Expert Systems With Applications 330 (2026) 133019

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Context-aware latent space mediation for inference-time unbiased semantic alignment in text-to-image models

Jili Chen [a,b], Huicheng Zeng [c], Changqin Huang [d,a,*], Qionghao Huang [a], Zilong Li [d,*], Xiaodi Huang [e]

[a] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, 321004, China
[b] School of Computer Science and Technology, Zhejiang Normal University, Jinhua, 321004, China
[c] Shanghai University of Finance and Economics Zhejiang College, Jinhua, 321015, China
[d] College of Education, Zhejiang University, Hangzhou, 310063, China
[e] School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, NSW, 2640, Australia

ARTICLE INFO

Keywords:
Text-to-image generation
Diffusion models
Inference-time optimization
Cross-attention
Semantic alignment

ABSTRACT

Text-to-image diffusion models have demonstrated remarkable capabilities in generating high-quality and diverse images. However, they remain vulnerable to semantic leakage arising from contextual bias, which often leads to misalignment between textual prompts and generated content. To address this issue, we propose **C**ontext-**A**ware **L**atent Space **M**ediation (CALM)—an inference-time optimization framework designed to enhance semantic alignment dynamically during inference. CALM identifies and mitigates contextual biases by comparing semantic similarities within and across semantic groups, adaptively reweighting contextual representations to reduce misalignment. To further resolve semantic conflicts in attention maps, we introduce a token–region alignment loss that minimizes rendering uncertainty in modifier relationships, thereby preserving consistency between objects and their attributes. Moreover, CALM employs supervised contrastive binding to optimize attention distributions across distinct semantic groups in the latent space, effectively enhancing inter-group distinction while preventing semantic entanglement and interference. Comprehensive experiments on five publicly available datasets demonstrate that CALM is architecture-agnostic, consistently improving both semantic alignment and image quality without retraining or fine-tuning. On GenEval and T2I-CompBench, CALM boosts the performance of multiple diffusion backbones by over 20%, confirming its effectiveness and generalizability. The code is available at https://github.com/TheShy-Dream/CALM-official.

1. Introduction

Diffusion models (Balaji et al., 2022; Esser et al., 2024; Podell et al., 2023; Ramesh et al., 2022; Rombach et al., 2022) have recently emerged as a powerful technique in text-to-image generation, leveraging their ability to progressively refine noise into meaningful images guided by free-form text prompts. Although they have made substantial progress, recent work (Chefer et al., 2023; Du et al., 2023) has highlighted their vulnerabilities, potentially resulting in catastrophic omission, semantic leakage, and other forms of semantic misalignment with the prompts (Agarwal et al., 2023; Bai et al., 2025; Zhang et al., 2024b).

Prior research has focused on enhancing semantic alignment between prompts and generated images, especially when prompts involve multiple distinct objects and intricate attribute-object relationships. Composable Diffusion (Liu et al., 2022) combines diffusion models to generate complex images from unseen concept combinations using operators. However, this often causes confusion or blending between multiple objects. Prompt-to-prompt (Hertz et al., 2022) inspires attention editing techniques in the latent space. Some works (Agarwal et al., 2023; Chefer et al., 2023; Guo et al., 2024) focus on the presence of objects and attributes. They improve the latent representation by introducing specific losses to control potential attention overlap, retention, and other conflicts. Other work (Feng et al., 2022; Li et al., 2023; Rassin et al., 2023; Zhang et al., 2024b) concentrates on improving the binding quality between attributes and objects. They prevent semantic leakage through techniques such as binding loss, energy-based latent optimization, and attention vector replacement.

Although prior work has improved semantic fidelity to prompts, these approaches overlook that such issues originate from biases in

* Corresponding authors.
E-mail addresses: irelia@zjnu.edu.cn (J. Chen), Z2015003@shufe-zj.edu.cn (H. Zeng), cqhuang@zju.edu.cn (C. Huang), qhhuang@m.scnu.edu.cn (Q. Huang), 0625921@zju.edu.cn (Z. Li), xhuang@csu.edu.au (X. Huang).

https://doi.org/10.1016/j.eswa.2026.133019
Received 24 January 2026; Received in revised form 16 March 2026; Accepted 25 May 2026
Available online 1 June 2026

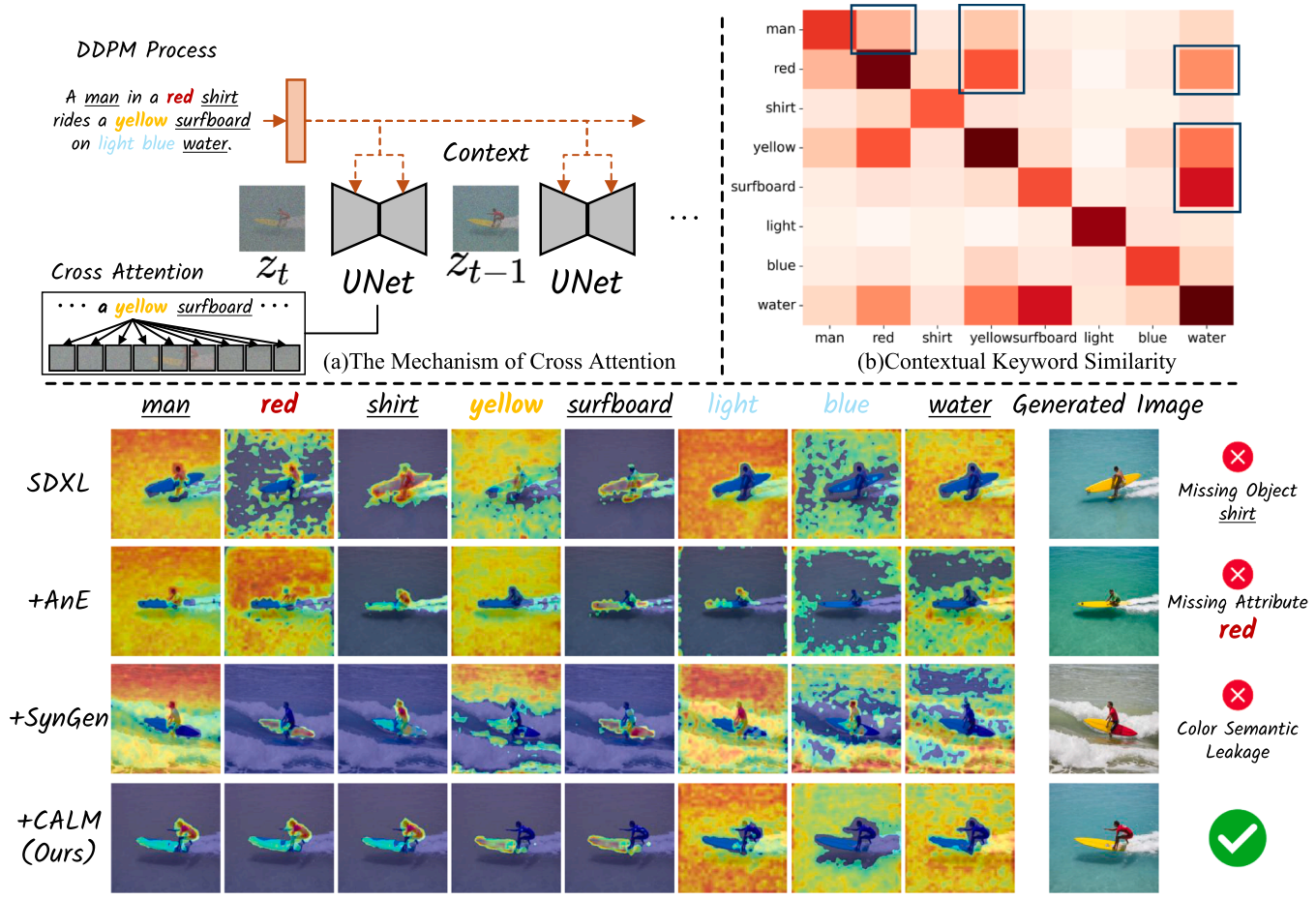


Fig. 1. (a) The cross attention mechanism between prompt tokens and patch tokens in the UNet during the reverse diffusion process. (b) In SDXL, the text encoder cannot effectively perceive modifier relationships, and pretrained biases cause large discrepancies in the contextual similarities of modified terms, resulting in unreasonably high similarities in certain contexts (highlighted using "blue boxes"). (c) Failure cases of the baselines and the attention visualizations of the keywords. The baselines suffer from issues such as object omission, attribute omission, and semantic leakage. CALM effectively addresses these problems.

the context encoder rather than the cross-attention mechanism itself (Chen et al., 2024), making their solutions merely palliative. As illustrated in Fig. 1(a), during the reverse diffusion process, each prompt token in the cross-attention modulecan influence all image tokens, potentially causing attention to leak to semantically irrelevant regions. Our analysis in Fig. 1(b) shows that the root problem lies in the text encoder: for example, we find that the token "yellow" is more similar to "red" and "man" than to the object it should modify, "surfboard". This bias is amplified during the reverse diffusion process in cross-attention, causing the rendered region for "yellow" to overlap with "red" and "man", which leads to semantic conflicts. We illustrate the impact of this contextual bias by visualizing the final-timestep cross-attention maps and their corresponding generated images using the SDXL 1.0 model (Podell et al., 2023) in Fig. 1(c). SDXL 1.0 omits objects such as a "shirt" (object omission). Attend-and-Excite (AnE) (Chefer et al., 2023) ensures the shirt appears but fails to render its color correctly (attribute omission). SynGen (Rassin et al., 2023) generates all specified objects and attributes but suffers from semantic leakage, where visual features of different attributes (e.g., colors) or objects become entangled or incorrectly influence one another. In addition, the overlap in attention can be observed to be closely related to the highlighted distributions in the context. For instance, "yellow", which is strongly associated with many keywords in the context, tends to be rendered across most of the attention regions. These examples directly reflect the limitations of prior work, which merely refines cross-attention without fully addressing contextual bias, failing to overcome the semantic consistency bottleneck.

To address the above issues, we propose **C**ontext-**A**ware **L**atent Space **M**ediation (CALM). CALM begins by constructing attribute-object semantic groups based on the syntactic parse tree. **1.** Then, we compute intra-group and inter-group similarities separately. Two dedicated augmentation mechanisms for objects and attributes are designed to ensure that objects and attributes within the same group are sufficiently similar in representation while being explicitly distinguishable from other semantic groups. **2.** Next, we extract cross-attention maps by associating object and attribute tokens with specific regions, maximizing their mutual alignment while minimizing uncertainty in region rendering, enabling objects and attributes to be simultaneously and clearly represented. **3.** Finally, to further reduce conflicts in the semantic space, CALM employs supervised contrastive learning by explicitly creating label mappings for the attention of each semantic group. This enhances the model's capacity to differentiate between distinct semantic pairs in the latent space, effectively decoupling the attention semantics of different object–attribute pairs. Our key contributions are summarized as follows:

- **Inference-time optimization framework:** We propose CALM, a novel inference-time optimization method for latent space mediation in Text-to-image models. CALM effectively mitigates semantic

misalignment and enhances the semantic consistency of the generated images.

- **Context-aware semantic binding:** To the best of our knowledge, this is the first work to identify and alleviate the influence of contextual biases on inference-time semantic binding in text-to-image generation. We introduce an adaptive context reweighting strategy and a supervised contrastive binding mechanism with a token–region semantic constraint loss, which jointly enhance semantic alignment and fidelity.
- **Comprehensive evaluation and generalization:** We evaluate CALM on five datasets featuring complex prompts and achieve state-of-the-art performance across mainstream text-to-image models, including both diffusion-based and flow-based architectures. The results demonstrate that CALM is architecture-agnostic, yielding consistent improvements across UNet and DiT backbones.

2. Related work

2.1. Text-to-image diffusion models

Text-to-image diffusion models aim to generate high-quality images based on textual descriptions, enabling the generation of diverse and realistic images that align with user-provided prompts. Milestone works like Stable Diffusion (Rombach et al., 2022), Imagen (Saharia et al., 2022), DALL·E 2 (Ramesh et al., 2022), ControlNet (Zhang et al., 2023) have significantly enhanced the capabilities of text-to-image generation, improving coherence between text and images (Esser et al., 2024; Peebles & Xie, 2023). However, recent research (Chefer et al., 2023; Guo et al., 2024; Li et al., 2023; Tang et al., 2024; Xie & Gong, 2025) has identified issues with semantic leakage in diffusion models, which can lead to missing objects or mismatched attributes in the generated images. Our approach effectively alleviates potential semantic leakage in prompts, enhancing the consistency between prompts and generated images without any fine-tuning.

2.2. Preference alignment for diffusion models

To better ensure that generative outputs conform to human expectations, a prevalent paradigm involves integrating external reward models to achieve preference alignment (Sun et al., 2025b). The principal works are categorized as follows: (1) Some methodologies perform parameter optimization via differentiable rewards (Clark et al., 2023; Guo et al., 2025; Prabhudesai et al., 2023; Wu et al., 2024; Xu et al., 2023), though they must overcome the computational overhead of differentiating through intricate backpropagation paths. (2) In a weighted SFT manner, RWR-based approaches reweight sampled trajectories to increase the likelihood of high-reward paths during training (Dong et al., 2023; Fan et al., 2025; Lee et al., 2023; Peters & Schaal, 2007). (3) By framing the diffusion trajectory as a dynamic Markov process, some work utilizes Proximal Policy Optimization (PPO) (Schulman et al., 2017) to maximize the expected rewards (Black et al., 2023; Chen & Kuo, 2025; Fan et al., 2023; Hu et al., 2025; Zhang et al., 2024a). (4) To alleviate the dependency on explicit reward models, some works learn directly from preference pairs via Direct Preference Optimization (DPO). These approaches align the model by maximizing the likelihood of preferred samples while concurrently suppressing the probability of non-preferred ones (Liang et al., 2025; Rafailov et al., 2023; Shen et al., 2025; Sun et al., 2025a,c,d,e; Wallace et al., 2024; Yang et al., 2024; Zhu et al., 2025). (5) GRPO-based methods utilize group-relative reward normalization as a baseline, thereby eliminating the dependency on an explicit value function. This architectural simplification facilitates a memory-efficient reinforcement learning paradigm (He et al., 2025c; Liu et al., 2025a; Luo et al., 2025a,b; Wang et al., 2025a,b; Xue et al., 2025). RL-based methods significantly improve T2I fidelity but suffer from heavy computational costs and reward hacking (He et al., 2025a; Zhai et al., 2025). Inspired by these works, we present a training-free method for alignment.

Table 1
Comparison of inference-time enhancement methods. Unlike prior works that focus on specific misalignment categories, **CALM** provides a holistic solution for object/attribute preservation and binding.

Method	Problems Addressed		
	Obj. Omis.	Attr. Omis.	Mis. Bind.
AnE (Chefer et al., 2023)	✓	×	×
SynGen (Rassin et al., 2023)	△	×	△
A-star (Agarwal et al., 2023)	×	×	✓
EBAMA (Zhang et al., 2024b)	×	△	✓
INITNO (Guo et al., 2024)	✓	×	△
CALM (Ours)	✓	✓	✓

* ✓: Fully Addressed; △: Partially Addressed; ×: Generally Overlooked.

2.3. Attention-based inference-time alignment

Prompt-to-prompt (Hertz et al., 2022) reveals the mechanism of cross-attention between text and images, providing a foundation for attention-based control in image generation. Through monitoring attention activations, Attend-and-Excite (Chefer et al., 2023) enhances subject token activations in cross-attention maps to ensure precise focus on key objects. Building on the shift from objects to modifiers, SynGen (Rassin et al., 2023) leverages syntactic analysis to encourage overlap between modifiers and their corresponding objects' attention maps, while simultaneously reducing interference from unrelated words. To explicitly address attention overlap and decay, A-star (Agarwal et al., 2023) introduces attention segregation and retention losses. Li et al. (2023) further mitigates improper object attention and attribute misbinding through a combination of total variation–based attention loss and Jensen-Shannon divergence–based binding loss. Building on this idea, EBAMA (Zhang et al., 2024b) optimizes object-centric attribute binding by maximizing the likelihood of an energy-based model, balancing attribute binding with an intensity regularizer. INITNO (Guo et al., 2024) partitions the initial latent space based on cross-attention responses and self-attention conflict scores, improving latent space optimization. As summarized in Table 1, while existing attention-based methods have made significant improvements, they primarily focus on posterior refinement of attention maps, treating the symptoms rather than the cause of semantic misalignment. In contrast, CALM introduces a paradigm shift by identifying contextual bias in the text encoder as the root of the problem. By integrating context-aware augmentation and supervised contrastive binding, CALM goes beyond simple attention reweighting. It is the first framework to achieve robust, architecture-agnostic semantic alignment that scales effectively from traditional UNet-based models to modern DiT and flow-matching architectures.

3. Preliminaries

3.1. Latent diffusion models

Latent Diffusion Models (LDMs) generate images by operating in the latent space of a pre-trained autoencoder (Kingma et al., 2013). Specifically, an encoder E first compresses an image x into a latent representation $z = E(x)$, and a decoder D reconstructs the image from this latent: $D(E(x)) \approx x$. Over the latent space, a Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) is trained to progressively denoise a latent variable z_t at each timestep t. The model can be conditioned on external inputs, e.g., a text prompt y encoded by a CLIP text encoder c (Radford et al., 2021). The denoising network ϵ_θ, typically a UNet or Transformer with self- and cross-attention layers, is trained to predict

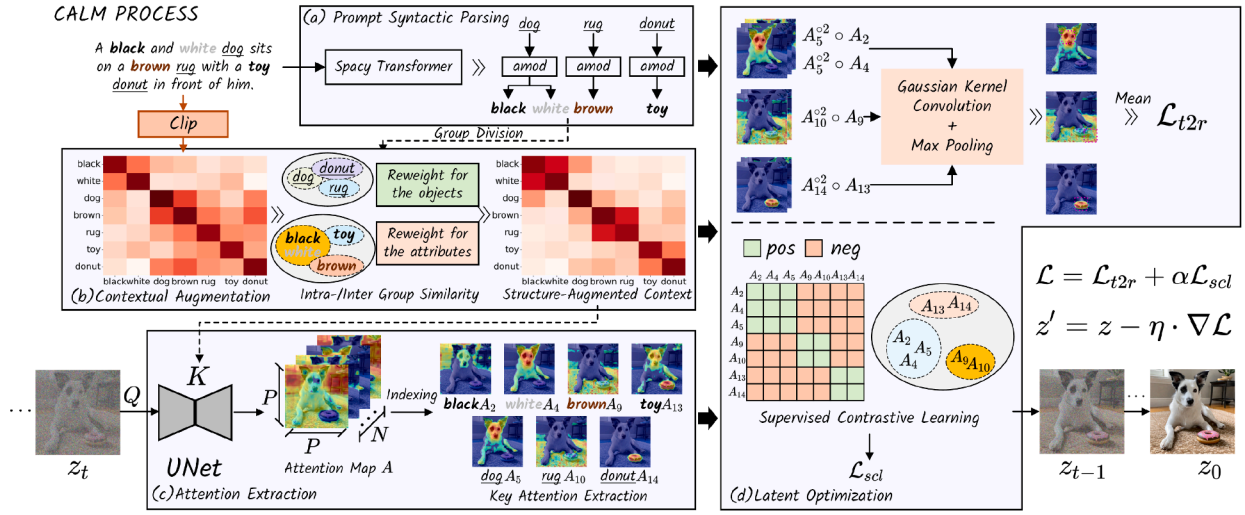


Fig. 2. Overview of our proposed CALM. The CALM framework begins by parsing a sentence into a semantic-dependency tree, from which the dependencies are organized into semantic groups. The token index corresponding to each object and modifier guides the model in reinforcing contextual representations and attention. The context-aware attention maps are used to refine the latent representation through $\mathcal{L}_{t2r}$ and $\mathcal{L}_{scl}$.

the noise ϵ using the objective:

$$\mathcal{L} = \mathbb{E}_{z\sim E(x),y,\varepsilon\sim\mathcal{N}(0,1),t}[\|\varepsilon - \varepsilon_\theta(z_t, t, c(y))\|_2^2], \tag{1}$$

at inference, a sample $z_t \sim \mathcal{N}(0,1)$ is progressively denoised to z_0, then decoded to an image $x' = D(z_0)$.

3.2. Attention mechanism in diffusion models

In latent diffusion models, cross-attention is crucial for aligning text and image features during generation. First, the text prompt $y \in \mathbb{R}^N$ of length N is encoded into a sequence of embeddings using a pre-trained encoder c, forming a condition $c(y)$. Then, the projected condition serves as the K and V, while the intermediate feature of the UNet acts as the Q to perform the cross-attention computation:

$$A = \mathrm{softmax}(\frac{QK^T}{\sqrt{d}}). \tag{2}$$

Let P denote the spatial resolution per side of the intermediate attention map. The attention map $A \in \mathbb{R}^{P\times P\times N}$ characterizes the similarities and the degree of alignment between the prompt tokens and the spatial locations of the feature map. Specifically, $A[i,j,n]$ quantifies the similarity between the nth token and the (i,j)th spatial patch token. Note that each value in A is bounded between 0 and 1.

4. CALM

CALM is dedicated to optimizing the latent representation in the early stages of the reverse diffusion process to prevent object omission, attribute omission, and semantic leakage. As shown in Fig. 2, the CALM pipeline consists of four components: prompt syntactic parsing, contextual augmentation, attention extraction, and latent optimization.

4.1. Prompt syntactic parsing

Following the preprocessing approach in Rassin et al. (2023), we employ spaCy's transformer-based dependency parser (Honnibal & Johnson, 2015) to analyze the syntactic structure of the prompt. We first identify all objects that are not syntactically dependent on other objects and treat them as object candidates. For each object, we then collect its associated modifiers by following specific dependency relations such as adjectival modifiers, nominal compounds, adverbial modifiers, complements, and coordinated modifiers. This procedure produces structured semantic groups that represent the prompt's compositional semantics.

4.2. Contextual augmentation

The primary source of semantic conflicts lies in contextual bias within the text encoder, where attributes often exhibit stronger similarity to unrelated entities than to the objects they are intended to modify. This weak attribute–object association is subsequently amplified during the diffusion process, leading to semantic misalignment.

To address this issue at its root, we propose Contextual Augmentation, a mechanism that explicitly strengthens the association between objects and their corresponding attributes within the text encoder. The key idea is to rescale the hidden representations based on intra-group and inter-group semantic similarities. We denote the context input generated by CLIP (Radford et al., 2021) as $H \in \mathbb{R}^{L\times d}$, where L is the sequence length and d is the hidden dimension. We first compute the pairwise similarity matrix $S = \mathrm{softmax}(HH^\top) \in \mathbb{R}^{L\times L}$. Let $\mathcal{N}$ denote the set of object indices. For each object $n_i \in \mathcal{N}$, let $\mathcal{M}_i$ denote the set of associated modifier indices. Define the union of all modifier indices as $\mathcal{M} = \bigcup_{n_i\in\mathcal{N}} \mathcal{M}_i$. The union of all these indices is denoted as $\mathcal{I} = \mathcal{N} \cup \mathcal{M}$.

4.2.1. Object augmentation

Since each semantic group contains only one object, we define the intra-group semantic similarity of an object as the similarity between the object and itself, and the inter-group semantic similarity as the aggregated similarity between the object and other objects. For each object $n \in \mathcal{N}$, we compute the semantic strength using:

$$r_n = \frac{S[n,n]}{\sum_{k\in\mathcal{N}} S[n,k] + \epsilon}, \tag{3}$$

where the numerator represents the intra-group semantic similarity, and the denominator denotes the combination of intra-group and inter-group semantic similarity (i.e., the aggregation of similarities with all objects). ϵ is a small constant for numerical stability. If r_n falls below an object-specific threshold τ_o, the representation of object n is rescaled:

$$\alpha_n = 1 + \tanh(\tau_o - r_n), \tag{4}$$

where an object token with lower semantic strength tends to be assigned a larger α_n, enhancing the self-coherence of its embedding and preserving its semantic prominence within the overall context. Notably, while the semantic strength $r_n \in [0,1]$ is inherently bounded, we employ the tanh function as a safeguard against distributional shifts in extreme scenarios (e.g., when using a large τ_o or encountering a near-zero r_n). Previous studies (Ho & Salimans, 2022; Kirstain et al., 2023; Saharia et al., 2022) have demonstrated that excessively strong text

guidance can drive latent representations off the manifold, resulting in over-saturation and visual artifacts. By leveraging the properties that $\tanh(x) < x$ and $\tanh(x) < 1$ for $x > 0$, the rescaling factor α_n is strictly constrained such that $\alpha_n = 1 + \tanh(\tau_o - r_n) < 1 + \tau_o - r_n$ within the valid interval $r_n \in [0, 1]$. Moreover, this ensures that the α_n is always bounded by 2. This mechanism effectively clamps the over-amplification while maintaining monotonicity, even for severely suppressed objects, the recalibrated embeddings remain within a safe region of the latent space.

4.2.2. Attribute augmentation

Given that each semantic group may include multiple attributes, the intra-group semantic similarity for an attribute is defined as the aggregated similarity between the current attribute and all items in the same semantic group, whereas the inter-group semantic similarity is defined as the aggregated similarity between the current attribute and items in all semantic groups. For each attribute $m \in \mathcal{M}_i$ associated with the object n_i, we compute the semantic strength using:

$$r_m = \frac{\sum_{j \in \mathcal{M}_i \cup \{n_i\}} S[m, j]}{\sum_{k \in \mathcal{I}} S[m, k] + \epsilon}, \tag{5}$$

where the numerator represents the aggregated intra-group semantic similarities, and the denominator represents the combination of intra-group and inter-group semantic similarity (i.e., the aggregation of similarities with all objects and all attributes). If r_m falls below an attribute-specific threshold τ_a, the hidden representation of token m is rescaled by:

$$\alpha_m = 1 + \tanh(\tau_a - r_m), \tag{6}$$

attribute tokens with lower semantic strengths are enhanced with a larger α_m, ensuring that attributes remain semantically anchored to intended objects rather than drifting toward unrelated tokens. The design of the tanh function here is consistent with the rationale in object augmentation to ensure stable rescaling and prevent potential distributional shifts.

Finally, we aggregate all rescaling factors (α_n and α_m) into a vector $\alpha \in \mathbb{R}^L$. The contexts H are updated by elementwise multiplication $H' = H \odot \alpha$ where $\odot$ denotes broadcasting along the hidden dimension. By amplifying underrepresented attributes and objects, contextual augmentation provides a clearer and more balanced semantic signal to the diffusion model. Therefore, it mitigates attribution drift and reduces entanglement between different objects and attributes. Subsequently, the contexts H' interact with the latent representations through the attention mechanism.

4.3. Attention extraction

Based on the object-attribute indices obtained from the syntactic parsing stage, we extract the corresponding cross-attention scores from the augmented attention map A at the specified resolution. In Stable Diffusion, the spatial resolutions of intermediate features are {64, 32, 16, 8}. Following Chefer et al. (2023), Guo et al. (2024), we collect attention values at the 16×16 resolution, as it contains rich semantic information (Hertz et al., 2022). Each object and its associated attributes are grouped as a list of token indices. For a given object-attribute group, we collect the attention scores $A[:, :, n]$ for each token index n in the group (denoted as A_n). More details of attention extraction are provided in Section 5.1.4.

4.4. Latent optimization

To mitigate the omission of objects and attributes, as well as semantic leakage, we introduce two loss functions to jointly optimize the latent representation: token-region alignment loss and supervised contrastive binding loss.

4.4.1. Token-region alignment loss

To concentrate attention on dominant regions and reduce the spatial uncertainty of each object, we first square its attention map, $\tilde{A}_{n_i} = A_{n_i}^{\circ 2}$. This operation suppresses weak activations and sharpens the focus, analogous to increasing purity in a Gini coefficient. To capture local co-activation with attributes, each modifier attention map $A_m : (m \in \mathcal{M}_i)$ is combined with the squared object map via element-wise multiplication and smoothed with a Gaussian kernel K:

$$s_i = \max\left(\frac{1}{|\mathcal{M}_i|} \sum_{m \in \mathcal{M}_i} (\tilde{A}_{n_i} \odot A_m) * K \right), \tag{7}$$

the multiplication effectively weights modifier activations by the object prominence, so that regions with stronger object presence contribute more, further concentrating attention and reducing uncertainty in the attention distribution. The overall token-region loss is then defined as

$$\mathcal{L}_{t2r} = \frac{1}{|\mathcal{N}|} \sum_{n_i \in \mathcal{N}} (1 - s_i), \tag{8}$$

which encourages strong spatial alignment between objects and their attributions by maximizing the peak coactivation scores.

4.4.2. Supervised contrastive binding loss

To further enhance the discriminability of different semantic groups in the latent space, we introduce a supervised contrastive binding loss. For each object or attribute token, we extract its cross-attention distribution as a latent embedding and normalize it to obtain a set of ℓ_2-normalized attention vectors $\{z_k\}_{k=1}^K$. Within each semantic group, the attention distributions of an object and its associated attributes share the same label y_k, whereas attention distributions from different semantic groups are assigned different labels. Following Khosla et al. (2020), for each anchor z_i, we treat all other tokens with the same label as positives, and the rest as negatives. The loss pulls positive pairs closer and pushes negative pairs apart in the latent embedding space:

$$\mathcal{L}_{scl} = -\frac{1}{K} \sum_{i=1}^{K} \log \frac{\sum_{j=1, j \neq i}^{K} \mathbb{I}[y_i = y_j] \cdot \exp\left(\frac{z_i \cdot z_j}{\tau}\right)}{\sum_{j=1, j \neq i}^{K} \exp\left(\frac{z_i \cdot z_j}{\tau}\right)}, \tag{9}$$

where τ is a temperature hyperparameter, K represents the total count of objects and attributes in the current prompt (i.e., $|\mathcal{I}|$), and $\mathbb{I}[y_i = y_j]$ is an indicator function that masks out negative units. By encouraging tokens within the same semantic group to exhibit similar attention distributions while separating them from unrelated tokens, a more structured latent space for prompt-grounded generation can be learned. This effectively disentangles attention representations and enhances object-attribute semantic bindings.

4.4.3. Loss function

To jointly optimize the latent representation, we combine the token-region alignment loss and the supervised contrastive binding loss into a unified objective. The hyperparameter β is used to control the weight of $\mathcal{L}_{scl}$:

$$\mathcal{L} = \mathcal{L}_{t2r} + \beta \mathcal{L}_{scl}, \tag{10}$$

the latent is then updated via gradient descent as:

$$z_t' \leftarrow z_t - \eta_t \nabla_{z_t} \mathcal{L}, \tag{11}$$

where η_t is the update step size at timestep t. This unified optimization encourages complete object rendering, accurate attribute localization, and accurate semantic bindings in the latent space. The algorithmic pipeline of CALM is illustrated in Algorithm 1. Following Zhang et al. (2024b), it operates in the first half of the reverse diffusion process, requiring only a single latent optimization at each iteration.

Algorithm 1 Context-aware latent space mediation (CALM).

Input: Text prompt y; pretrained T2I model SD; inference steps T; image decoder D; text encoder with a time embedding mapping network E_t; update step size η, the trade-off hyperparameter β.
Output: Image x aligned with prompt y.
Initialize $z_T \sim \mathcal{N}(0,1)$
Parse y to extract object tokens N and attribute tokens $\{M(n)\}_{n\in N}$
for $t = T$ to $\lfloor T/2 \rfloor + 1$ **do**
  $H \leftarrow E_t(t)$
  $\alpha_n, \alpha_m \leftarrow \text{Compute}(H, S, \{M(n)\}_{n\in N})$ // see Eqs. (4) and (6)
  $H' = \text{Augment}(H, \alpha_n, \alpha_m)$
  $\_, A \leftarrow \text{SD}(z_t, t, H')$ // Extract attention maps
  $\mathcal{L}_{f2r}, \mathcal{L}_{scl} \leftarrow \text{Compute}(A, N, \{M(n)\}_{n\in N})$ // see Eqs. (8) and (9)
  $z_t' \leftarrow z_t - \eta \nabla_{z_t} \mathcal{L}$
  $z_{t-1}, \_ \leftarrow \text{SD}(z_t', t, y)$
end for
for $t = \lfloor T/2 \rfloor$ to 1 **do**
  $z_{t-1}, \_ \leftarrow \text{SD}(z_t, t, y)$
end for
$x \leftarrow D(z_0)$
return x

5. Experiments

5.1. Experimental settings

The backbone models include Stable Diffusion v1.4, SDXL 1.0 Base, SD3.5 Medium, and Flux.1 Dev. We compare our method with SD (Rombach et al., 2022) and several inference-time enhancement methods, including AnE (Chefer et al., 2023), SynGen (Rassin et al., 2023), EBAMA (Zhang et al., 2024b), CONFORM (Meral et al., 2024), and T-SAM (Kim et al., 2025) across three datasets: **AnE** (Chefer et al., 2023), **ABC-6K** (Feng et al., 2022), and **DVMP** (Rassin et al., 2023). In addition, we compare CALM with mainstream text-to-image generative models (PixArt-α (Chen et al., 2023), DALLE-2 (Ramesh et al., 2022), DALLE-3 (Betker et al., 2023), Emu (Wang et al., 2024) , Show-o (Xie et al., 2024), Janus (Wu et al., 2025), JanusFlow (Ma et al., 2025b), Dream Engine (Chen et al., 2025), DiT-ST Medium (Zhang et al., 2025), MARS (He et al., 2025b), and LLM4GEN (Liu et al., 2025b)) on two benchmark datasets (**GenEval** (Ghosh et al., 2023) and **T2I-CompBench** (Huang et al., 2025, 2023)).

5.1.1. Datasets

AnE (Chefer et al., 2023) comprises 66 Animal-Animal pairs and 66 Object-Object pairs, and 144 Animal-Object pairs. The **ABC-6K** (Feng et al., 2022) consists of natural compositional prompts from MSCOCO with at least two color words modifying different objects. The **DVMP** (Rassin et al., 2023) aims to challenge models by using prompts with numerous and uncommon attributes. We randomly sample 600 prompts from DVMP and ABC-6K to conduct a representative and efficient evaluation (Rassin et al., 2023). **GenEval** (Ghosh et al., 2023) contains 553 samples and assesses properties such as object co-occurrence, position, count, and color through fine-grained object-level analysis. **T2I-CompBench** (Huang et al., 2025, 2023) has 8000 samples, from which we use the attribute-binding (color, shape, texture) subset to evaluate compositional capabilities in text-to-image models.

5.1.2. Evaluation metrics

For each prompt, 10 images are generated using the same seed group. Following Chefer et al. (2023), we evaluate our method using three text-alignment metrics: **Full**, measuring similarity between the generated image and the entire text prompt in CLIP space (Radford et al., 2021); **Min**, the average CLIP similarity of the least well-represented object; and **T-C**, the average similarity between the original prompt and BLIP-generated captions (Li et al., 2022). For dataset-specific evaluation, GenEval relies on a pre-trained object detector to assess object existence, attributes, spatial relations, and counts, while T2I-CompBench uses BLIP-based VQA to evaluate attribute binding and object relationships. We also measure overall image quality using ImageReward (IR) (Xu et al., 2023), Aesthetic Score (AES) (Murray et al., 2012), HPSv2 (Wu et al., 2023), and PickScore (Kirstain et al., 2023), which collectively reflect human preference and perceptual quality.

5.1.3. Hyperparameter configuration

Following Chefer et al. (2023), Stable Diffusion v1.4 is used as the base text-to-image diffusion model. Experiments are also conducted using SDXL 1.0-BASE, SD3.5-Medium, and Flux.1 Dev to demonstrate that the method generalizes to other UNet-based diffusion models as well as DiT-based flow matching models. The token dependency relations are extracted using spaCy's *en_core_web_trf* model. Furthermore, we employ the DDIM sampler (Song et al., 2020) for the denoising process in the SD1.4 and SDXL 1.0 backbones with the number of sampling steps fixed at 50 and a classifier-free guidance scale (CFG scale) of 7.5. For the SD3.5 and the Flux.1 Dev model, we utilize the ODE Sampler, setting the classifier-free guidance scale to 4.5. We perform latent optimization during the first $T/2$ steps (i.e., the first 25 timesteps) of the denoising process.

5.1.4. Details of attention extraction

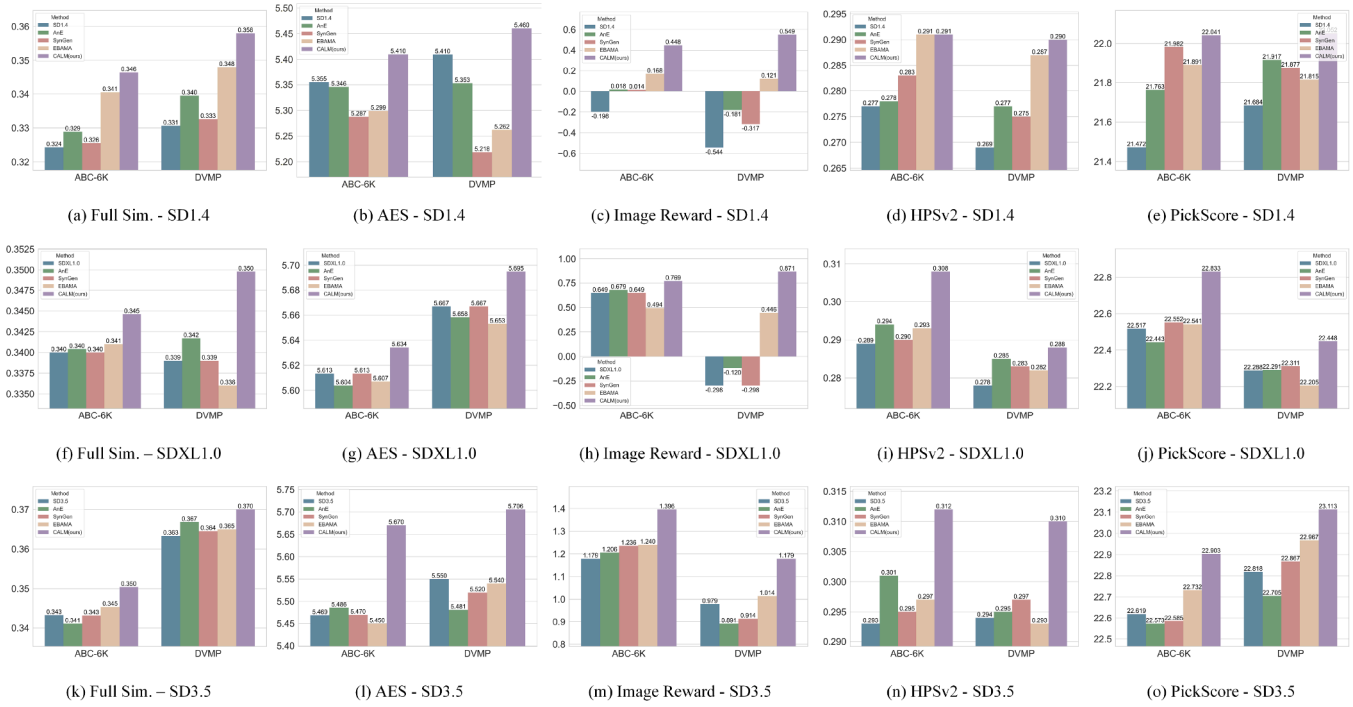
The aggregated attention features A consist of N spatial attention maps, each corresponding to a token in the input prompt y. The CLIP text encoder adds a special start-of-text token $\langle SOT \rangle$ at the beginning of y to mark the start. It has been found that in Stable Diffusion, such $\langle SOT \rangle$ token consistently attracts the highest attention among all tokens. Following the approach in Chefer et al. (2023), we remove the attention assigned to $\langle SOT \rangle$ and then apply a softmax function over the remaining tokens to compute the normalized attention scores. For SDXL, the spatial resolutions are {128, 64, 32}, and we utilize attention values at the 32×32 resolution (Tian et al., 2025). For SD3.5 (Esser et al., 2024), which employs two text encoders (CLIP and T5) within the MM-DiT architecture, the attention scores between prompt tokens and spatial locations are computed via self-attention rather than cross-attention, inherently capturing the correspondence between text tokens and spatial feature representations. To balance performance with computational and memory costs during inference, attention values are extracted from the last five MM-DiT blocks containing contextually encoded hidden states. These values are indexed separately according to the maximum token length of each text encoder. The resulting attention map has a shape of $(4429, 4429)$, where image representations occupy the first 4096 positions, followed by text representations. Accordingly, we extract the matrix block corresponding to rows $[:4096, 4096:]$, yielding a $(4096, 333)$ submatrix. The first 77 columns correspond to CLIP text embeddings, and the remaining 256 columns correspond to T5 text embeddings. These two parts are indexed and treated as attention, while all subsequent operations follow the same procedure as in cross-attention. For Flux.1 Dev, attention is extracted from the last five single-stream blocks where text and image tokens are processed jointly. Unlike SD3.5, Flux.1 Dev utilizes only the T5 encoder for token-level representation. In the $(4608, 4608)$ joint attention matrix, we slice the $(4096, 512)$ submatrix from rows $[512:]$ and columns $[:512]$, where 512 corresponds to the default text sequence length of the Flux T5 encoder. Transposing this submatrix yields the final cross-modal alignment mapping each text token to its corresponding spatial latents. All attention visualizations and extractions in this paper are based on averaged results across all attention heads.

5.2. Quantitative result

Table 2 and Fig. 3 provide a comprehensive summary of the quantitative comparison results across the AnE, DVMP, and ABC-6K datasets.

Table 2
Comparison results across different methods on the AnE dataset based on SD 1.4, SDXL 1.0, and SD 3.5. Bold and underlined values indicate the best and second-best performances.

Base Model	Method	Animal-Animal			Animal-Object			Object-Object		
		Full/Min/T-C	IR	AES	Full/Min/T-C	IR	AES	Full/Min/T-C	IR	AES
SD 1.4	SD	0.311/0.213/0.767	−0.237	5.384	0.340/0.246/0.793	0.098	5.398	0.335/0.235/0.765	−0.661	5.151
	AnE	0.332/0.248/0.806	1.110	5.459	0.353/0.265/0.830	1.352	5.447	0.360/0.270/0.811	0.788	5.211
	SynGen	0.311/0.213/0.767	−0.237	5.384	0.355/0.264/0.830	1.527	5.525	0.355/0.262/0.811	0.662	5.170
	EBAMA	0.340/0.256/0.817	0.628	5.474	0.362/0.270/0.851	1.312	5.479	0.366/0.274/0.836	0.725	5.232
	CONFORM	0.340/0.248/0.820	–	–	0.360/0.269/0.850	–	–	0.360/0.268/0.820	–	–
	T-SAM	–	–	–	0.370/**0.280**/0.850	–	–	0.370/**0.280**/0.810	–	–
	CALM(ours)	**0.348/0.262/0.822**	**1.326**	**5.636**	**0.371**/0.276/**0.856**	**1.689**	**5.566**	**0.375**/0.279/**0.848**	**1.538**	**5.291**
SDXL 1.0	SDXL	0.309/0.219/0.778	0.395	5.625	0.349/0.256/0.815	1.274	5.564	0.350/0.259/0.784	0.402	5.342
	AnE	0.311/0.217/0.784	0.578	5.763	0.350/0.257/0.829	1.411	5.756	0.355/0.261/0.800	0.761	5.476
	SynGen	0.309/0.219/0.778	0.395	5.625	0.349/0.257/0.816	1.353	5.596	0.352/0.258/0.785	0.416	5.513
	EBAMA	0.310/0.218/0.785	0.543	5.893	0.350/0.256/0.827	1.242	5.858	0.357/0.260/0.798	0.459	5.514
	CALM(ours)	**0.317/0.224/0.797**	**1.448**	**5.988**	**0.356/0.260/0.836**	**1.499**	**5.921**	**0.363/0.268/0.811**	**0.967**	**5.530**
SD 3.5	SD3.5	0.345/0.252/0.830	1.543	5.691	0.361/0.264/0.845	1.609	5.564	0.368/0.270/0.862	1.623	5.087
	AnE	0.339/0.254/0.837	1.556	5.703	0.359/0.263/0.844	1.658	5.668	0.369/0.271/0.855	1.722	5.197
	SynGen	0.345/0.252/0.830	1.543	5.691	0.362/0.265/0.849	1.671	5.577	0.365/0.267/0.849	1.691	5.170
	EBAMA	0.346/0.252/0.835	1.616	5.712	0.362/0.264/0.847	1.643	5.664	0.365/0.267/0.849	1.600	0.520
	CALM(ours)	**0.354/0.261/0.864**	**1.707**	**5.897**	**0.368/0.272/0.869**	**1.826**	**5.812**	**0.378/0.282/0.875**	**1.841**	**5.263**



(a) Full Sim. - SD1.4 (b) AES - SD1.4 (c) Image Reward - SD1.4 (d) HPSv2 - SD1.4 (e) PickScore - SD1.4

(f) Full Sim. – SDXL1.0 (g) AES - SDXL1.0 (h) Image Reward - SDXL1.0 (i) HPSv2 - SDXL1.0 (j) PickScore - SDXL1.0

(k) Full Sim. – SD3.5 (l) AES - SD3.5 (m) Image Reward - SD3.5 (n) HPSv2 - SD3.5 (o) PickScore - SD3.5

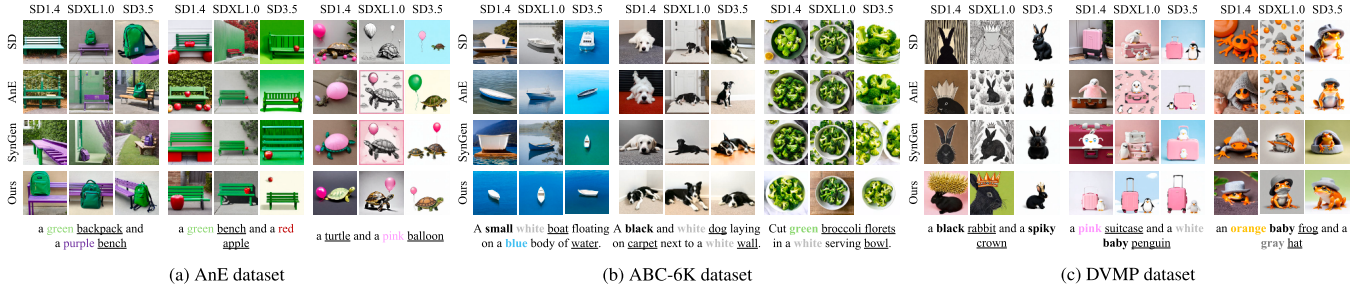
Fig. 3. Quantitative comparison of CALM against baselines using **SD1.4** (top), **SDXL 1.0** (middle), and **SD3.5** (bottom) as backbones on ABC-6K and DVMP datasets. Each columns represent specific metrics (from left to right: Full Sim., AES, ImageReward, HPSv2, and PickScore).

Traditional methods often encounter limitations such as object or attribute omission and attribute leakage because they struggle to handle complex mappings between attributes and objects. Such issues lead to weaker alignment between the generated images and the input prompts. CALM effectively overcomes these challenges by enforcing precise token and region alignment, which results in superior performance across all benchmarks. A significant observation in Fig. 3 is that while existing methods show improvements on SD1.4 but fail to scale to larger backbones like SDXL or the DiT-based SD3.5, CALM remains highly effective across these advanced architectures. This adaptability aligns with previous findings suggesting that smaller models like SD1.4 can match or even exceed the performance of larger models when paired with appropriate inference time scaling techniques (Ma et al., 2025a). Furthermore, the overall results demonstrate that CALM manages the trade-off between semantic binding and visual fidelity more effectively than its counterparts. While other methods often improve prompt alignment at the cost of image quality, CALM ensures a steady enhancement in both semantic consistency and visual realism. This indicates that our approach achieves a more balanced optimization that strengthens textual alignment while simultaneously promoting higher image quality across diverse diffusion architectures.

Moreover, as shown in Table 3, our results on the GenEval and T2I-CompBench datasets demonstrate that while many existing models perform well in single-object rendering, their accuracy drops sharply when faced with tasks involving multiple objects, multiple colors, specified object counts, multi-object semantic binding, or spatial constraints. Beyond strengthening semantic binding, CALM also improves related abilities such as accurately rendering object count, position, size, and texture. Across various backbone architectures, CALM consistently improves performance metrics, achieving nearly double the scores in ob-

Table 3
Results across GenEval and T2I-CompBench. Bold and underlined values denote the best and second-best performances.

Model	GenEval							T2I-CompBench		
	Mean	Single	Two	Counting	Colors	Position	Color Attribution	Color	Shape	Texture
PixArt-α	0.48	0.98	0.50	0.44	0.80	0.08	0.07	0.69	0.56	0.70
DALL-E 2	0.52	0.94	0.66	0.49	0.77	0.10	0.19	0.57	0.55	0.64
DALL-E 3	0.67	0.96	0.87	0.47	0.83	0.43	0.45	0.81	**0.68**	0.81
Emu	0.54	0.98	0.71	0.34	0.81	0.17	0.21	0.79	0.58	0.74
Show-o	0.53	0.95	0.52	0.49	0.82	0.11	0.28	–	–	–
Janus	0.61	0.97	0.68	0.30	0.84	0.46	0.42	–	–	–
JanusFlow	0.63	0.97	0.59	0.45	0.83	**0.53**	0.42	–	–	–
Dream Engine	0.69	**1.00**	**0.94**	0.64	0.81	0.27	0.49	–	–	–
DiT-ST Medium	0.69	0.99	0.88	0.70	0.75	0.36	0.46	–	–	–
MARS	–	–	–	–	–	–	–	0.69	0.54	0.71
LLM4GEN$_{SD1.5}$	–	–	–	–	–	–	–	0.47	0.45	0.55
LLM4GEN$_{SDXL1.0}$	–	–	–	–	–	–	–	0.77	0.59	0.71
SD 1.4	0.38	0.95	0.35	0.24	0.64	0.03	0.05	0.38	0.36	0.42
SD 1.4$_{+CALM}$	0.56	0.99	0.71	0.46	0.87	0.16	0.14	0.60	0.47	0.61
SDXL 1.0	0.55	0.98	0.74	0.39	0.85	0.15	0.23	0.59	0.47	0.53
SDXL 1.0$_{+CALM}$	0.64	0.99	0.75	0.59	**0.94**	0.23	0.31	0.77	0.64	0.67
SD 3.5	0.63	0.98	0.78	0.50	0.81	0.24	0.52	0.80	0.57	0.73
SD 3.5$_{+CALM}$	**0.80**	0.99	0.91	0.88	0.85	0.42	**0.72**	**0.86**	**0.68**	0.80
Flux.1 Dev	0.66	0.98	0.81	0.74	0.79	0.22	0.45	0.74	0.57	0.69
Flux.1 Dev$_{+CALM}$	0.78	0.99	0.92	**0.90**	0.82	0.35	0.69	0.85	0.66	**0.83**



(a) AnE dataset (b) ABC-6K dataset (c) DVMP dataset

Fig. 4. Qualitative comparison results on AnE, ABC-6K, and DVMP datasets, with each column sharing the same random seed.

ject counting and spatial localization tasks, with an overall average improvement of around 20% for each backbone. Paired with CALM, text-to-image models can match and even surpass current state-of-the-art approaches.

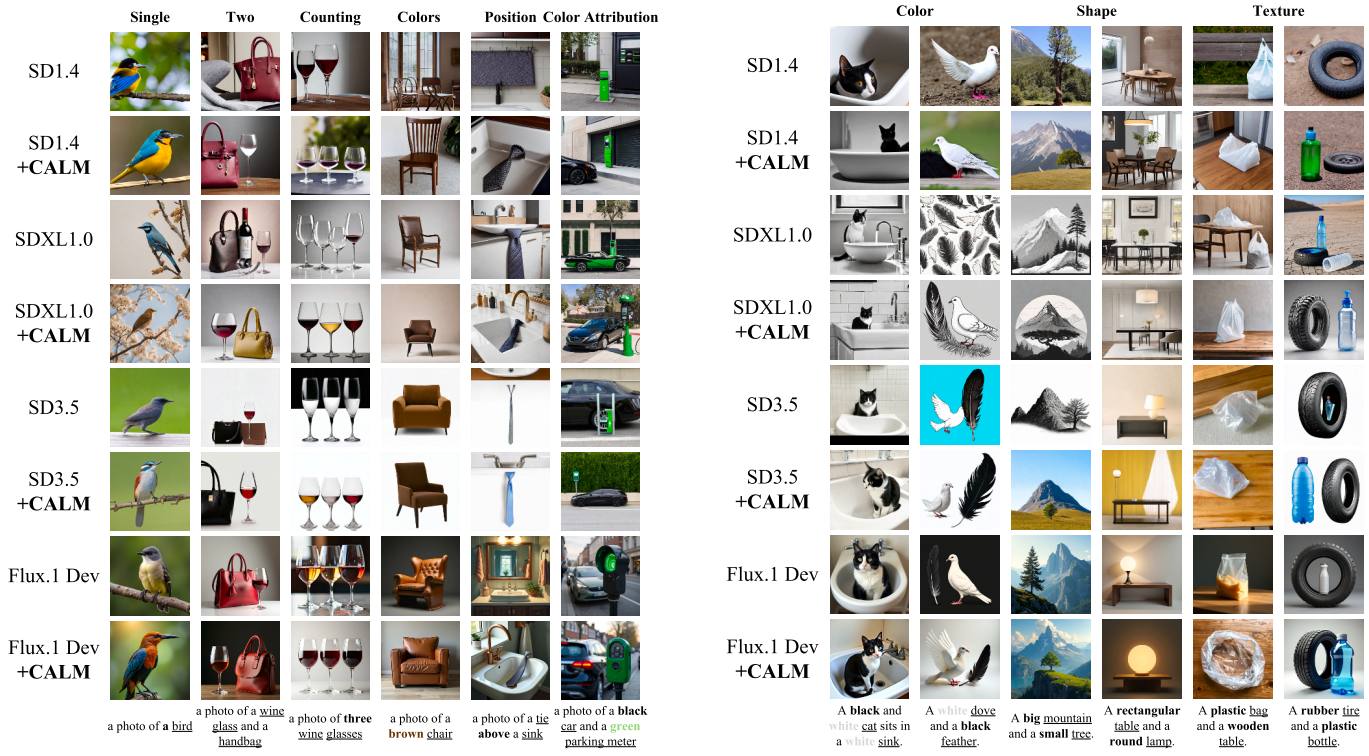
5.3. Qualitative result

In Fig. 4a–c, we present the image generation results on different datasets, based on various backbones and latent optimization methods, respectively. Issues like object omission, attribute omission, and semantic leakage are common in prior methods. Specifically, the baseline models often exhibit object omission when generating objects or animals, such as “backpack” in Fig. 4a, etc.. In addition, due to semantic conflicts between the object and the attribute information in the prompt, some baseline models exhibit attribute omission issues, for instance, “blue” in Fig. 4b, etc.. Semantic leakage is the most commonly observed issue. In Fig. 4c, there is incorrect binding of “spiky” and “rabbit”, etc.. These examples reflect broader issues of semantic misalignment and incoherent attribute rendering. We observe that the issues of semantic leakage and insufficient fidelity persist not only in SD 1.4, but also in more advanced models such as SDXL 1.0 and SD 3.5. While these newer models generate more photorealistic images than SD 1.4, they still suffer from semantic leakage. Our CALM consistently demonstrates superior fidelity and image quality across diverse datasets and model backbones, outperforming baseline methods that fail to achieve comparable results.

Fig. 5a presents the qualitative analysis results on the GenEval dataset. We observe that although most models can accurately generate single-object images, as the number of objects increases and their colors become more complex, even employing a stronger backbone cannot fully overcome the challenge of semantic entanglement. After applying the CALM framework, the model can significantly mitigate issues of semantic leakage and object omission, which improves semantic fidelity across various tasks. Moreover, we observe that our semantic modulation also facilitates other tasks, such as object counting and spatial reasoning. Fig. 5b presents the qualitative analysis results on the T2I-CompBench dataset. The results are consistent with observations on the AnE, DVMP, and ABC-6K datasets. In the multi-color binding task, our model enhances the semantic fidelity of existing models during image generation. Moreover, CALM effectively addresses semantic binding issues related to object size, shape, and texture. Compared with the base model, the CALM-enhanced model appropriately amplifies the semantics of modifiers, which enhances contrast and visual impact.

5.4. Ablation study

The ablation study conducted across the AnE, ABC-6K, and DVMP datasets (Table 4) highlights the indispensable contribution of each component within the CALM framework. Visualized examples for SD1.4, SDXL1.0, and SD3.5 backbones are further provided in Fig. 6. Across all three backbones, removing the complete Contextual Augmentation (w/o CA) leads to the most significant drop in both IR and AES scores. This observation suggests that the lack of contextual refinement causes substantial semantic misalignment, including frequent object omission and attribute leakage. Specifically, the results show that removing object augmentation (w/o CA$_o$) generally leads to a more pronounced performance decrease than removing attribute augmentation (w/o CA$_a$). This indicates that establishing a solid foundation for key objects is a prerequisite for preserving text-to-image semantic alignment. However,



(a) Qualitization result on GenEval dataset.

(b) Qualitization result on T2I-CompBench dataset.

Fig. 5. Comparison of qualitative results. Left: GenEval; Right: T2I-CompBench.

Table 4
Ablation study on AnE, ABC-6K, and DVMP datasets across SD1.4, SDXL1.0, and SD3.5 backbones. Bold indicates the maximum values per column within each base model group.

Base Model	Method	Animal-Animal			Animal-Object			Object-Object			ABC-6K			DVMP		
		Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES
SD 1.4	CALM	**0.348**	**1.326**	**5.636**	**0.371**	**1.689**	5.566	**0.375**	**1.538**	**5.291**	**0.346**	**0.448**	**5.410**	**0.358**	0.549	**5.460**
	w/o CA_o	0.337	1.211	5.590	0.363	1.560	5.432	0.369	1.476	5.096	0.343	0.437	5.394	0.355	0.496	5.376
	w/o CA_a	0.341	1.237	5.536	0.368	1.673	**5.572**	0.372	1.492	5.218	0.344	0.442	5.387	0.357	**0.624**	5.412
	w/o CA	0.329	1.105	5.353	0.349	1.327	5.437	0.325	1.033	5.130	0.337	0.131	5.275	0.348	−0.063	5.327
	w/o $\mathcal{L}_{t2r}$	0.332	1.203	5.432	0.355	1.406	5.467	0.361	1.376	5.260	0.339	0.237	5.249	0.340	0.034	5.319
	w/o $\mathcal{L}_{scl}$	0.326	1.234	5.421	0.357	1.562	5.507	0.360	1.496	5.221	0.342	0.233	5.320	0.342	0.317	5.213
	w/o $\mathcal{L}$	0.324	0.943	5.362	0.347	0.872	5.416	0.346	1.251	5.202	0.329	0.112	5.304	0.340	−0.160	5.351
	w/o all	0.311	−0.237	5.384	0.340	0.098	5.398	0.335	−0.661	5.151	0.324	−0.198	5.355	0.331	−0.544	5.410
SDXL 1.0	CALM	**0.317**	**1.448**	**5.988**	**0.356**	**1.499**	**5.921**	**0.363**	**0.967**	**5.530**	**0.345**	**0.769**	**5.634**	**0.350**	**0.871**	**5.695**
	w/o CA_o	0.315	1.397	5.803	0.353	1.459	5.842	0.360	0.883	5.521	0.342	0.759	5.504	0.347	0.859	5.549
	w/o CA_a	**0.317**	1.416	5.954	0.354	1.456	5.881	0.362	0.933	5.517	0.342	0.760	5.576	0.348	0.868	5.608
	w/o CA	0.314	1.313	5.786	0.353	1.442	5.637	0.350	0.783	5.351	0.340	0.757	5.493	0.342	0.843	5.547
	w/o $\mathcal{L}_{t2r}$	0.313	1.435	5.769	0.355	1.420	5.792	0.359	0.851	5.402	0.339	0.732	5.577	0.341	0.845	5.438
	w/o $\mathcal{L}_{scl}$	0.313	1.423	5.752	0.351	1.388	5.767	0.353	0.829	5.576	0.340	0.715	5.544	0.341	0.828	5.511
	w/o $\mathcal{L}$	0.313	0.925	5.719	0.349	1.291	5.601	0.355	0.654	5.284	0.343	0.747	5.591	0.345	0.724	5.527
	w/o all	0.309	0.395	5.625	0.349	1.274	5.564	0.350	0.402	5.342	0.340	0.649	5.613	0.339	−0.298	5.667
SD 3.5	CALM	**0.354**	**1.707**	**5.897**	**0.368**	**1.826**	5.812	**0.378**	1.841	**5.263**	**0.350**	**1.396**	**5.670**	**0.370**	**1.179**	**5.706**
	w/o CA_o	0.347	1.596	5.739	0.366	1.720	5.663	0.375	1.813	5.154	0.348	1.324	5.396	0.368	1.118	5.493
	w/o CA_a	0.351	1.691	5.857	0.366	1.809	**5.815**	0.377	**1.848**	5.192	0.347	1.337	5.387	0.369	1.175	5.517
	w/o CA	0.343	1.576	5.625	0.353	1.460	5.573	0.367	1.764	5.093	0.344	1.123	5.396	0.357	1.120	5.476
	w/o $\mathcal{L}_{t2r}$	0.344	1.632	5.783	0.360	1.773	5.741	0.376	1.825	5.213	0.349	1.384	5.433	0.367	1.166	5.554
	w/o $\mathcal{L}_{scl}$	0.347	1.614	5.764	0.363	1.755	5.723	0.372	1.840	5.184	0.345	1.353	5.372	0.364	1.135	5.483
	w/o $\mathcal{L}$	0.344	1.550	5.756	0.362	1.636	5.482	0.371	1.717	5.217	0.346	1.201	5.402	0.363	1.095	5.501
	w/o all	0.345	1.543	5.691	0.361	1.609	5.564	0.368	1.623	5.087	0.343	1.179	5.469	0.363	0.979	5.550

because the objectives of semantic alignment and image quality optimization are not always perfectly aligned, some quality-related metrics reach their peak under the w/o CA_a configuration. The structural losses also play a vital role in the final performance. Removing the token region loss (w/o $\mathcal{L}_{t2r}$) results in moderate degradation because it weakens the link between specific text tokens and their corresponding image regions. This loss is essential to ensure that each object and its modifiers are properly reflected in the generated spatial layout. Similarly, removing the supervised contrastive loss (w/o $\mathcal{L}_{scl}$) leads to a noticeable decline in accuracy. This loss contributes to fine-grained differentiation by separating distinct semantic elements and preventing them from blending incorrectly. To further verify that semantic leakage originates from the internal context bias of text encoders, we evaluate the (w/o $\mathcal{L}$) configuration, which employs the context augmentation mechanism without

Table 5
Sensitivity analysis of the parameter β.

Base Model	β	Animal-Animal			Animal-Object			Object-Object			ABC-6K			DVMP		
		Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES
SD 1.4	0.0	0.326	1.234	5.421	0.357	1.562	5.507	0.360	1.496	5.221	0.342	0.233	5.320	0.342	0.317	5.213
	0.1	**0.348**	**1.326**	**5.636**	**0.371**	**1.689**	**5.566**	**0.375**	**1.538**	**5.291**	**0.346**	**0.448**	**5.410**	0.347	0.532	5.437
	0.5	0.342	1.283	5.581	0.363	1.622	5.533	0.371	1.521	5.284	0.342	0.421	5.257	**0.358**	**0.549**	**5.460**
	1.0	0.335	1.250	5.540	0.355	1.590	5.512	0.365	1.523	5.257	0.339	0.414	5.335	0.351	0.525	5.449
SDXL 1.0	0.0	0.313	1.423	5.752	0.351	1.388	5.767	0.353	0.829	5.576	0.340	0.715	5.544	0.341	0.828	5.511
	0.1	**0.317**	**1.448**	**5.988**	0.349	1.370	5.740	**0.363**	**0.967**	5.530	0.342	0.702	5.521	**0.350**	**0.871**	**5.695**
	0.5	0.315	1.435	5.943	**0.356**	**1.499**	**5.921**	0.356	0.950	5.506	**0.345**	**0.769**	**5.634**	0.348	0.851	5.653
	1.0	0.312	1.376	5.907	0.352	1.473	5.890	0.355	0.930	**5.680**	0.342	0.723	5.565	0.347	0.830	5.620
SD 3.5	0.0	0.347	1.614	5.764	0.363	1.755	5.723	0.372	1.840	5.184	0.345	1.353	5.472	0.364	1.135	5.483
	0.1	**0.354**	**1.707**	**5.897**	**0.368**	**1.826**	**5.812**	0.370	1.830	5.209	0.342	1.340	5.564	**0.370**	1.179	**5.706**
	0.5	0.349	1.689	5.866	0.362	1.803	5.791	**0.378**	**1.841**	**5.263**	**0.350**	**1.396**	**5.670**	0.365	1.163	5.584
	1.0	0.345	1.651	5.822	0.358	1.773	5.752	0.373	1.820	5.240	0.348	1.371	5.542	0.366	**1.253**	5.551

(a) SD1.4 results.

(b) SDXL1.0 results.

(c) SD3.5 results.

Fig. 6. Visualized ablation results across SD1.4, SDXL1.0, and SD3.5 backbones. These results demonstrate that contextual enhancement improves modifier relationship understanding, while token-region alignment and semantic alignment binding ensure precise object preservation and attribute control.

any auxiliary loss functions. This variant outperforms the vanilla baseline (w/o all) across all tested backbones, which confirms that direct recalibration of the embedding space is intrinsically effective for mitigating semantic interference, even in the absence of explicit loss constraints. Overall, these components are highly complementary. Whether applied to the traditional UNet of SD1.4 or the DiT-based architecture of SD3.5, the combination of these modules enables the generation of coherent images that faithfully reflect complex input prompts.

5.5. Sensitivity analysis of hyperparameter

Sensitivity of β. The quantitative results in Table 5 and the qualitative visualizations in Fig. 7a collectively demonstrate the impact of the weighting parameter β on model performance. Introducing the supervised contrastive loss $\mathcal{L}_{scl}$ through a positive β value consistently improves performance across all models and datasets, with $\beta = 0.1$ providing the most stable gains. While most subsets peak at this value, certain categories, such as DVMP for SD 1.4 and Animal-Object for SDXL 1.0, benefit from a stronger contribution. For SD 3.5, performance is optimal at 0.1, although complex tasks like ABC-6K peak at 0.5 before slightly degrading at higher values. Visualized examples further clarify these trade-offs. When β is set to 0, the inactive supervised contrastive loss results in visible semantic leakage. Conversely, an excessively high β allows $\mathcal{L}_{scl}$ to dominate the entire optimization, which renders other loss functions insignificant and leads to missing objects or failed attribute binding. Therefore, $\beta = 0.1$ serves as the robust default setting for maintaining semantic integrity across diverse generative tasks.

Sensitivity of τ_o. The sensitivity of the object level augmentation threshold τ_o is reported through the quantitative results in Table 6 and the representative examples in Fig. 7b. This parameter determines which objects are enhanced within the contextual augmentation module. A threshold that is too low results in few objects being enhanced, which limits the potential benefits of the augmentation process. Conversely, an excessively high threshold amplifies nearly all objects and consequently dilutes the attention for descriptive modifiers. This imbalance disrupts the coordination among co-occurring objects in the prompt and leads to less coherent generations. The results indicate that moderate thresholds achieve the best trade-off between object prominence and attribute preservation. For instance, SD 1.4 reaches peak performance between 0.2 and 0.8. In more advanced architectures like SDXL 1.0 and SD 3.5, sensitivity is dataset dependent, where lower values, such as 0.2, are optimal for Animal-Animal subsets, but higher thresholds between 0.4 and 0.8 are superior for complex Object-Object tasks. As illustrated in Fig. 6, τ_o serves as a critical mechanism for managing the relative rendering strength of objects. It places greater emphasis on entities with initially weaker representations to ensure they are not omitted from the final scene. This control effectively regulates spatial proportions and visual contrast in the generated image. Overall, these findings confirm that moderate thresholds enhance semantic coverage without overwhelming modifier information or distorting inter-object relations, ensuring that each component of the prompt is rendered with appropriate clarity.

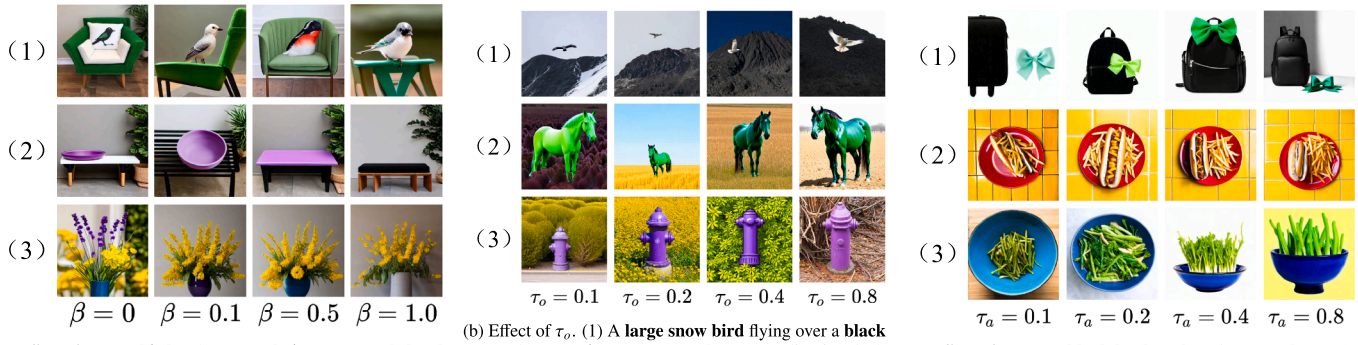
Sensitivity of τ_a. The sensitivity analysis of the modifier level augmentation threshold τ_a is presented through the quantitative data in Table 7 and the visual evidence in Fig. 7c. This parameter regulates the strength of attributes within the contextual augmentation module by ensuring that modifiers are better distinguished across different objects while preserving their semantic affinity with the associated object. In contrast to the object level threshold τ_o, the primary goal here is to enhance only the modifiers to improve the clarity of descriptive cues. The results indicate that moderate thresholds generally provide the best balance between attribute clarity and semantic prominence across various backbones. For instance, SD 1.4 typically peaks at 0.4 while more

Table 6
Sensitivity analysis of the parameter τ_o.

Base Model	τ_o	Animal-Animal			Animal-Object			Object-Object			ABC-6K			DVMP		
		Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES
SD 1.4	0.1	0.342	1.253	5.581	0.355	1.314	5.472	0.364	1.287	5.129	0.345	0.404	5.351	0.348	0.423	5.376
	0.2	0.342	1.274	5.608	0.358	1.337	5.463	0.362	1.331	5.167	0.348	0.427	5.376	**0.355**	**0.560**	**5.670**
	0.4	**0.348**	**1.326**	**5.636**	0.367	1.679	5.560	**0.375**	**1.538**	**5.291**	**0.350**	**0.470**	**5.420**	0.353	0.554	5.463
	0.8	0.346	1.303	5.611	**0.371**	**1.689**	**5.566**	0.373	1.534	5.165	0.346	0.448	5.413	0.352	0.437	5.395
SDXL 1.0	0.1	0.311	1.327	5.726	0.342	1.355	5.727	0.353	0.795	5.453	**0.345**	**0.769**	**5.634**	0.343	0.721	5.304
	0.2	**0.317**	**1.448**	**5.988**	0.352	1.386	5.874	0.355	0.914	5.530	0.342	0.742	5.613	**0.350**	**0.871**	**5.695**
	0.4	0.315	1.243	5.976	**0.356**	**1.499**	**5.921**	**0.363**	**0.967**	**5.530**	0.343	0.767	5.626	0.348	0.865	5.571
	0.8	0.314	1.373	5.945	0.354	1.487	5.913	0.361	0.953	5.522	0.340	0.763	5.632	0.347	0.859	5.434
SD 3.5	0.1	0.345	1.631	5.754	0.364	1.714	5.749	0.375	1.812	5.188	0.343	1.321	5.379	0.343	1.038	5.239
	0.2	**0.354**	**1.707**	**5.897**	0.366	1.753	5.748	0.376	1.787	5.199	0.348	1.398	5.436	0.361	1.067	5.545
	0.4	0.352	1.697	5.799	**0.368**	**1.826**	**5.812**	0.374	1.829	**5.264**	**0.350**	**1.396**	**5.670**	**0.370**	**1.179**	**5.706**
	0.8	0.350	1.670	5.793	0.367	1.792	5.797	**0.378**	**1.841**	5.263	0.347	1.370	5.457	0.364	1.045	5.571

Table 7
Sensitivity analysis of the parameter τ_a.

Base Model	τ_a	Animal-Animal			Animal-Object			Object-Object			ABC-6K			DVMP		
		Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES	Full	IR	AES
SD 1.4	0.1	0.341	1.115	5.472	0.355	1.313	5.475	0.357	1.282	5.197	0.335	0.312	5.298	0.348	0.387	5.337
	0.2	0.342	1.275	5.614	0.358	1.334	5.530	0.362	1.403	5.239	0.344	0.437	5.396	0.352	0.435	5.396
	0.4	**0.348**	**1.326**	**5.636**	0.360	1.400	5.550	**0.375**	**1.538**	**5.291**	**0.346**	**0.448**	**5.410**	**0.358**	**0.549**	**5.460**
	0.8	0.345	1.316	5.597	**0.371**	**1.689**	**5.566**	0.373	1.523	5.289	0.345	0.439	5.372	0.355	0.514	5.458
SDXL 1.0	0.1	0.305	1.314	5.594	0.331	1.355	5.728	0.357	0.792	5.349	0.337	0.726	5.627	0.348	0.819	5.543
	0.2	**0.317**	**1.448**	**5.988**	0.343	1.387	5.862	0.355	0.877	5.514	**0.345**	**0.769**	**5.634**	**0.350**	**0.871**	**5.695**
	0.4	0.310	1.423	5.933	**0.356**	**1.499**	**5.921**	0.360	0.946	**5.533**	0.337	0.757	5.621	0.348	0.861	5.670
	0.8	0.308	1.318	5.893	0.355	1.485	5.870	**0.363**	**0.967**	5.530	0.331	0.653	5.395	0.335	0.772	5.323
SD 3.5	0.1	0.345	1.610	5.743	0.355	1.700	5.696	0.376	1.795	5.164	0.337	1.221	5.363	0.353	0.869	5.459
	0.2	**0.354**	**1.707**	**5.897**	0.365	1.750	5.750	**0.378**	1.841	**5.263**	0.348	1.373	5.417	0.361	0.953	5.546
	0.4	0.352	1.690	5.870	**0.368**	**1.826**	**5.812**	0.373	1.769	5.246	**0.350**	**1.396**	**5.670**	**0.370**	**1.179**	**5.706**
	0.8	0.337	1.489	5.593	0.353	1.738	5.757	0.378	**1.857**	5.263	0.348	1.376	5.437	0.365	1.064	5.567



(a) Effect of β. (1) a **bird** and a **green chair**. (2) a **purple bowl** and a **black bench**. (3) A **slender**, **purple vase** containing many **yellow flowers**.

(b) Effect of τ_o. (1) A **large snow bird** flying over a **black** covered **mountain**. (2) A **green horse** standing in a lush brown **field**. (3) A purple **fire hydrant** is covered by yellow **flowers**.

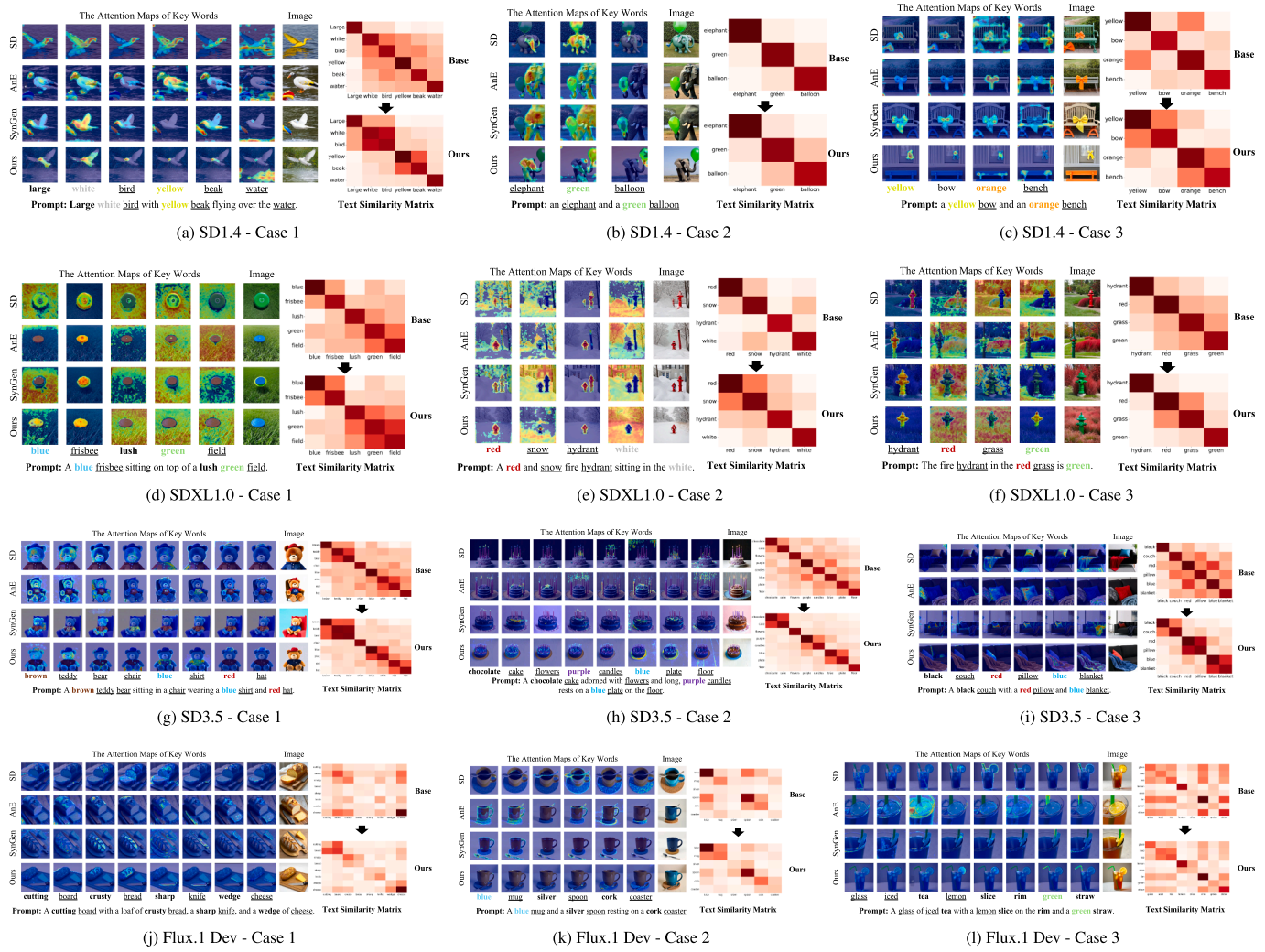
(c) Effect of τ_a. (1) a **black backpack** and a **green bow**. (2) **Hot dog** with **fries** on **red plate** on **yellow tile**. (3) The **blue bowl** has **skinny green vegetables** in it.

Fig. 7. Qualitative sensitivity analysis on the effects of β, τ_o, and τ_a. Each row in the subfigures shares the same random seed.

advanced models like SDXL 1.0 and SD 3.5 exhibit a dataset-dependent sensitivity where values of 0.2 are optimal for Animal-Animal but values of 0.4 favor complex Animal-Object and DVMP tasks. These findings suggest that while strengthening modifiers improves faithfulness by separating cues for different objects, an excessively high threshold may diminish the prominence of individual attributes. As illustrated in Fig. 7c, the hyperparameter τ_a controls the rendering strength of attributes by ensuring that modifiers within the same semantic group receive consistent enhancement. This mechanism effectively helps distinguish attributes associated with different objects and improves overall visual contrast. Similar to the observations for the object-level threshold, the optimal value for τ_a varies across different image domains and should be adjusted according to the characteristics of the data to ensure the most robust generative results.

5.6. Attention visualization analysis

We evaluate the quality of latent representations by examining keyword cross-attention interactions at the final timestep, where closer alignment between attribute and object attention maps reflects reduced semantic leakage. Regarding the SD1.4 backbone in the top row of Fig. 8, baseline models often exhibit significant semantic leakage, e.g., the incorrect transfer of color semantics from a balloon to an elephant or attribute omission caused by global attention interference. For the SDXL1.0 backbone shown in the middle row, the performance heavily depends on learned visual priors, where the model handles common patterns well but fails when prompt semantics conflict with pretraining distributions, e.g., assigning unrealistic colors to fire hydrants or grass that contradict internal visual priors. For the SD3.5, despite having



(a) SD1.4 - Case 1 (b) SD1.4 - Case 2 (c) SD1.4 - Case 3

(d) SDXL1.0 - Case 1 (e) SDXL1.0 - Case 2 (f) SDXL1.0 - Case 3

(g) SD3.5 - Case 1 (h) SD3.5 - Case 2 (i) SD3.5 - Case 3

(j) Flux.1 Dev - Case 1 (k) Flux.1 Dev - Case 2 (l) Flux.1 Dev - Case 3

Fig. 8. Visualization of cross attention maps grouped by diffusion backbones. The rows from top to bottom display the attention distributions for SD1.4, SDXL1.0, SD3.5, and Flux.1 Dev, respectively. Each column corresponds to a specific prompt scenario, demonstrating the internal consistency and semantic focus of our CALM framework across various architectures.

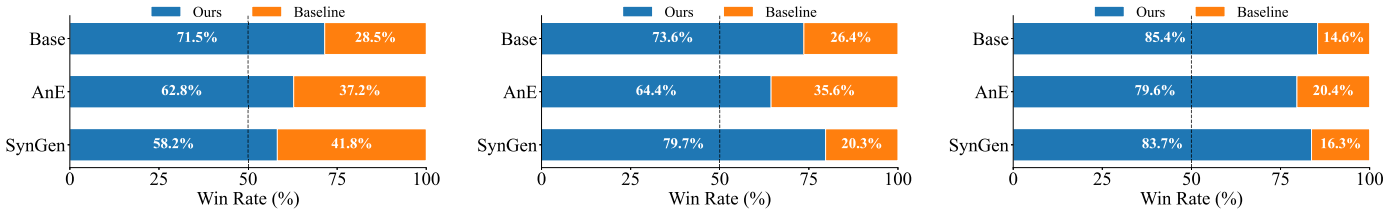
more refined attention distributions, persistent issues remain in complex environments, e.g., semantic interference from surrounding furniture or the omission of background elements like the floor. Regarding Flux.1 Dev, the attention encounters attention leakage and erroneous rendering when processing complex scenes with multiple objects that demand strict positional and morphological accuracy. Across all three backbones, CALM achieves superior semantic fidelity by ensuring that attention for related pairs overlaps significantly while remaining distinct for different semantic groups, which effectively mitigates the leakage and entanglement observed in baseline methods. In summary, the baseline models exhibit frequent semantic leakage and attribute-object entanglement, particularly under conflicting or uncommon prompt compositions. These phenomena stem from the inherent biases and limitations of their learned representations. Our CALM framework achieves better semantic fidelity and more precise attention alignment, effectively mitigating semantic leakage and improving the consistency between generated images and the textual prompts.

5.7. Human evaluation

Recent work suggests that objective metrics alone often fail to capture the nuances of image quality and human preference, which frequently leads to mismatches between automated scores and actual visual appeal. To address this, we conducted a comprehensive human evaluation using 100 randomly selected prompts from each dataset to incorporate critical dimensions of context and aesthetics. We recruited 92 participants with diverse disciplinary backgrounds, including 60 master's students and 23 PhD students, as well as 9 professors. All annotators possessed normal visual and color perception abilities, and their evaluations were conducted under strict confidentiality.

To ensure a rigorous and unbiased assessment, we adopted a double blind binary forced-choice protocol. In each trial, an annotator was presented with a single textual prompt and two images generated by different methods, such as CALM versus a baseline. The identities of the methods were randomized and hidden from the raters to maintainthe integrity of the results. Participants were required to answer the specific question of which image best matches the prompt description. By forcing a selection between the two images, we ensured a clear relative preference for each comparison. This process was facilitated by a dedicated web-based tool illustrated in Fig. 10, which supported real-time data synchronization and efficient evaluation. The resulting win rates across all three backbones are presented in Fig. 9.

The human preference win rates across different backbones are illustrated in Fig. 9. On the SD1.4 backbone, although it is a relatively smaller model, CALM achieves significant preference margins with win rates of 58.2% against SynGen and 71.0% against the vanilla base



(a) SD1.4 (b) SDXL1.0 (c) SD3.5

Fig. 9. Human preference results across different backbones. We report the win rates of CALM against various baselines through a binary forced-choice evaluation. The preference rates demonstrate the consistent superiority of CALM in generating images that align better with human perception and textual descriptions.

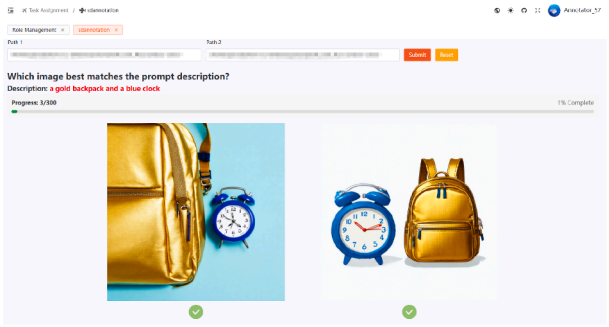


Fig. 10. Interface of the annotation tool. The clean, web-based interface presents the textual prompt and the generated image, requiring the rater to perform a binary forced-choice evaluation on the alignment quality.

model. When moving to the more advanced SDXL 1.0 architecture, CALM further amplifies the potential of the backbone by reaching a dominant win rate of 79.7% against SynGen and 73.6% against the base model. This trend suggests that our method is particularly effective at leveraging the increased capacity of larger diffusion models to meet user expectations for precise and visually appealing generation. The results for the DiT-based SD3.5 backbone further confirm the consistent superiority of our approach. While SD3.5 possesses formidable generative capabilities, our evaluations indicate that traditional refinement methods often struggle to fully unlock its potential for high fidelity alignment. In contrast, CALM consistently dominates the preference evaluations on this platform by achieving substantial win rates of 83.7% against SynGen and 85.4% against the base model. These findings empirically demonstrate that CALM is better optimized for flow matching architectures and significantly enhances both perceived semantic consistency and visual quality. The overwhelming preference for CALM across all three backbones highlights its effectiveness in refining complex textual prompts into precise visual representations that align more closely with human expectations than existing state-of-the-art methods.

5.8. Computation efficiency

We randomly sampled 100 prompts from the DVMP dataset to evaluate computation efficiency, and the results are shown in Table 8. SD represents the baseline performance obtained by using only the backbone network without any latent optimization, serving as a lower bound on inference time. It can be observed that AnE requires the longest time due to its iterative latent refinement process, which continues until either the latent quality reaches a predefined threshold or the maximum number of iterations is exhausted. While CONFORM and T-SAM adopt a strategy of iterating only at specific timesteps, they still require multiple internal iterations, resulting in a significantly higher computational overhead compared to CALM. Even compared to methods like SynGen and EBAMA, which perform optimization only once at a single timestep, our method achieves superior efficiency. This is because CALM effectively leverages information from the attention mechanism, significantly

Table 8
The average time (min) for generating images based on 100 random prompts on DVMP.

Method	SD1.4	SDXL1.0	SD3.5
SD	15.68	142.28	223.33
AnE	42.53	319.48	408.72
SynGen	26.01	214.01	275.38
EBAMA	24.93	205.75	267.93
CONFORM	25.30	–	–
T-SAM	24.02	–	–
Ours	**22.96**	**197.13**	**261.29**

Table 9
Impact of the *spaCy* coreference resolution component (f) on ABC-6K (Full Sim.) across different base models. **CALM**$_f$ denotes the use of *fastcoref* (Otmazgin et al., 2022) as a custom *spaCy* component for enhanced coreference resolution.

Model	Method	ABC-6K
SD1.4	CALM	0.346
	CALM$_f$	**0.349**
SDXL1.0	CALM	0.345
	CALM$_f$	**0.351**
SD3.5	CALM	0.350
	CALM$_f$	**0.351**

reducing the extensive pairwise attention operations typically employed in other approaches. Our method achieves the lowest time consumption among all methods. The shortest inference time enhances the practical applicability of CALM.

5.9. Limitation and future work

While CALM is effective on structured prompts (e.g., AnE, DVMP, GenEval, and T2I-CompBench), Table 9 demonstrates that it becomes a performance bottleneck on unstructured data (e.g., ABC-6K) due to the inherent parsing limitations of spaCy in complex semantic resolution as noted in prior studies. Additionally, the current reliance on manual hyperparameter tuning across different scenarios limits the model's "adaptive" potential in open-domain tasks. Future work will focus on replacing traditional parsers with generative-side attention sharing constraints or entity-centric scene graph modeling to ensure semantic consistency, while simultaneously implementing learnable hyperparameters.

6. Conclusion

We introduce CALM, an inference-time optimization framework that enhances semantic fidelity in text-to-image generative models. We point out that contextual bias acts as a driving force behind semantic misalignment. We leverage semantic group decomposition for adaptive context

enhancement. By representing prompts as structured attribute–object relations, we guide attention dynamics through token–region alignment with contrastive objectives. Experiments on five datasets confirm that CALM enforces more accurate and corrects semantic failures while simultaneously improving image quality on the fly. Moreover, it provides an enhancement that is agnostic to both model architecture and training methods. We believe CALM paves the way for more reliable and semantically grounded image generation.

CRediT authorship contribution statement

Jili Chen: Conceptualization, Methodology, Investigation, Formal analysis, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Huicheng Zeng:** Conceptualization, Methodology, Validation, Visualization, Investigation, Writing – review & editing; **Changqin Huang:** Conceptualization, Methodology, Investigation, Formal analysis, Software, Validation, Writing – review & editing; **Qionghao Huang:** Methodology, Validation, Visualization, Writing – review & editing; **Zilong Li:** Validation, Visualization, Investigation, Writing – review & editing; **Xiaodi Huang:** Methodology, Validation, Visualization, Writing – review & editing.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant Nos. 62337001 and U25A20440) and the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (Grant No. 2025C02022).

References

Agarwal, A., Karanam, S., Joseph, K. J., Saxena, A., Goswami, K., & Srinivasan, B. V. (2023). A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2283–2293).

Bai, L., Sugiyama, M., & Xie, Z. (2025). Weak-to-strong diffusion with reflection. arXiv:2502.00473.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S. et al. (2022). ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv:2211.01324.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y. et al. (2023). Improving image generation with better captions. *Computer Science*. https://cdn.openai.com/papers/dall-e-3.pdf, *2*(3), 8.

Black, K., Janner, M., Du, Y., Kostrikov, I., & Levine, S. (2023). Training diffusion models with reinforcement learning. arXiv:2305.13301.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., & Cohen-Or, D. (2023). Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, *42*(4), 1–10.

Chen, C.-S., & Kuo, E.-J. (2025). Quantum reinforcement learning-guided diffusion model for image synthesis via hybrid quantum-classical generative model architectures. arXiv:2509.14163.

Chen, C.-Y., Tseng, C., Tsao, L.-W., & Shuai, H.-H. (2024). A cat is a cat (not a dog!): Unraveling information mix-ups in text-to-image encoders through causal analysis and embedding optimization. *Advances in Neural Information Processing Systems*, *37*, 57944–57969.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H. et al. (2023). Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv:2310.00426.

Chen, L., Bai, S., Chai, W., Xie, W., Zhao, H., Vinci, L., Lin, J., & Chang, B. (2025). Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6146–6156).

Clark, K., Vicol, P., Swersky, K., & Fleet, D. J. (2023). Directly fine-tuning diffusion models on differentiable rewards. arXiv:2309.17400.

Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., & Zhang, T. (2023). Raft: Reward ranked finetuning for generative foundation model alignment. arXiv:2304.06767.

Du, C., Li, Y., Qiu, Z., & Xu, C. (2023). Stable diffusion is unstable. *Advances in Neural Information Processing Systems*, *36*, 58648–58669.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F. et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Fan, J., Shen, S., Cheng, C., Chen, Y., Liang, C., & Liu, G. (2025). Online reward-weighted fine-tuning of flow matching with wasserstein regularization. In *The thirteenth international conference on learning representations*.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., & Lee, K. (2023). DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, *36*, 79858–79885.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., & Wang, W. Y. (2022). Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv:2212.05032.

Ghosh, D., Hajishirzi, H., & Schmidt, L. (2023). Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, *36*, 52132–52152.

Guo, X., Cui, M., Bo, L., & Huang, D. (2025). ShortFT: Diffusion model alignment via shortcut-based fine-tuning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 678–687).

Guo, X., Liu, J., Cui, M., Li, J., Yang, H., & Huang, D. (2024). Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9380–9389).

He, H., Ye, Y., Liu, J., Liang, J., Wang, Z., Yuan, Z., Wang, X., Mao, H., Wan, P., & Pan, L. (2025a). Gardo: Reinforcing diffusion models without reward hacking. arXiv:2512.24138.

He, W., Fu, S., Liu, M., Wang, X., Xiao, W., Shu, F., Wang, Y., Zhang, L., Yu, Z., Li, H. et al. (2025b). Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 17123–17131). (*vol. 39*).

He, X., Fu, S., Zhao, Y., Li, W., Yang, J., Yin, D., Rao, F., & Zhang, B. (2025c). Tempflow-GRPO: When timing matters for GRPO in flow models. arXiv:2508.04324.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. arXiv:2208.01626.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv:2207.12598.

Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Conference on empirical methods in natural language processing, EMNLP 2015* (pp. 1373–1378). Association for Computational Linguistics (ACL).

Hu, Z., Zhang, F., Chen, L., Kuang, K., Li, J., Gao, K., Xiao, J., Wang, X., & Zhu, W. (2025). Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards. In *Proceedings of the computer vision and pattern recognition conference* (pp. 23604–23614).

Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., & Liu, X. (2025). T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(5), 3563–3579.

Huang, K., Sun, K., Xie, E., Li, Z., & Liu, X. (2023). T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, *36*, 78723–78747.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, *33*, 18661–18673.

Kim, J., Esmaeili, E., & Qiu, Q. (2025). Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. In *Proceedings of the computer vision and pattern recognition conference* (pp. 8031–8040).

Kingma, D. P., Welling, M. (2013). Auto-encoding variational Bayes. arXiv:1312.6114.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., & Levy, O. (2023). Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, *36*, 36652–36663.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., & Gu, S. S. (2023). Aligning text-to-image models using human feedback. arXiv:2302.12192.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888–12900). PMLR.

Li, Y., Keuper, M., Zhang, D., & Khoreva, A. (2023). Divide & bind your attention for improved generative semantic nursing. arXiv:2307.10864.

Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., & Zheng, L. (2025). Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the computer vision and pattern recognition conference* (pp. 13199–13208).

Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., & Ouyang, W. (2025a). Flow-GRPO: Training flow matching models via online RL. arXiv:2505.05470.

Liu, M., Ma, Y., Yang, Z., Dan, J., Yu, Y., Zhao, Z., Hu, Z., Liu, B., & Fan, C. (2025b). LLM4GEN: Leveraging semantic representation of LLMs for text-to-image generation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 5523–5531). (*vol. 39*).

Liu, N., Li, S., Du, Y., Torralba, A., & Tenenbaum, J. B. (2022). Compositional visual generation with composable diffusion models. In *European conference on computer vision* (pp. 423–439). Springer.

Luo, Y., Du, P., Li, B., Du, S., Zhang, T., Chang, Y., Wu, K., Gai, K., & Wang, X. (2025a). Sample by step, optimize by chunk: Chunk-level GRPO for text-to-image generation. arXiv:2510.21583.

Luo, Y., Hu, X., Fan, K., Sun, H., Chen, Z., Xia, B., Zhang, T., Chang, Y., & Wang, X. (2025b). Reinforcement learning meets masked generative models: Mask-GRPO for text-to-image generation. arXiv:2510.13418.

Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.-C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X. et al. (2025a). Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv:2501.09732.

Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X. et al. (2025b). Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7739–7751).

Meral, T. H. S., Simsar, E., Tombari, F., & Yanardag, P. (2024). Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9005–9014).

Murray, N., Marchesotti, L., & Perronnin, F. (2012). Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2408–2415). IEEE.

Otmazgin, S., Cattan, A., & Goldberg, Y. (2022). F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd conference of the Asia-Pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: System demonstrations* (pp. 48–56).

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4195–4205).

Peters, J., & Schaal, S. (2007). Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on machine learning* (pp. 745–750).

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952.

Prabhudesai, M., Goyal, A., Pathak, D., & Fragkiadaki, K. (2023). Aligning text-to-image diffusion models with reward backpropagation. arXiv:2310.03739.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PmLR.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*, 53728–53741.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125, *1*(2), 3.

Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., & Chechik, G. (2023). Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, *36*, 3536–3559.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, *35*, 36479–36494.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.

Shen, X., Li, Z., Yang, Z., Zhang, S., Zhang, Y., Li, D., Wang, C., Lu, Q., & Tang, Y. (2025). Directly aligning the full diffusion trajectory with fine-grained human preference. arXiv:2509.06942.

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv:2010.02502.

Sun, H., Liang, B., Xia, B., Wu, J., Zhao, Y., Qin, K., Chang, Y., & Wang, X. (2025a). Diffusion-rainbowpa: Improvements integrated preference alignment for diffusion-based text-to-image generation. *Transactions on Machine Learning Research*. ISSN: 2835-8856. Available at: https://openreview.net/forum?id=KY0TSY2bx8.

Sun, H., Wu, J., Xia, B., Luo, Y., Zhao, Y., Qin, K., Lv, X., Zhang, T., Chang, Y., & Wang, X. (2025b). Reinforcement fine-tuning powers reasoning capability of multimodal large language models. arXiv:2505.18536.

Sun, H., Xia, B., Chang, Y., & Wang, X. (2025c). Generalizing alignment paradigm of text-to-image generation with preferences through f-divergence minimization. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 27644–27652). (*vol. 39*).

Sun, H., Xia, B., Zhao, Y., Chang, Y., & Wang, X. (2025d). Identical human preference alignment paradigm for text-to-image models. In *Icassp 2025-2025 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5). IEEE.

Sun, H., Xia, B., Zhao, Y., Chang, Y., & Wang, X. (2025e). Positive enhanced preference alignment for text-to-image models. In *Icassp 2025-2025 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5). IEEE.

Tang, Z., Peng, J., Tang, J., Hong, M., Wang, F., & Chang, T.-H. (2024). Inference-time alignment of diffusion models with direct noise optimization. arXiv:2405.18881.

Tian, Y., Xia, X., Ren, Y., Lin, S., Wang, X., Xiao, X., Tong, Y., Yang, L., & Cui, B. (2025). Training-free diffusion acceleration with bottleneck sampling. arXiv:2503.18940.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., & Naik, N. (2024). Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8228–8238).

Wang, J., Liang, J., Liu, J., Liu, H., Liu, G., Zheng, J., Pang, W., Ma, A., Xie, Z., Wang, X. et al. (2025a). GRPO-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. arXiv:2510.22319.

Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q. et al. (2024). Emu3: Next-token prediction is all you need. arXiv:2409.18869.

Wang, Y., Li, Z., Zang, Y., Zhou, Y., Bu, J., Wang, C., Lu, Q., Jin, C., & Wang, J. (2025b). Pref-GRPO: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. arXiv:2508.20751.

Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C. et al. (2025). Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the computer vision and pattern recognition conference* (pp. 12966–12977).

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., & Li, H. (2023). Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv:2306.09341.

Wu, X., Hao, Y., Zhang, M., Sun, K., Huang, Z., Song, G., Liu, Y., & Li, H. (2024). Deep reward supervisions for tuning text-to-image diffusion models. In *European conference on computer vision* (pp. 108–124). Springer.

Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q., Gu, Y., Chen, Z., Yang, Z., & Shou, M. Z. (2024). Show-o: One single transformer to unify multimodal understanding and generation. arXiv:2408.12528.

Xie, X., & Gong, D. (2025). DyMO: Training-free diffusion model alignment with dynamic multi-objective scheduling. In *Proceedings of the computer vision and pattern recognition conference* (pp. 13220–13230).

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., & Dong, Y. (2023). ImageReward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, *36*, 15903–15935.

Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W. et al. (2025). DanceGRPO: Unleashing GRPO on visual generation. arXiv:2505.07818.

Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., & Li, X. (2024). Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8941–8951).

Zhai, K., Singh, U., Thatipelli, A., Chakraborty, S., Sahu, A. K., Huang, F., Bedi, A. S., & Shah, M. (2025). Mira: Towards mitigating reward hacking in inference-time alignment of T2I diffusion models. arXiv:2510.01549.

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3836–3847).

Zhang, Y., Tzeng, E., Du, Y., & Kislyuk, D. (2024a). Large-scale reinforcement learning for diffusion models. In *European conference on computer vision* (pp. 1–17). Springer.

Zhang, Y., Yu, P., & Wu, Y. N. (2024b). Object-conditioned energy-based attention map alignment in text-to-image diffusion models. In *European conference on computer vision* (pp. 55–71). Springer.

Zhang, Y., Zhou, J., Li, X., Zhang, Q., Wan, Z., Miao, D., Wang, C., & Cao, L. (2025). Enhancing text-to-image diffusion transformer via split-text conditioning. In *The thirty-ninth annual conference on neural information processing systems*.

Zhu, H., Xiao, T., & Honavar, V. G. (2025). DSPO: Direct score preference optimization for diffusion model alignment. In *The thirteenth international conference on learning representations*.