T-MOG: TEMPORALLY STRUCTURED MULTI-OBJECTIVE GUIDANCE FOR TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality image generation must satisfy multiple objectives, including semantic consistency and aesthetic preference, etc. However, existing multi-objective guidance mainly assigns weights across objectives by introducing multiple rewards at each step, implicitly assuming all objectives remain useful throughout generation. This parallel modeling overlooks the dynamics of diffusion, where semantics and spatial structure are usually established early, while local details are refined later. Thus, ignoring diffusion dynamics and continuously optimizing all objectives together can introduce both objective conflict and cross-stage interference. To address this, we propose **T-MOG**, a training-free multi-objective guidance framework with temporal decoupling. Unlike existing methods that coordinate objectives simultaneously, **T-MOG** redefines multi-objective guidance from an orthogonal perspective. It adaptively determines when and in what order different objectives should be optimized according to diffusion dynamics. Specifically, it selects the suitable guidance objective for each generation stage, enabling each objective to act when most effective. Further, we introduce a cross-stage memory protection mechanism that monitors the influence of a new objective's guidance direction on previous objectives after switching. It suppresses updates that may degrade earlier gains, balancing new-objective improvement with prior-optimization preservation. Comprehensive experiments demonstrate that T-MOG delivers more balanced joint improvements than single-objective guidance and static reward fusion while preserving previously optimized objectives.

1 INTRODUCTION

High-quality image generation (Ling et al., 2026; Ma et al., 2026; Honavar, 2025; van Krieken et al., 2026) must satisfy multiple requirements, including text-semantic structural consistency and visual aesthetics. In recent studies (Yan et al., 2026b), reward-model-based inference-time guidance has enabled external evaluation signals to directly influence the diffusion sampling process, improving target generation attributes without retraining the generative model. As multiple rewards are introduced into training-free methods, the generation process is optimized toward multiple objectives simultaneously. However, existing studies usually formulate this problem as multi-objective weighted fusion, coordinating multiple objectives at every denoising step (Cheng et al., 2026b). Since the optimization directions of different objectives are not fully aligned, improving one metric may suppress or even degrade another, leading to a typical 'seesaw effect' (Galashov et al., 2025).

The underlying cause of the above seesaw effect lies in a strong implicit assumption of current multi-objective fusion methods, namely that different objectives have similar temporal validity throughout the generation process (Yan et al., 2026c). However, diffusion generation is not temporally homogeneous. Recent frequency-domain analyses show that, in high-noise stages, usable data signals are mainly concentrated in low frequencies, while high-frequency components are still dominated by noise. As generation proceeds, the recoverable effective frequency spectrum gradually extends toward higher frequencies (Haridas et al., 2026; Kim et al., 2026). This dynamic reveals an issue overlooked by existing multi-objective guidance paradigms, where different optimization objectives are not equally effective at all timesteps. Objectives related to global semantics and major structures

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.

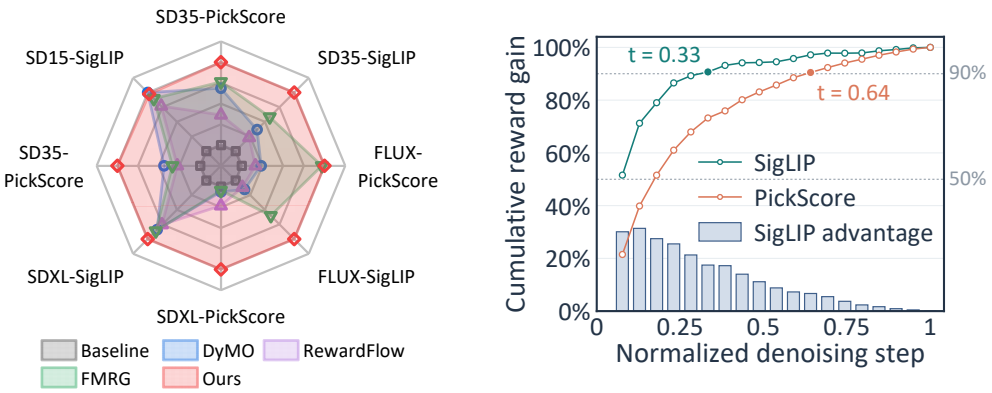


Figure 1: **Performance and Motivation**. **Left**: Overall performance of T-MOG across different generative backbones, demonstrating consistently strong multi-objective optimization. **Right**: Temporal evolution of structure- and preference-oriented rewards during unguided SD3.5 Medium sampling. The cumulative gains of SigLIP and PickScore are normalized by their respective final gains, while the bars indicate the gap between the two reward trajectories. SigLIP accumulates most of its gains during early denoising, whereas PickScore continues to develop during the middle and late stages. This temporal specialization motivates T-MOG to match each objective with its effective generation stage instead of optimizing all objectives throughout the entire trajectory.

can exert decisive influence on the trajectory in early generation stages. In contrast, objectives relying on local textures, visual quality, and human preferences often require clearer middle-to-late states to provide reliable evaluation and effective guidance. Therefore, multi-objective guidance involves not only conventional objective conflicts but also a more fundamental temporal validity mismatch (Golan et al., 2026). If all objectives act jointly from the start, some objectives may be applied prematurely before reliable evaluation. In contrast, others that have played their main role may become redundant or even interfere with other objectives in later stages.

As discussed above, an ideal multi-objective guidance method should align with the dynamics of image generation. In addition, it is equally important to protect earlier optimization gains while optimizing later objectives (Golan et al., 2026; Yan et al., 2026a; Li et al., 2026; Yu et al., 2026). Since image generation proceeds along a continuous trajectory, later objectives are not optimized from a state independent of earlier ones. When the model starts enhancing visual aesthetics or fine details, its guidance direction may still alter object relations, text consistency, and other information, thereby damaging structural attributes established earlier. In other words, beyond optimization conflicts among objectives at the same timestep, we should also consider cross-stage interference. During generation, later-introduced detail objectives may weaken earlier structure objectives through the trajectory. Therefore, a multi-objective guidance method following diffusion dynamics should not only decide when to optimize which objective, but also ensure later-objective improvements do not compromise attributes established earlier. More related work is provided in Appendix. A.

Based on the above observations, we propose **T-MOG** (**T**emporally **S**tructured **M**ulti-**O**bjective **G**uidance), a temporal multi-objective guidance framework for generative models. **T-MOG** no longer treats multiple objectives as static, parallel competitors throughout sampling; instead, it reorganizes multi-objective guidance into a temporally ordered optimization process based on stage-wise effectiveness along the generation trajectory. Specifically, we first characterize each objective's effective window along the denoising trajectory and dynamically determine the primary guidance objective based on the current generation state, allowing each objective to act mainly at the stage where it provides the most optimization value. As the gains of early objectives gradually saturate, **T-MOG** adaptively shifts the optimization focus to subsequent objectives instead of relying on predefined fixed temporal boundaries. Finally, we introduce a cross-stage anti-forgetting protection mechanism, which detects the impact of later-stage guidance on previously established objectives before applying it and suppresses updates that may significantly degrade historical objectives, thereby balancing new-objective improvement with the preservation of existing optimization gains.

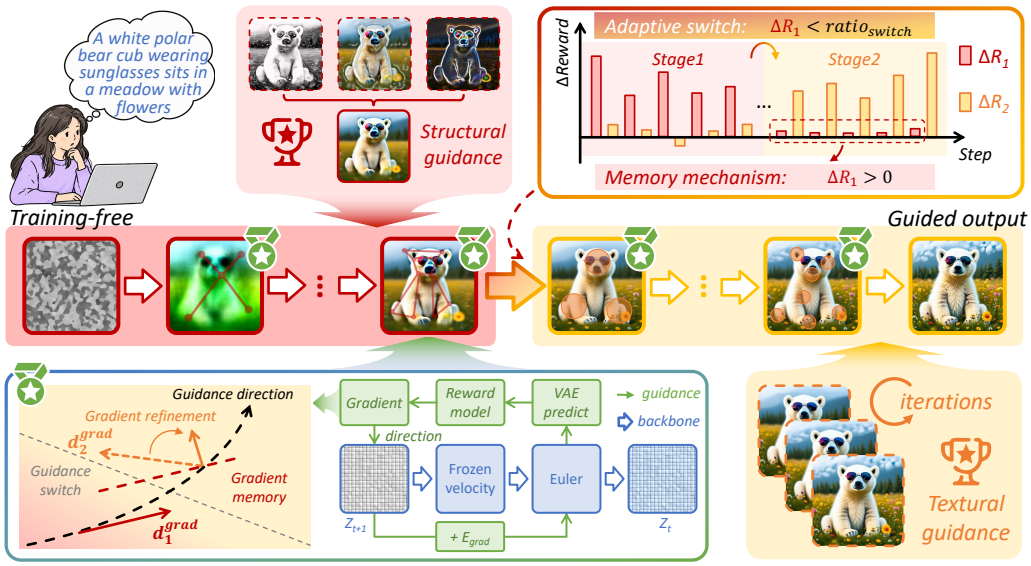


Figure 2: **Overview of T-MOG.** T-MOG adaptively switches from structure-oriented to detail-oriented guidance along the generation trajectory. Cross-stage protection suppresses updates that improve the current objective at the expense of previously established gains.

Therefore, **T-MOG** shifts the multi-objective problem in diffusion models away from the prevailing view of reducing optimization conflicts in multi-objective fusion. It instead focuses on two questions directly tied to the generation process. When should each objective be optimized? How can a new objective be improved without erasing previous ones? By jointly modeling the temporal order of objectives and cross-stage dependencies, we provide a new perspective for multi-objective guidance in generative models. This perspective differs from the current mainstream training-free multi-objective guidance paradigm. In summary, our contributions are threefold (1) We identify temporal mismatch in multi-objective guidance, so continuously optimizing them together can cause objective conflicts. (2) We transform multi-objective guidance from objective fusion into objective routing. **T-MOG** selects the most suitable objective for guidance based on the current generation state, reducing direct competition among objectives. (3) We introduce a cross-stage objective protection mechanism. It converts previous objectives from active optimization signals into preservation criteria for subsequent updates. Benefiting from the above innovations and our core idea of matching different objectives to appropriate generation stages for on-demand scheduling, experiments show that T-MOG effectively alleviates the 'seesaw effect' in conventional multi-objective optimization.

2 METHOD

T-MOG views multi-objective diffusion guidance as a temporal optimization process with three consecutive decisions, including which objective to optimize at the current stage, when to switch to the next objective, and how to preserve previously achieved gains while optimizing a new objective. Specifically, given multiple objectives, we first select the objective with the highest optimization value at the current generation stage for guidance. Then, by continuously monitoring the current objective's optimization gain and diffusion dynamics, **T-MOG** adaptively triggers objective transitions when the previous objective saturates and generation enters a new stage. Meanwhile, optimized objectives are retained as protection criteria for subsequent updates. When a new objective's guidance direction may degrade previous objectives, this update is suppressed. In this way, **T-MOG** transforms multi-objective guidance from full-process parallel optimization into adaptive on-demand guidance involving objective activation, transition, and cross-stage protection.

2.1 Problem Setup and Diffusion Guidance

2.1.1 Diffusion Models

The notation in this paper is defined as flow matching (Lipman et al., 2023). Given real data $\mathbf{x}_0$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ sampled from a standard Gaussian distribution, the forward interpolation trajectory is defined as $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\epsilon}, t \in [0,1]$, where $t = 0$ corresponds to a clean image and $t = 1$ corresponds to noise. The model learns a conditional velocity field $\mathbf{v}_t = \boldsymbol{\epsilon} - \mathbf{x}_0$, where c denotes the text condition. For a linear flow path, the target velocity can be written as $\mathbf{v}_t = \boldsymbol{\epsilon} - \mathbf{x}_0$. During inference, we start from $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$ and numerically integrate from $t = 1$ to $t = 0$ along the learned velocity field. For adjacent timesteps t and $t - \Delta t$, a basic first-order update can be written as $\mathbf{x}_{t-\Delta t} = \mathbf{x}_t - \Delta t\, \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c})$. To compute image-level rewards on intermediate generation states (Zhong et al., 2025), we further recover the corresponding clean-sample estimate using the current velocity:

$$\hat{\mathbf{x}}_{0|t} = f_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c}) = \mathbf{x}_t - t\, \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c}). \tag{1}$$

2.1.2 Reward Gradient Guidance

Given a differentiable reward function $R(\cdot)$, the generation quality is evaluated using the clean sample $\hat{\mathbf{x}}_{0|t}$ predicted from the current state. Instead of modifying model parameters, we adjust the current generation state only during inference along the reward gradient. Specifically, before performing the flow update, $\mathbf{x}_t$ is updated as follows:

$$\mathbf{q}_t = \nabla_{\mathbf{x}_t} R(\mathcal{D}(f_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c}))), \qquad \mathbf{x}_t \leftarrow \mathbf{x}_t + \eta_t \frac{\mathbf{q}_t}{\|\mathbf{q}_t\|_2}, \qquad \|\mathbf{q}_t\|_2 \neq 0. \tag{2}$$

Here, $\mathcal{D}$ denotes the VAE decoder, $\mathbf{q}_t$ is the image-reward gradient of the current latent $\mathbf{x}_t$, $\|\cdot\|_2$ denotes the Euclidean norm, and $\eta_t > 0$ is an adaptive step size. The reward is computed on the image decoded from the clean latent estimate, and its gradient is backpropagated to $\mathbf{x}_t$ through the decoder and $f_{\boldsymbol{\theta}}$. For simplicity, we denote the m consecutive reward-gradient refinement steps as $\tilde{\mathbf{x}}_t = g(\mathbf{x}_t, R, m)$. where m denotes the number of inner refinement steps at the same sampling timestep, g represents the refinement process with candidate acceptance, and $\tilde{\mathbf{x}}_t$ is the final retained state. The reward-guided state $\tilde{\mathbf{x}}_t$ is then used to continue the flow-matching generation update $\mathbf{x}_{t-\Delta t} = \tilde{\mathbf{x}}_t - \Delta t\, \mathbf{v}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t, t, \mathbf{c})$. Therefore, reward-gradient guidance can be understood as performing local test-time optimization at each sampling timestep, after which the pretrained flow model continues to move the updated state along the generation trajectory toward the data distribution.

2.2 When to Switch: Adaptive Objective Transition

Existing multi-objective guidance methods usually apply different rewards simultaneously throughout the generation trajectory, ignoring that the visual attributes associated with different objectives form at different stages. For the two objectives considered in this paper, we treat R_1 as the early-stage structural objective and R_2 as the late-stage detail objective. **T-MOG** does not predefine fixed stage boundaries; instead, it adaptively determines the separation point based on the evolution trend of the structural objective along the generation trajectory. Specifically, for adjacent generation states, we first compute the change in the structural objective:

$$\Delta r_t^{(1)} = R_1\big(\hat{\mathbf{x}}_{0|t-\Delta t}\big) - R_1\big(\hat{\mathbf{x}}_{0|t}\big). \tag{3}$$

This quantity measures the gain of the structural objective as generation progresses from t to $t - \Delta t$. In early generation stages, global semantics and spatial structure are still being formed, so $\Delta r_t^{(1)}$ usually changes noticeably. As the structure gradually stabilizes, the structural gain between adjacent timesteps also becomes smoother. Inspired by trend-aware generation-stage modeling, we further examine the trend of the structural gain itself:

$$\lim_{\Delta t \to 0} \frac{\Delta r_{t-\Delta t}^{(1)} - \Delta r_t^{(1)}}{\Delta t} \to 0. \tag{4}$$

Along the generation direction from $t = 1$ to 0, we denote the first position satisfying the above stability condition as $t^\star$. Since t decreases numerically, $t^\star$ can be written as follows:

$$t^\star = \max\left\{ t \in [0,1] \;\middle|\; \lim_{\Delta t \to 0} \frac{\Delta r^{(1)}_{t-\Delta t} - \Delta r^{(1)}_{t}}{\Delta t} \to 0 \right\}. \tag{5}$$

This adaptive boundary divides the generation trajectory into two stages. For $t > t^\star$, generation is in the structure-formation stage; the structural objective R_1 is used for reward-gradient guidance. For $t \leq t^\star$, the structure has largely stabilized, and the optimization focus shifts to the detail- and aesthetics-oriented objective R_2. Therefore, the guidance objective can be uniformly written as:

$$R_t = \begin{cases} R_1, & t > t^\star, \\ R_2, & t \leq t^\star. \end{cases} \tag{6}$$

Here R_1 denotes a metric reflecting image structure and semantics, and R_2 denotes a metric reflecting image details. Accordingly, the current reward-gradient refinement is given by: $\tilde{\mathbf{x}}_t = g(\mathbf{x}_t, R_t, m)$. Unlike fixed-timestep partitioning, $t^\star$ is adaptively determined by the actual evolution of the structural objective in the current generation trajectory. Thus, different prompts and generated samples can yield different objective switching points. Through this mechanism, **T-MOG** explicitly decouples two objectives that may compete under simultaneous optimization into a structure-formation stage and a detail-refinement stage. This allows each objective to mainly act within the temporal region matching its associated generative attribute.

2.3 How to Preserve: Cross-Stage Objective Protection

When optimization enters the second detail-refinement stage, the reward gradient produced by R_2 still acts on the same continuous generation trajectory as in the previous stage. As a result, updates that improve the detail metric may alter the semantic structure established earlier and degrade R_1. To address this, we introduce a cross-stage objective protection mechanism that makes second-stage reward-guided refinement aware of structural consistency. For $t \leq t^\star$, we compute the two reward gradients separately. For $i \in \{1, 2\}$, we have: $\mathbf{q}_t^{(i)} = \nabla_{\mathbf{x}_t} R_i(\mathcal{D}(f_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c})))$. Here i is the objective index, $\mathbf{q}_t^{(1)}$ denotes the gradient of the previous structural objective, and $\mathbf{q}_t^{(2)}$ denotes the gradient of the current detail objective. The gradient of R_1 is used only to constrain the current guidance direction, while the active optimization objective in the later stage remains R_2. We measure the degree of directional conflict using the inner product of the two gradients $s_t = \left\langle \mathbf{q}_t^{(2)}, \mathbf{q}_t^{(1)} \right\rangle$. Here s_t is the gradient inner product, and $\langle \cdot, \cdot \rangle$ denotes the inner product. When $s_t < 0$, we remove the component of the R_2 gradient that conflicts with R_1:

$$\overline{\mathbf{q}}_t^{(2)} = \begin{cases} \mathbf{q}_t^{(2)} - \dfrac{s_t}{\|\mathbf{q}_t^{(1)}\|_2^2} \mathbf{q}_t^{(1)}, & s_t < 0, \\ \mathbf{q}_t^{(2)}, & s_t \geq 0. \end{cases} \tag{7}$$

We use $\overline{\mathbf{q}}_t^{(2)}$ to denote the corrected gradient of R_2. When generating the candidate state $\tilde{\mathbf{x}}_t$, we use $\overline{\mathbf{q}}_t^{(2)}$ in the reward-gradient refinement above to substitute $\mathbf{q}_t$ in Eq. 2. We obtain the candidate state $\tilde{\mathbf{x}}_t$ using the R_2 gradient after removing the conflicting direction, and check whether it affects R_1:

$$\Delta r_t^{(1\leftarrow 2)} = R_1\left(f_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t, t, \mathbf{c})\right) - R_1\left(\hat{\mathbf{x}}_{0|t}\right). \tag{8}$$

where $\mathbf{x}_t$ is the current state before guidance, and $\Delta r_t^{(1\leftarrow 2)}$ measures the influence of the current R_2 guidance on the previous structural metric R_1. When $\Delta r_t^{(1\leftarrow 2)} \geq 0$, the detail optimization does not damage the established structural attributes, so the current reward-guided update is accepted. Conversely, when $\Delta r_t^{(1\leftarrow 2)} < 0$, the current R_2 guidance may benefit the detail objective but degrade the previous structural objective. In this case, we discard this guidance step and revert to the original state $\mathbf{x}_t$. Thus, the state actually used for the subsequent flow update is written as:

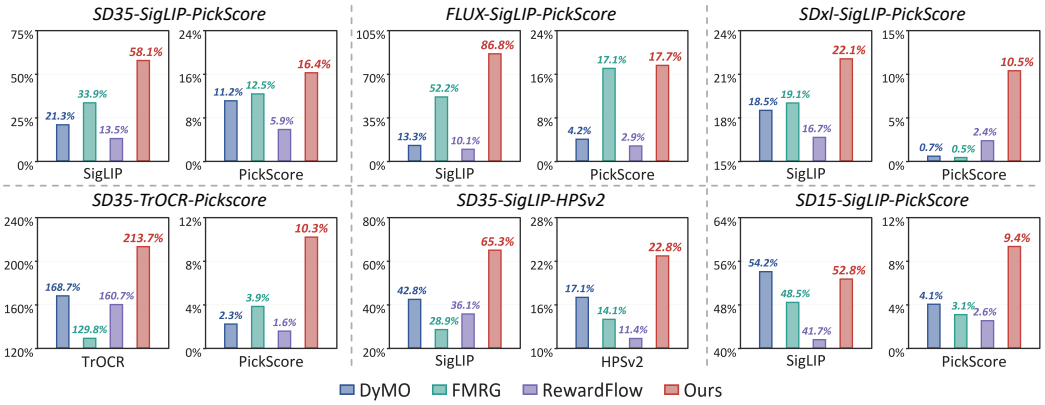


Figure 3: **Relative gains under multi-objective guidance.** We report the improvement of each method's multi-objective configuration over the corresponding unguided backbone for both objectives. T-MOG consistently achieves stronger joint gains across different backbones and reward pairs.

$$\bar{\mathbf{x}}_t = \begin{cases} \tilde{\mathbf{x}}_t, & \Delta r_t^{(1\leftarrow 2)} \geq 0, \\ \mathbf{x}_t, & \Delta r_t^{(1\leftarrow 2)} < 0. \end{cases} \tag{9}$$

The generation process then proceeds as follows: $\mathbf{x}_{t-\Delta t} = \bar{\mathbf{x}}_t - \Delta t\, \mathbf{v}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_t, t, \mathbf{c})$. This design ensures that R_1 does not disappear after early active optimization, but instead transforms from an active objective into a preservation criterion for later stages.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Backbones. We evaluate **T-MOG** on four text-to-image generative backbones: SD3.5 (Esser et al., 2024), FLUX (Black Forest Labs, 2024), SDXL (Podell et al., 2024), and SD1.5 (Rombach et al., 2022). These backbones cover both recent flow-based models and conventional latent diffusion models, allowing us to assess whether the proposed temporal multi-objective guidance generalizes across different model architectures and generation dynamics.

Baselines. We compare **T-MOG** against DyMO (Xie & Gong, 2025), FMRG (Huang et al., 2026), and RewardFlow (Susladkar et al., 2026). For each method, we consider guidance using only the first reward, only the second reward, and their multi-objective fusion. We additionally include the corresponding base model without reward guidance. These unguided references reveal the gains or degradations introduced by reward optimization, while the single-objective and fusion variants expose whether a method improves both objectives or merely specializes in one of them.

Optimization Objectives and Metrics. We consider three complementary objective pairs. The standard setting combines SigLIP cosine similarity (Zhai et al., 2023), which evaluates prompt–image semantic alignment, with PickScore (Kirstain et al., 2023), which reflects human visual preference. We further evaluate TrOCR (Li et al., 2023) with PickScore to study text rendering and SigLIP with HPSv2 (Wu et al., 2023) to test an alternative preference objective. For every setting, we report both reward-model scores rather than only the objective favored by each method, and higher values indicate better performance. This evaluation directly measures whether improving a new objective preserves the gains established by the preceding objective.

3.2 QUANTITATIVE COMPARISON

Comparisons across Backbones. We first evaluate the standard SigLIP–PickScore setting on SD3.5, FLUX, SDXL, and SD1.5. As shown in the SigLIP–PickScore panels of Fig. 3 and the first eight rows of Tab. 1, existing guidance methods can be highly effective for the reward they directly favor, but this specialization frequently provides limited benefit to the paired objective. Direct

Table 1: Quantitative comparison across different base models and rewards. Reward 1 (R1)/Reward 2 (R2) denote SigLIP/PickScore for the standard benchmarks, TrOCR/PickScore for OCR, and SigLIP/HPSv2 for the HPSv2 setting. Rank is computed among all reported methods in each row, with higher values ranked better.

Base	Reward	Baseline	DyMO			FMRG			RewardFlow			Ours	
			R1 Only	R2 Only	Fusion	R1 Only	R2 Only	Fusion	R1 Only	R2 Only	Fusion	Score	Rank
SD3.5	R1:SigLIP	0.1871	0.2386	0.1939	0.2270	0.2614	0.1899	0.2506	0.2160	0.1906	0.2123	**0.2958**	1
	R2:PickScore	23.914	23.320	27.576	26.597	23.534	27.753	26.900	24.341	25.836	25.335	**27.824**	1
FLUX	R1:SigLIP	0.1778	0.2125	0.1849	0.2015	0.2866	0.1860	0.2706	0.1996	0.1800	0.1958	**0.3321**	1
	R2:PickScore	24.543	24.279	26.586	25.566	24.295	**29.658**	28.750	24.728	25.477	25.265	28.890	2
SDXL	R1:SigLIP	0.1766	0.2125	0.1848	0.2109	0.2164	0.1840	0.2118	0.2067	0.1841	0.2076	**0.2172**	1
	R2:PickScore	23.770	22.860	25.930	23.970	23.470	25.640	23.930	23.950	25.910	24.390	**26.300**	1
SD1.5	R1:SigLIP	0.1403	**0.2189**	0.1643	0.2163	0.2090	0.1515	0.2084	0.1965	0.1455	0.1988	0.2144	3
	R2:PickScore	21.726	21.910	23.442	22.616	21.778	**24.243**	22.410	21.805	23.847	22.289	23.765	3
SD3.5	R1:TrOCR	0.2096	**0.6610**	0.2086	0.5631	0.5099	0.2185	0.4816	0.5922	0.2440	0.5465	0.6575	2
	R2:PickScore	22.442	21.166	23.805	22.954	22.072	23.760	23.316	22.131	23.195	22.809	**24.744**	1
SD3.5	R1:SigLIP	0.1808	0.2683	0.1878	0.2581	0.2410	0.1812	0.2330	0.2623	0.1874	0.2460	**0.2988**	1
	R2:HPSv2	0.3416	0.3091	**0.4335**	0.4000	0.3399	0.4067	0.3897	0.3369	0.3959	0.3807	0.4195	2

reward fusion only partly alleviates this issue, since heterogeneous objectives still compete along the same generation trajectory. In contrast, **T-MOG** improves both semantic alignment and visual preference, achieving a stronger overall trade-off across backbones. This consistency supports our core motivation: objectives are better coordinated by matching them to effective generation stages rather than optimizing them simultaneously at every timestep.

Generalization across Different Objective Pairs. We further evaluate **T-MOG** under TrOCR–PickScore and SigLIP–HPSv2, extending the optimization from semantic alignment and visual preference to text rendering and alternative preference modeling. In Fig. 3 and the last four rows of Tab. 1, **T-MOG** preserves the same joint-improvement pattern across heterogeneous reward pairs, whereas baselines tend to favor one objective or obtain limited joint gains. The consistent behavior indicates that T-MOG is not a reward-specific combination rule. Instead, adaptive objective transition routes each reward to the stage where its signal is informative, while cross-stage protection prevents later preference optimization from erasing attributes established earlier. These results validate temporal organization as a general mechanism for training-free multi-objective guidance.

3.3 Qualitative Comparison

Semantic Alignment and Visual Preference. As the top panel of Fig. 4, **T-MOG** consistently preserves prompt-critical subjects, attributes, and styles while producing coherent and visually detailed results on both FLUX and SD3.5. This balance shows that later preference optimization improves visual quality without overwriting semantic structure established by the earlier alignment objective.
Generalization to Text Rendering and Alternative Preference. The bottom panel of Fig. 4 shows similar behavior under TrOCR–PickScore and SigLIP–HPSv2. **T-MOG** preserves the task-specific text or semantic content while maintaining strong perceptual quality, demonstrating that its temporal organization generalizes beyond a particular backbone or reward pair.

3.4 Ablation Studies

We conduct four groups of ablations covering objective coordination, stage-wise scheduling, cross-stage protection, and reward-view assignment. All variants are compared with the full **T-MOG** to evaluate balanced improvements across objectives rather than isolated gains.
Temporal Coordination of Multiple Objectives. In Fig. 5-middle, it can be observed that optimizing either SigLIP or PickScore alone improves its favored objective while degrading the other, reproducing the seesaw effect. By assigning the two objectives to their effective generation stages, **T-MOG** maintains a more balanced multi-objective solution.
Stage-wise Objective Scheduling. Fig. 5-right show that stage division alone is insufficient: the objective order and guidance budget must follow the coarse-to-fine generation dynamics. Excessive

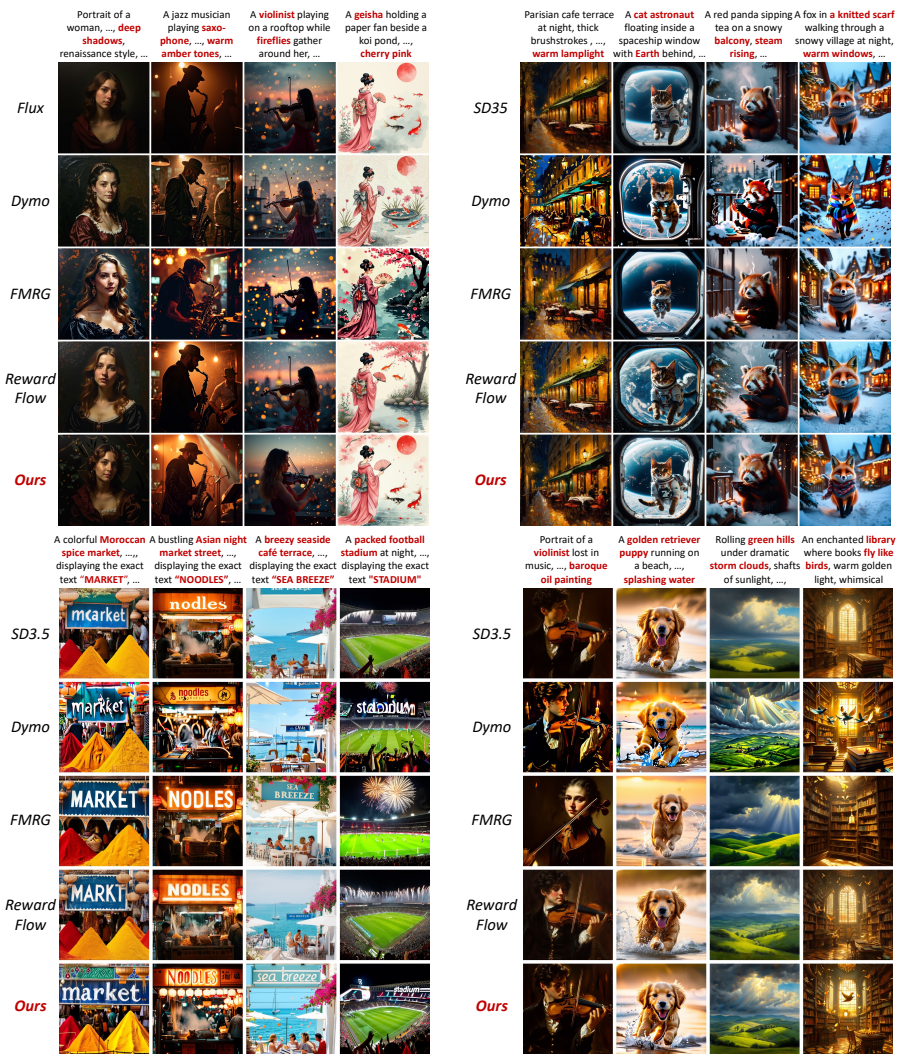


Figure 4: **Qualitative comparisons across backbones and objective pairs.** The top panel compares SigLIP–PickScore guidance on FLUX and SD3.5, while the bottom panel evaluates TrOCR–PickScore and SigLIP–HPSv2 on SD3.5. **T-MOG** achieves a favorable balance between prompt consistency and visual quality across backbones and objective pairs.

early structural guidance becomes redundant after saturation, whereas premature preference guidance acts before a stable semantic structure is formed, supporting the structural-first scheduling.

Complementary Cross-stage Protection. Fig. 6-middle validates the complementary roles of gradient and score memory. Gradient memory removes conflicting directions before refinement, while score memory rejects residual updates that damage previously established objectives. Combining both mechanisms enables later refinement while preserving earlier structural gains.

Stage-specific Reward Views. Fig. 6-right shows that using the same reward view throughout generation overlooks the changing reliability of visual signals. Augmented views guide early generation robustly, whereas the canonical view better preserves fine appearance during later refinement. This stage-specific assignment complements temporal objective routing.

3.5 Subjective Evaluation

We further employ ChatGPT as an LVM judge to evaluate six perceptual dimensions: Structural Faithfulness (SF), Aesthetic Appeal (AA), Fine-grained Detail (FD), Semantic Alignment (SA), Prompt Responsiveness (PR), and Anti-Hacking (AH). As shown in Fig. 7, evaluations across four configurations involving different generation backbones, visual encoders, and rewards consistently

	SigLIP	PickScore
A1	✓	✗
A2	✗	✓
Ours	✓	✓

	Stage 1	Stage 2
A3	SigLIP 3R	Fusion 1R
A4	PickScore 1R	Fusion 3R
Ours	SigLIP 1R	Fusion 3R

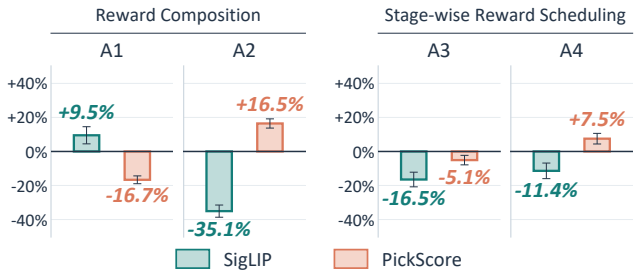


Figure 5: **Ablation analysis of different design variants.** Left: Summarizes the configurations of the ablation variants. Middle: Compares different reward compositions. Right: Compares different stage-wise reward scheduling strategies. ***R*** denotes Round.

	Grad. Mem.	Score Mem.
A5	✓	✗
A6	✗	✓
Ours	✓	✓

	Aug. Views	Canon. View
A7	Throughout	–
A8	–	Throughout
Ours	Structural	Textural

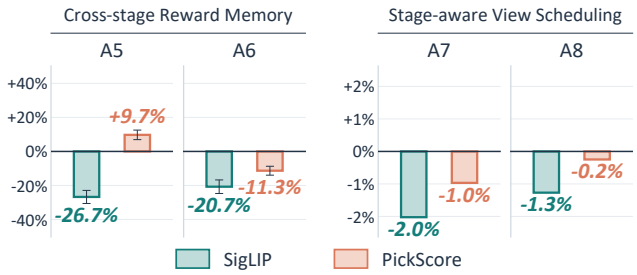


Figure 6: **Ablation analysis of different design variants.** Left: Summarizes the configurations of the ablation variants. Middle: Compares gradient- and score-memory strategies. Right: Compares different assignments of augmented and canonical views.

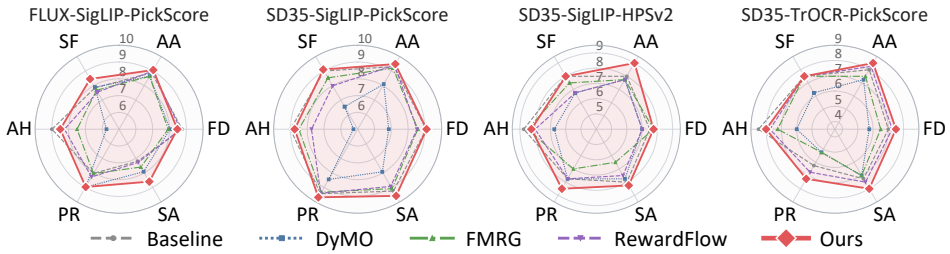


Figure 7: **LVM-based subjective evaluation.** Advanced MLLM (i.e., GPT5.5) evaluates six perceptual dimensions across diverse backbone, encoder, and reward configurations. **T-MOG** consistently achieves superior subjective quality.

favor **T-MOG** over competing approaches. In particular, **T-MOG** exhibits clear gains in semantic alignment and prompt responsiveness while maintaining strong structural fidelity, aesthetics, and fine-grained details, indicating that its reward improvements translate into perceptually meaningful generation quality.

4 Conclusion

We revisit multi-objective guidance for text-to-image diffusion models from a temporal perspective. Rather than applying multiple rewards in parallel throughout sampling, **T-MOG** organizes objectives according to their stage-wise effectiveness along the generation trajectory. It adaptively transitions from structure-oriented to detail-oriented guidance by monitoring the evolution of the early objective, and employs cross-stage protection to suppress later updates that would compromise previously established semantic and structural attributes.

AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were employed mainly to refine the language and improve the overall clarity and flow of the writing.

REFERENCES

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. *Advances in Neural Information Processing Systems*, 38:8583–8609, 2026.

Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. *Advances in neural information processing systems*, 38:8522–8549, 2026.

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

Jikang Cheng, Hao Shen, Xueyi Zhang, Guangcheng Wang, Zhongyuan Wang, Renye Yan, and Baojin Huang. Infodense: Density-aware regional decisive replay for memory-efficient incremental face forgery detection. *arXiv preprint arXiv:2607.16873*, 2026a.

Min Cheng, Fatemeh Doudi, Dileep Kalathil, Mohammad Ghavamzadeh, and Panganamala Kumar. Diffusion blend: Inference-time multi-preference alignment for diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 51151–51181, 2026b.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, pp. 12606–12633, 2024.

Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A fourier space perspective on diffusion models. *arXiv preprint arXiv:2505.11278*, 2025.

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Diffusion models and representation learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Alexandre Galashov, Ashwini Pokle, Arnaud Doucet, Arthur Gretton, Mauricio Delbracio, and Valentin De Bortoli. Learn to guide your diffusion model. *arXiv preprint arXiv:2510.00815*, 2025.

Shelly Golan, Michael Finkelson, Ariel Bereslavsky, Yotam Nitzan, and Or Patashnik. Paretoslider: Diffusion models post-training for continuous reward control. *arXiv preprint arXiv:2604.20816*, 2026.

Akash Haridas, Utkarsh Saxena, Parsa Ashrafi Fashi, Mehdi Rezagholizadeh, Vikram Appia, and Emad Barsoum. Dc-dit: Adaptive compute and elastic inference for visual generation via dynamic chunking. *arXiv preprint arXiv:2603.06351*, 2026.

Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. *ICLR*, 2025.

Jerry Y. Huang, Justin Lin, Sheel Shah, Kartik Nair, and Nicholas M. Boffi. How to guide your flow: Few-step alignment via flow map reward guidance. In *International Conference on Machine Learning*, 2026.

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. *Advances in Neural Information Processing Systems*, 38:97394–97430, 2026.

Dahye Kim, Deepti Ghadiyaram, and Raghudeep Gadde. Ddit: Dynamic patch scheduling for efficient diffusion transformers. *arXiv preprint arXiv:2602.16968*, 2026.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663, 2023.

Haodong Li, Tianfei Ren, Xiaoxiao Ma, Chunmei Qing, Zhen Fang, Sipeng He, Ziyu Guo, Haoyu Wu, Juanxi Tian, Yihang Zou, et al. Videococo: Code-as-cot for physically-consistent video generation via an agentic dual-engine system. *arXiv preprint arXiv:2607.27380*, 2026.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102, 2023.

Zeyu Ling, Xinyao Yu, Renye Yan, Jikang Cheng, Zhanke Wang, Qing Shuai, and Changqing Zou. Gentrack: Physical alignment for robot-native motion generation and zero-shot humanoid tracking. *arXiv preprint arXiv:2608.01410*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Qichao Ma, Jikang Cheng, Ling Liang, Zhaofei Yu, Tiejun Huang, and Renye Yan. Can we perform online rl for image editing without editing rewards? *arXiv preprint arXiv:2608.22780*, 2026.

David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng, and Angjoo Kanazawa. Flow matching policy gradients. In *International Conference on Learning Representations*, volume 2026, pp. 36352–36372, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Yair Schiff, Subham Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 43776–43821, 2025.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1185–1194. IEEE, 2025.

Onkar Kishor Susladkar, Dong-Hwan Jang, Tushar Prakash, Adheesh Sunil Juvekar, Vedant Shah, Ayush Barik, Nabeel Bashir, Muntasir Wahed, Ritish Shrirao, and Ismini Lourentzou. Rewardflow: Generate images by optimizing what you reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20066–20076, 2026.

Emile van Krieken, Pasquale Minervini, Edoardo Maria Ponti, and Antonio Vergari. Neurosymbolic diffusion models. *Advances in neural information processing systems*, 38:135889–135931, 2026.

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.

Xin Xie and Dong Gong. Dymo: Training-free diffusion model alignment with dynamic multi-objective scheduling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13220–13230, 2025.

Renye Yan, Jikang Cheng, Shikun Sun, Yi Sun, You Wu, Wei Peng, Zongwei Wang, Ling Liang, Junliang Xing, and Yimao Cai. Do less, achieve more: Do we need every-step optimization for rl fine-tuning of diffusion models? *arXiv preprint arXiv:2605.15855*, 2026a.

Renye Yan, Jikang Cheng, You Wu, Bojin Huang, Wei Peng, Zongwei Wang, Ling Liang, and Yimao Cai. Swiftexplorer: Training-free diffusion model alignment with swift diversity exploration. *arXiv preprint arXiv:2609.06651*, 2026b.

Renye Yan, Jikang Cheng, You Wu, Wei Peng, Zongwei Wang, Ling Liang, and Yimao Cai. Explore or converge? stage-guided per-step optimization for diffusion models. *arXiv preprint arXiv:2608.06768*, 2026c.

Jiayang Yu, Jialun Zhong, and Lei Zou. Ragocr: Optical compression of retrieval-augmented text via visual representation. *arXiv preprint arXiv:2608.00765*, 2026.

Stefano Zampini, Jacob K Christopher, Luca Oneto, Davide Anguita, and Ferdinando Fioretto. Training-free constrained generation with stable diffusion models. *Advances in Neural Information Processing Systems*, 38:27285–27316, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*, 2025.

APPENDIX

A RELATED WORK EXPANSION

A.1 DIFFUSION MODELS

Diffusion models (Jo & Hwang, 2026; Schiff et al., 2025; Cheng et al., 2026a; Falck et al., 2025; Singhal et al., 2025; Fuest et al., 2026; Balcerak et al., 2026) have recently become an important class of generative modeling methods. Their core idea is to construct a forward process that gradually adds noise and learn the corresponding reverse denoising process, thereby generating samples from random noise toward the real data distribution. Early diffusion probabilistic models achieved stable and high-quality generation through maximum-likelihood training, showing strong sample quality and training stability in image generation. As research progressed, latent diffusion models further reduced the computational cost of high-resolution image generation. By performing denoising in a compressed latent space, they enabled large-scale text-conditioned image generation. Diffusion models trained on large-scale text-image data then advanced text-to-image generation and substantially improved semantic consistency and visual quality. In recent years, researchers have further explored continuous forms of diffusion generation. Flow Matching (Stoica et al., 2025; McAllister et al., 2026; Bertrand et al., 2026) learns a continuous-time velocity field rather than relying on the traditional stochastic denoising process, offering a new modeling approach for efficient, controllable generation. Compared with discrete diffusion processes, flow-based generative models directly learn a continuous mapping from the noise distribution to the data distribution. This allows more flexible sampling strategies and has become an important direction for advanced text-to-image generation models. This paper studies multi-objective guidance based on flow-matching text-to-image generation models.

A.2 TRAINING-FREE GUIDANCE WITH MULTIPLE OBJECTIVES

Recently, guidance methods have provided an effective way to optimize generation results without additional training of generative models Zampini et al. (2026); Yan et al. (2026b). These methods usually use additional evaluation signals from pretrained models, such as text-image alignment, aesthetic quality, or other semantic attributes. By computing reward gradients and applying them to latent states during sampling, they guide generative models toward outputs that better match target preferences. Compared with directly retraining generative models, training-free guidance has lower computational cost, broader applicability, and greater flexibility for transfer to different generative models. As generation requirements have expanded from single-attribute optimization to joint optimization of multiple objectives, recent studies have further explored multi-objective guidance, where multiple reward signals are used simultaneously to guide the generation process. However, different objectives often correspond to different or even conflicting optimization directions. Existing methods therefore mainly focus on building reasonable trade-offs among objectives. For example, some methods combine multiple reward functions using fixed or dynamic weights, while others seek coordinated update directions through gradient fusion. These methods mainly address how to coordinate multiple objectives within the same generation stage. Nevertheless, optimization conflicts among objectives remain common, where improving one objective may degrade another. Unlike these methods, this paper focuses on a complementary yet underexplored question: whether different objectives should continuously participate in optimization throughout the entire generation process. Since diffusion generation exhibits clear stage-wise characteristics, different objectives may have different effective optimization windows. T-MOG does not further design objective-weight fusion strategies. Instead, it dynamically adjusts when each objective participates according to the generation stage and uses a cross-stage objective protection mechanism to prevent later objectives from damaging generative attributes established earlier. Therefore, T-MOG is complementary to existing multi-objective guidance methods and orthogonal in the optimization dimension.

A.3 IMAGE GENERATION DYNAMICS

Recently, attention has increasingly focused on the dynamic evolution of the image generation process, rather than only the quality of the final generated results (Sclocchi et al., 2025; Falck et al., 2025). Existing studies show that diffusion generation is not a simple uniform denoising process.

Table 2: Prompts used in the FLUX qualitative comparison.

ID	Prompt
1	Portrait of a woman, oil painting, chiaroscuro lighting, deep shadows, renaissance style, rich brushstrokes
2	A jazz musician playing saxophone, smoky bar, low key lighting, warm amber tones, candid moment
3	A violinist playing on a rooftop while fireflies gather around her, twilight, warm bokeh
4	A geisha holding a paper fan beside a koi pond, Japanese watercolor, delicate washes, cherry pink

Table 3: Prompts used in the SD3.5 qualitative comparison.

ID	Prompt
1	A Parisian cafe terrace at night, impressionist oil painting, thick brushstrokes, warm lamplight
2	A cat astronaut floating inside a spaceship window with Earth behind, soft cinematic light
3	A red panda sipping tea on a snowy balcony, steam rising, cozy winter evening, soft glow
4	A fox in a knitted scarf walking through a snowy village at night, warm windows, festive mood

Instead, it gradually restores different levels of visual information as sampling proceeds. In the early generation stage, the model mainly recovers low-frequency information such as global structure, object layout, and semantic relationships (Yan et al., 2026c). As the noise level decreases, the process further shifts toward constructing high-frequency information such as textures, edges, and local details. This coarse-to-fine generation pattern has been verified from multiple perspectives. Some studies analyze the diffusion process in the frequency domain and find clear differences in the recovery speed of different frequency components, with low-frequency structures usually forming before high-frequency details. Other studies further observe that semantic layout, spatial structure, and local appearance in images do not stabilize simultaneously, but are progressively established at different generation stages. These findings suggest that diffusion timesteps not only indicate changes in noise level, but also correspond to the dynamic formation of different visual attributes. However, although existing studies have recognized the stage-wise nature of image generation, most methods mainly use this property for sampling acceleration and do not further investigate the matching relationship between different optimization objectives and generation stages under the training-free guidance paradigm. Based on this observation, this paper explores how to adaptively schedule the timing of training-free guidance for different objectives according to image generation dynamics.

B More Results

B.1 Further Visual Results

Fig. 8 provides additional qualitative comparisons on SD3.5 across diverse subjects, scenes, and visual styles. Compared with the base model and existing guidance methods, T-MOG more consistently preserves prompt-specific content while maintaining coherent composition, lighting, and fine-grained appearance, showing that its balanced reward improvements translate into perceptually meaningful generation quality.

B.2 Intact Prompt Showcases

For completeness, we provide the intact prompts used in our qualitative comparisons across different backbones and objective pairs. The full prompt lists for FLUX, SD3.5, SigLIP–HPSv2, and TrOCR–PickScore are presented in Tables 2, 3, 4, and 5, respectively.

B.3 Hyperparameter

The full list of hyper-parameters in our paper is shown in Table 6.

Table 4: Prompts used in the SigLIP–HPS qualitative comparison.

ID	Prompt
1	Portrait of a violinist lost in music, candlelight, rich shadows, baroque oil painting
2	A golden retriever puppy running on a beach, splashing water, joyful, sunset glow
3	Rolling green hills under dramatic storm clouds, shafts of sunlight, oil painting
4	An enchanted library where books fly like birds, warm golden light, whimsical

Table 5: Prompts used in the TrOCR–PickScore qualitative comparison.

ID	Target	Prompt
1	MARKET	A colorful Moroccan spice market with pyramids of red and yellow spices, hanging lanterns, woven baskets and shoppers bargaining, with a flat painted wooden sign above the stalls displaying the exact text "MARKET" in large bold white uppercase letters on a deep blue board; warm dappled sunlight, rich saturated colors, travel photography, and no other readable text.
2	NOODLES	A bustling Asian night market street crowded with lantern-lit food stalls, steam rising from woks, a cook tossing noodles and a ginger cat sleeping on a crate, with a large flat wooden signboard above the main stall displaying the exact text "NOODLES" in large bold white uppercase letters; warm neon and lantern glow, shallow depth of field, cinematic street photography, and no other readable text.
3	SEA BREEZE	A breezy seaside café terrace with white umbrellas, couples sipping drinks, sailboats in the bay and bougainvillea on whitewashed walls, with a flat blue signboard above the terrace displaying the exact text "SEA BREEZE" in large bold white uppercase letters; bright Mediterranean sunlight, airy pastel palette, lifestyle travel photography, and no other readable text.
4	STADIUM	A packed football stadium at night with cheering fans waving scarves, players on the bright green pitch and fireworks overhead, with a large flat scoreboard panel displaying the exact text "STADIUM" in large bold white uppercase letters; floodlight glow, electric atmosphere, dramatic sports photography, and no other readable text.

Table 6: Hyper-parameters and metrics in our experiment. All other hyper-parameters follow the SD3.5 sampler or the baseline methods.

Name	Description	Value
T	denoising steps of the FlowMatch Euler sampler	20
w	classifier-free guidance weight	5.0
$\mathrm{NFE}_{\max}$	maximum number of function evaluations per image	29
mp	mixed precision	fp16
GPU	hardware	NVIDIA A800 80G
t_s	Reward switch stage	adaptive
r_t	Reward rounds per step （structural / textural）	1 / 3



Figure 8: Qualitative comparison between **T-MOG** and other methods.