {objects}.
Color binding	A {color} {object A} and a {color} {object B}.
Spatial relation	A {object A} to the {relation} of a {object B}.
Mixed composition	{number} {color} {objects A} beside a {object B}.

Table 5: Prompt templates used to construct the training set. Each template specifies one or more constraints that are evaluated during training.

Constraint representation. Each record contains the original prompt, canonical object names, detector queries, and the corresponding existence, count, color, and spatial-relation constraints. Spatial relations are represented as ordered subject–predicate–object triples.

For example, "a red cup to the left of a blue bowl" produces two existence constraints, two color constraints, and the ordered relation $(\text{cup}, \text{left of}, \text{bowl})$. Similarly, "four zebras" defines an existence constraint and an exact target count of four. A singular object mention is treated as an existence constraint unless the prompt explicitly requires exactly one object.

We normalize plural forms and common aliases during parsing and check that all objects referenced by attributes or relations are present in the object list. Prompts with unsupported or inconsistent constraints are removed before training.

Prompt optimization. The semantic branch rewrites the original instruction into a clearer visual description while preserving its content. We use the following instruction:

"Rewrite the instruction into a clear visual description for image synthesis. Preserve every requested object, count, color, and spatial relation. Clarify ambiguous phrasing without replacing the requested scene. Return only the revised description."

The original prompt P is appended to this instruction. The model generates an optimized description P', with a maximum output length of 128 tokens. Training is applied to the generated description and its termination token. If the generated description is empty or cannot be parsed, we fall back to the original prompt P and do not apply the semantic loss to that sample.

A.4 Constraint-Anchored Outcome Evaluator

Object detection and instance matching. We use GroundingDINO for object detection. The box threshold is set to 0.30, the text threshold to 0.25, and class-wise non-maximum suppression uses an IoU threshold of 0.50. Aliases are merged before suppression to avoid counting the same object more than once.

For multiple instances of the same class, detections are first sorted by confidence and assigned to different object instances without replacement. When several objects of the same class have different target colors, we use one-to-one matching based on object–color compatibility. The same assignment is used for the corresponding spatial-relation checks.

If a required object is not detected, its existence, color, and spatial-relation scores are zero. Count constraints use the number of detected instances, including zero.

Existence and counting. Existence is measured by

$$m_e = \mathbf{1}[\hat{n} \geq 1].$$

For counting, we use

$$m_n(\hat{n}, n^*) = \exp(-|\hat{n} - n^*|/\tau_n), \qquad \tau_n = 1, \tag{16}$$

where $\hat{n}$ is the detected count and n^* is the target count from the original prompt. The score is one when the count is correct and decreases smoothly as the counting error increases.

Spatial relations. We normalize bounding boxes to $[0, 1]^2$. Let (x_a, y_a) and (x_b, y_b) denote the centers of two detected objects. We use a spatial margin of $\eta_s = 0.05$. The left/right relations are evaluated as

$$x_b - x_a > \eta_s \quad \text{and} \quad x_a - x_b > \eta_s,$$

and above/below relations are evaluated as

$$y_b - y_a > \eta_s \quad \text{and} \quad y_a - y_b > \eta_s.$$

These relations are based on object centers and are therefore suitable for the relative 2D layouts considered in our training prompts. Relations that require depth, contact, or containment are not included in this constraint family.

Color binding. For color evaluation, we crop each detected object and compare it with text templates of the form "a photo of a {color} {object}" using CLIP ViT-L/14. We use a 5% margin around the detected box. The color vocabulary is

$$\{\text{black, white, gray, red, orange, yellow, green, blue, purple, pink, brown}\}.$$

The color with the highest CLIP similarity is taken as the predicted color and is compared with the target using an exact-match score. Object-level crops are used instead of the full image so that background color has less influence on the result.

Holistic preference. We use HPS v2.1 to provide a complementary image-level preference score with respect to the original prompt P. We normalize the raw HPS score as

$$\mathcal{R}_h(I, P) = \text{clip}\left(\frac{h(I, P) - a}{b - a}, 0, 1\right), \qquad b > a, \tag{17}$$

where a and b are the 5th and 95th percentiles measured on a held-out development set. These values are fixed throughout training.

A.5 Matched Rollouts and Semantic Attribution

Shared randomness. To compare P and P' under the same generation conditions, we use matched random streams for the two branches. Random variables are indexed by the prompt, trajectory, decoding step, visual-token position, and sampling type. This allows both branches to reuse the same random draws even when their decoding states differ.

During these comparisons, model parameters, tokenizer, decoding schedule, temperature, and evaluator remain fixed. Dropout is disabled during rollout generation.

Semantic pairs. For each optimized prompt P', we sample $K = 2$ matched image pairs from P and P'. All images are evaluated using the constraint record extracted from the same original prompt P. We then compute the semantic advantage using Eq. 9.

We use the mean paired reward difference directly and stop gradients through the reward. The resulting scalar provides supervision for all generated tokens in P'.

A.6 Joint Optimization and Training Configuration

Committed-token likelihoods. For each visual action, all newly committed visual tokens are evaluated at the same predecessor state using the bidirectional masked context. The policy ratio is computed from the joint score of the committed tokens:

$$\log \rho_t^g = \sum_{i \in \mathcal{U}_t^g} \left[\log p_\theta(\hat{Y}_{t,i}^g \mid s_t^g) - \log p_{\theta_{\text{old}}}(\hat{Y}_{t,i}^g \mid s_t^g) \right]. \tag{18}$$

We keep the sum over committed positions rather than averaging the token-wise ratios, since all tokens in $\mathcal{U}_t^g$ belong to one parallel visual action.

Loss and KL regularization. We average the semantic loss over generated text tokens and prompts. Outcome and process losses are averaged over their corresponding visual actions. The three terms are then combined using Eq. 15.

We apply the process and outcome advantages in separate clipped objectives. The rollout policy is kept fixed while advantages are computed and during each policy-update epoch.

For $\mathcal{K}$, we use the forward KL divergence from the current policy to the initial Show-o model. The KL is computed on sampled text states and committed visual-token positions. We average the text and visual terms separately and then take their mean. The initial Show-o model remains frozen throughout training.

Configuration	Value
Backbone / parameter count	Show-o / 1.3B
Training prompts	15,000
Learning rate	10^{-6}
Semantic pairs K / trajectories G / alternatives J	2 / 4 / 4
Reward models	GroundingDINO and HPS v2.1
Optimizer	AdamW
Adam coefficients / epsilon	$(0.9, 0.999)$ / 10^{-8}
Weight decay / gradient clipping	0.01 / 1.0
Effective prompt batch size	8
Training epochs	1
Update steps	125
Warmup / learning-rate schedule	3% / cosine decay
Precision / checkpointing	BF16 / enabled
Hardware	8 NVIDIA H20 GPUs
Fine-tuning scope	Unified transformer; tokenizer frozen
Policy-update epochs per rollout batch	1
Clipping range ϵ	0.2
Stabilizer δ	10^{-6}
λ_v / λ_p / λ_{KL}	1.0 / 0.5 / 0.01
Visual decoding steps	32
Process probe steps	$\{0, 4, 8, 12\}$
Text temperature / top-p	0.7 / 0.95
Maximum text output length	128 tokens
Visual-token temperature / CFG scale	1.0 / 3.5
Initial confidence temperature	1.0, linearly decayed
Base random seed	42

Table 6: Training and implementation settings used in our experiments.

Training loop. For each minibatch, we first load the original prompts and their cached constraint records. The current rollout policy generates one optimized prompt P' for each input. We then generate two matched semantic pairs and four original visual trajectories.

At each process probe, four alternative commitments are generated from the same predecessor state. These branches are completed using matched continuation randomness and are used to compute C_t^g and β_t^g. We then form the semantic, outcome, and process advantages, replay the stored states to compute policy ratios, and optimize Eq. 15. The rollout policy is refreshed after each update.

A.7 Evaluation Protocol and Ablations

Benchmark evaluation. We evaluate CODA on GenEval, T2I-CompBench, and WISE using the evaluator and aggregation protocol provided by each benchmark. The training reward $\mathcal{O}$ is used only during optimization and is not used as a replacement for benchmark metrics.

All compared methods use the same evaluation prompts and sample budget. For each run, we keep the decoding settings, resolution, and evaluation seeds fixed across methods. Checkpoints are selected using held-out development data rather than benchmark test results.

Metric reporting. We report benchmark values on their original scales. For percentage-valued metrics, absolute changes are reported in percentage points. For example, GenEval scores of 84.25 and 68.29 differ by 15.96 percentage points.

When confidence intervals are estimated, samples generated from the same prompt are grouped together during resampling so that correlated outputs are not treated as independent examples.

Controlled comparisons. For all ablations, we keep the backbone initialization, training prompts, optimizer, effective batch size, evaluation seeds, and training schedule fixed. We additionally compare matched and independent continuation randomness using the same values of K, G, and J, and study different probe locations while keeping the total number of probes unchanged.

These settings allow the effects of semantic supervision, process supervision, influence weighting, and matched rollouts to be measured under the same training conditions.

A.8 Discussion and Limitations

Our outcome evaluator focuses on object existence, counting, color binding, spatial relations, and holistic image quality. Its accuracy therefore depends on the underlying detector and image–text models. Missed detections or incorrect instance matching can affect several constraint scores at the same time. Color prediction can also be influenced by lighting, occlusion, or background content within an object crop.

The spatial evaluator is designed for simple 2D relations such as left, right, above, and below. More complex relations involving depth, contact, containment, material properties, or external knowledge are not explicitly modeled.

The process reward is estimated from a finite number of counterfactual branches. It therefore measures the relative effect of the sampled alternative commitments rather than all possible decoding decisions. The influence term measures how much a commitment changes the following token distribution, but a large distribution change alone does not imply a better generation. CODA combines this influence with the corresponding outcome change to obtain the process-level training signal.

Counterfactual rollouts also increase training cost. We probe only a subset of decoding steps and use four alternatives per probe to balance process-level supervision with computation. In addition, equal-sized alternatives are not always available near the end of decoding, so these late steps are excluded from process supervision when necessary.

A.9 Additional Visual Examples

Figures 5 and 6 present additional qualitative results from our experiments. All panels are generated by the evaluated models under the corresponding experimental settings. The images are shown using a common layout for easier visual comparison; no additional semantic image editing is applied.

The examples complement the quantitative results by showing typical differences in object completeness, counting, attribute binding, spatial composition, and overall visual fidelity.

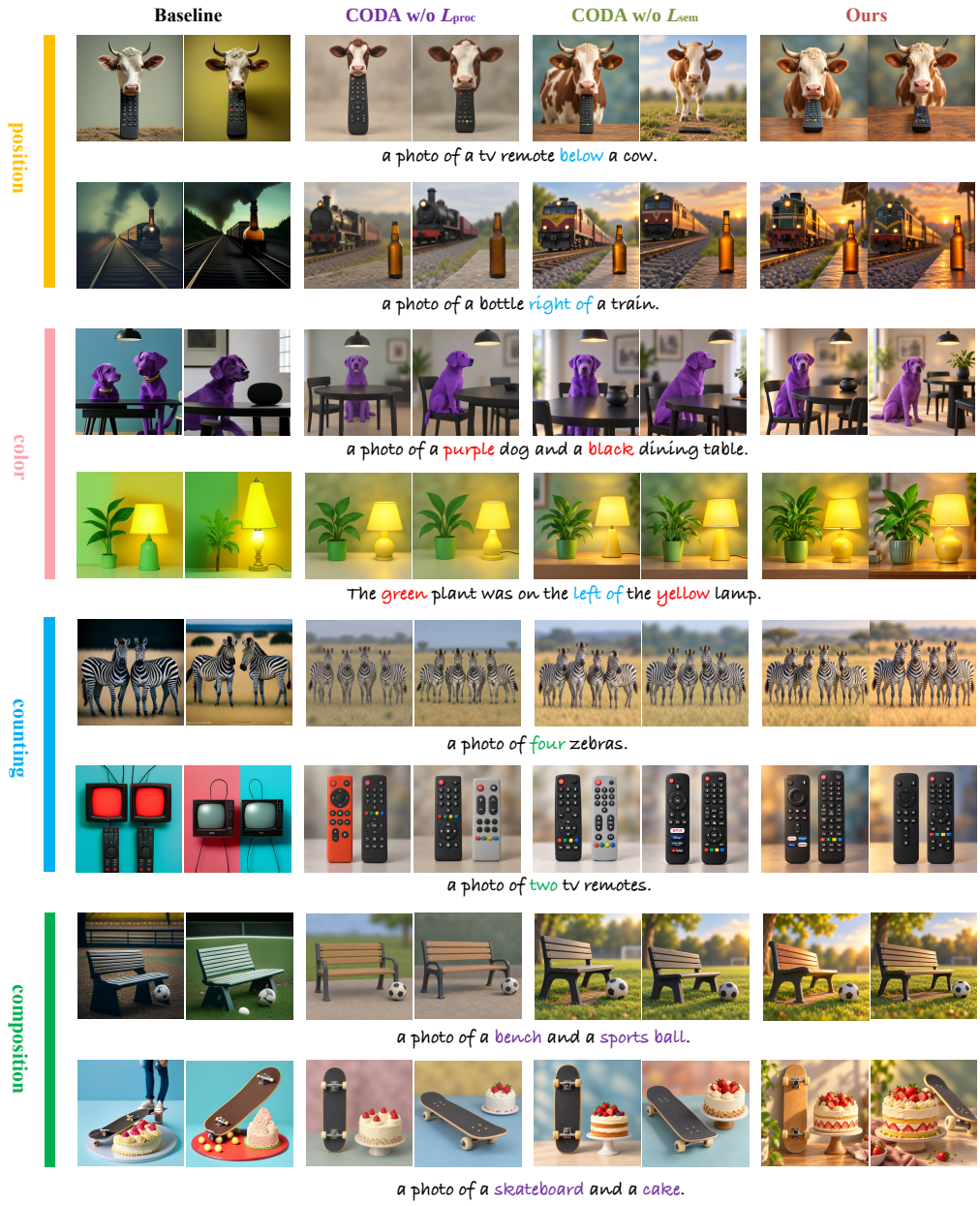


Figure 5: **Qualitative comparison of CODA and its ablations.** Results are shown for prompts covering spatial relations, color attributes, object counting, and multi-object composition. From left to right: baseline, CODA w/o L_{proc}, CODA w/o L_{sem}, and the full CODA model. Each setting shows two generations for the same prompt.

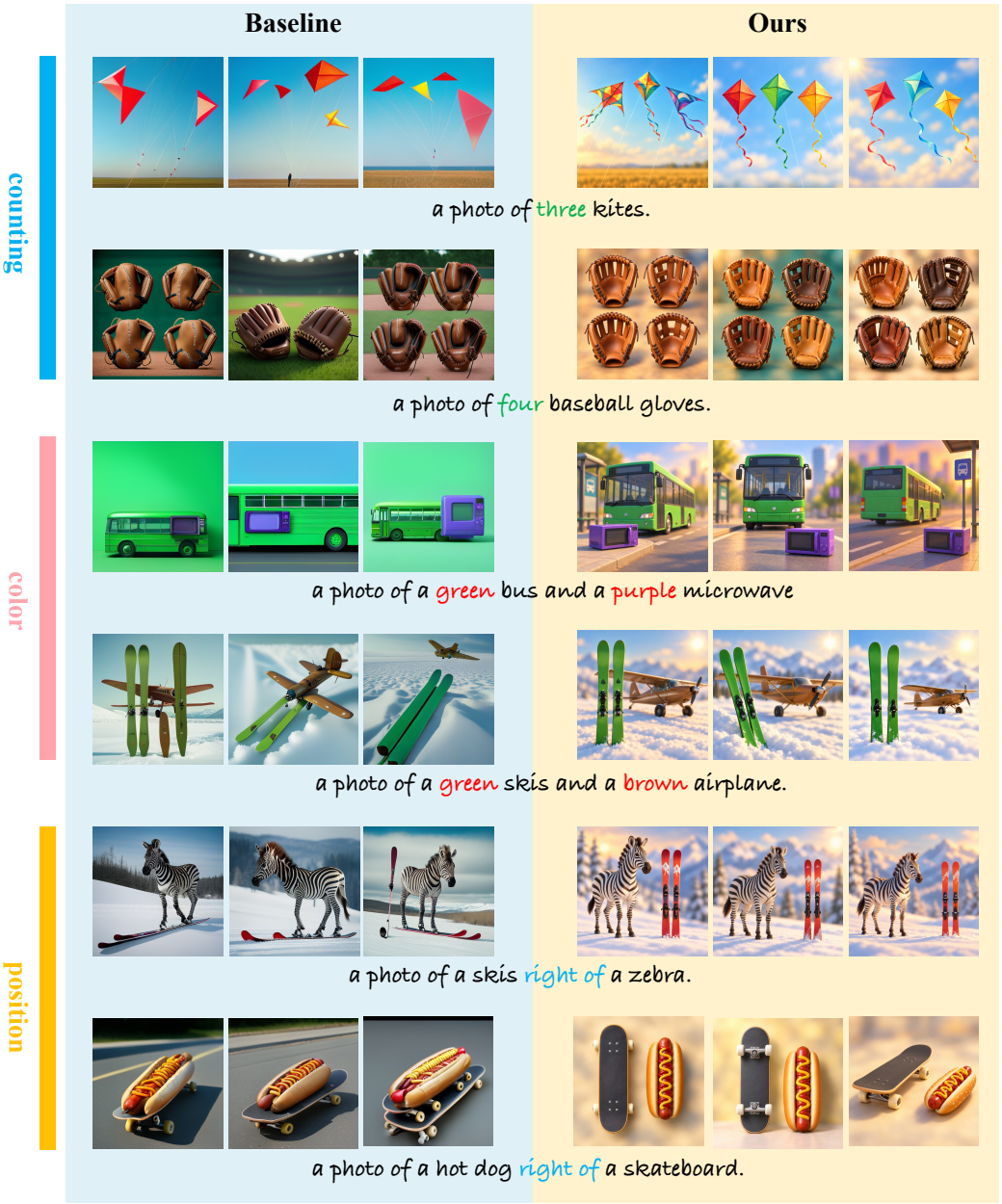


Figure 6: **Additional qualitative comparison between the baseline and CODA.** We show three generations per prompt for counting, color-binding, and spatial-relation examples. Both methods are evaluated using the same prompts and generation settings.