CO3: CONTRASTING CONCEPTS COMPOSE BETTER

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose to improve multi-concept prompt fidelity in text-to-image diffusion models. We begin with common failure cases—prompts like "a cat and a clock" that sometimes yields images where one concept is missing, faint, or colliding awkwardly with another. We hypothesize that this happens when the diffusion model drifts into mixed modes that over-emphasize a single concept it learned strongly during training. Instead of re-training, we introduce a *corrective sampling strategy* that steers away from regions where the joint prompt behavior overlaps too strongly with any single concept in the prompt. The goal is to steer towards "pure" joint modes where all concepts can coexist with balanced visual presence. We further show that existing multi-concept guidance schemes can operate in unstable weight regimes that amplify imbalance; we characterize favorable regions and adapt sampling to remain within them. Our approach, **CO3**, is plug-and-play, requires no model tuning, and complements standard classifier-free guidance. Experiments on diverse multi-concept prompts indicate improvements in concept coverage, balance and robustness, with fewer dropped or distorted concepts compared to standard baselines and prior compositional methods. Results suggest that lightweight corrective guidance can substantially mitigate brittle semantic alignment behavior in modern diffusion systems.

1 INTRODUCTION

Recent diffusion models (Ho et al., 2020; Ramesh et al., 2022a; Rombach et al., 2022) have ushered significant breakthroughs in Text-to-Image (T2I) synthesis, producing high-fidelity images from textual descriptions. However, ensuring the generated images faithfully adhere to the prompt, a challenge known as semantic alignment (Chefer et al., 2023; taihang Hu et al., 2024; Liu et al., 2023), remains a problem. Concretely, for a given prompt C, T2I models like StableDiffusion (Rombach et al., 2022) sample from the modes (or high probability regions) of the learned distribution, $p(x|C)$. While such models can produce high resolution images in general, every so often, the results are surprisingly misaligned even for very simple prompts containing few concepts, e.g., C="a cat and a dog". Diagnosing exactly why this behavior emerges periodically is difficult. It is conceivable that the complex training process in high dimensions, especially in conjunction with text embeddings, creates some problematic modes in $p(x|C)$.

We hypothesize that problematic modes in $p(x|C)$ arise when they overlap with modes of individual concept distribution $p(x|c_i)$. Such an overlap biases the generation toward a single concept c_i, reducing the prominence of others. For instance, across images of $c_1 =$ "cat" in the training dataset, a few may have an inconspicuous or partial $c_2 =$ "dog" in the background. This image may still fall under the mode of $p(x|C)$. We attribute this to training instabilities and relatively less coverage of multi-concept prompts C, which cause the model to assign high probability even to weakly conforming images. Said differently, an image of a big cat and an inconspicuous dog can get assigned high probabilities under $p(x|C)$, causing semantic misalignment.

Preventing such problematic modes warrants strict and specialized training paradigms; a difficult task for such large models. However, "curing" them after their occurrence is a more viable approach Assuming our hypothesis is true, we propose a cure for problematic modes. Our intuitive idea is to go away from problematic modes and move towards modes under which none of the individual concepts are strong. To realize this, we propose to design a *corrector* that generates samples from the following distribution:

$$\tilde{p}(x \mid C) \;\propto\; \frac{p(x \mid C)}{\prod_i p(x \mid c_i)}. \tag{1}$$

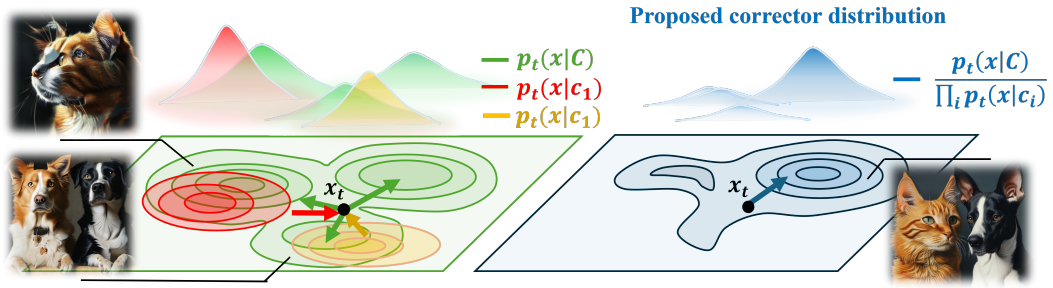

Figure 1: The figure illustrates our hypothesis on mode overlap using a simple 2D toy example. (a) Two modes of the distribution $p_t(x|$*"a cat and a dog"*$)$ (in green contour) has significant overlap with the modes of the individual concept distributions $p_t(x|$*"a cat"*$)$ (in red contour) and $p_t(x|$*"a dog"*$)$ (in orange contour). (b) The proposed corrector distribution $p_t(x|$`"a cat and a dog"`$)/(p_t(x|$*"a cat"*$)p_t(x|$*"a dog"*$))$ suppresses these overlaps, steering the generation away from problematic modes. The arrows indicate the denoising directions.

Figure 1 illustrates the intuition behind our proposal. Our corrector distribution $\tilde{p}(x|C)$ assigns low probability to regions where $p(x|C)$ overlaps with individual $p(x|c_i)$; we deem them as degenerate modes dominated by a single concept. By suppressing these overlaps, the corrector emphasizes *pure* $p(x|C)$ modes where all concepts coexist without one overwhelming the others. From a probabilistic perspective, this acts as a corrective factor: while $p(x|C)$ may assign high probability to weakly conforming images due to training noise or limited multi-concept data, dividing by the marginals removes this bias and sharpens the distribution toward genuine multi-concept samples. As a result, the modes we target are more semantically aligned and less prone to concept suppression or distortion.

Correction sampling from $\tilde{p}(x|C)$ can be achieved by composing scores from constituent component distributions $\nabla_x \log p(x|C), \{\nabla_x \log p(x|c_i)\}_i$ (Liu et al., 2022). While there are many ways to compose, we analyse and show that composition through weighted sum of Tweedie-means—in the Tweedie denoised space—offers a more general framework that subsumes existing approaches.

In particular, we show that two classes of correction sampling—noise-resampling and latent correction (Bao et al., 2024; Rassin et al., 2023; taihang Hu et al., 2024; Kwon & Ye, 2025)—become special cases of Tweedie-mean composition under different weighting schemes. This allows us to design a hybrid composition framework that serves the purpose of resampling at time T, and then toggles to latent correction at later steps. By latent, we mean that the correction steps are accomplished in between each DDIM time step, allowing **CO3** to be a plug-and-play, model-agnostic and gradient-free approach for T2I models. Comparison of **CO3** against SOTA baselines shows stronger semantic alignment to prompts (measured using multiple metrics), giving empirical evidence of our hypothesis.

2 Background

■ **Conditional generation using Classifier-Free Guidance (CFG).** In diffusion-based Text-to-Image (T2I) generation (Rombach et al., 2022; Saharia et al., 2022a; Ramesh et al., 2022b), given the noisy latent x_t at timestep t, a denoised estimate can be derived using Tweedie's formula:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, c, t)}{\sqrt{\bar{\alpha}_t}}, \tag{2}$$

where ϵ_θ denotes the predicted noise conditioned on the text prompt c, and $\bar{\alpha}_t$ is the cumulative product of the noising schedule. This step corresponds to the *denoising* stage, recovering an estimate of the clean signal x_0. In the DDIM sampler (Song et al., 2021), under the noise-free condition, the subsequent step deterministically evolves $\hat{x}_0$ to x_{t-1} without introducing additional stochasticity:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(x_t, c, t). \tag{3}$$

Here, the same predicted noise ϵ_θ is reused, eliminating the renoisification step present in stochastic samplers such as DDPM(Ho et al., 2020). This perspective highlights how the Tweedie decompo-

sition denoising via $\hat{x}_0$, followed by deterministic reconstruction of x_{t-1}, is naturally aligned with DDIM sampling.

In practice, most T2I models adopt *classifier-free guidance* (CFG) (Ho & Salimans, 2022), where predictions from both the conditional and unconditional models are combined:

$$\epsilon_t^{\lambda,c} \;=\; \lambda\,\epsilon_\theta(x_t, c, t) \;+\; (1-\lambda)\,\epsilon_\theta(x_t, \varnothing, t) \tag{4}$$

where $\lambda > 1$ controls the guidance strength. $\epsilon_t^{\lambda,c}$ is the composed noise prediction under CFG. Then the denoising and DDIM steps proceed as before, but using $\epsilon_t^{\lambda,c}$ in place of $\epsilon_\theta(x_t, c, t)$.

Tweedie View with CFG. Substituting $\epsilon_t^{\lambda,c}$ into the Tweedie denoising and DDIM update yields:

$$\hat{x}_0^{\lambda,c} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_t^{\lambda,c}}{\sqrt{\bar{\alpha}_t}} = \frac{1}{\sqrt{\bar{\alpha}_t}}\hat{x}_{tweedie}[\epsilon_t^{\lambda,c}], \tag{5}$$

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}\,\hat{x}_{tweedie}[\epsilon_t^{\lambda,c}] + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_t^{\lambda,c} \tag{6}$$

where $\hat{x}_{tweedie}[\epsilon_t^{\lambda,c}] := x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_t^{\lambda,c}$ is the Tweedie mean from the CFG noise at t. Thus, CFG in the Tweedie framework can be interpreted as modifying the denoised estimate $\hat{x}_0$ to $\hat{x}_0^{\lambda,c}$ before the deterministic step to x_{t-1} (Chung et al., 2024; Kwon & Ye, 2023).

■ **Correction-based approaches for conditional generation.** A number of recent works have addressed the challenge of compositional text-to-image generation using correction-based approaches (Chefer et al., 2023; Liu et al., 2023; Bao et al., 2024; Rassin et al., 2023; taihang Hu et al., 2024). During sampling with classifier-free guidance, these methods iteratively correct the latent variable by applying gradient updates of the form

$$x_t^{k+1} = x_t^k - s\,\nabla_{x_t}\mathcal{L}(x_t, c), \quad k = 1, 2, \ldots, M-1. \tag{7}$$

and then use the final refined latent x_t^M to predict the next DDIM step x_{t-1}, using Eqs. 5 and 6. Here $\mathcal{L}(x_t, c)$ is a task-specific loss function that enforces better alignment with the target condition c, and s is a step size. This iterative update can be interpreted as correcting the diffusion guidance process using a corrector distribution of the form:

$$\tilde{p}_t(x_t, c) \propto \exp\big(-\mathcal{L}(x_t, c)\big), \tag{8}$$

which refines the generative process at each timestep. In the special case where the step size $s = \sigma_t^2 = 1 - \bar{\alpha}_t$ (the noise scale at timestep t), the update rule becomes equivalent to iterative Tweedie-mean correction (see Eq. 5 above), where x_t^{k+1} is reinterpreted as the next estimated Tweedie mean given x_t^k for the distribution $\tilde{p}_t(x_t, c)$.

■ **Composable-diffusion.** Generating samples that satisfy multiple conditions $\{c_1, \ldots, c_K\}$ can be formulated as sampling from the joint distribution

$$\tilde{p}_0(x_0 \mid c_1, \ldots, c_K) \;\propto\; \prod_{k=1}^{K} p_0(x_0 \mid c_k). \tag{9}$$

To achieve this, Liu et al. (2022) proposed *composable diffusion*, which directly composes the output scores (predicted noises) from different conditional diffusion models using CFG during sampling.

Specifically, in the text-to-image setting, if the prompt C can be decomposed into constituent concepts $\{c_1, c_2, \ldots, c_K\}$, the outputs of the corresponding diffusion chains can be combined as

$$\tilde{\epsilon}_t^{\lambda,C} \;=\; \epsilon_t^{\phi} + \lambda_1\big(\epsilon_t^{c_1} - \epsilon_t^{\phi}\big) + \lambda_2\big(\epsilon_t^{c_2} - \epsilon_t^{\phi}\big) + \cdots + \lambda_K\big(\epsilon_t^{c_K} - \epsilon_t^{\phi}\big) \tag{10}$$

where ϵ_t^{ϕ} denotes the unconditional score, and λ_k controls the guidance strength for concept c_k. The next sample is then predicted via the Tweedie formulation:

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}\,\hat{x}_{\text{tweedie}}\Big[\tilde{\epsilon}_t^{\lambda,C}\Big] \;+\; \sqrt{1-\bar{\alpha}_{t-1}}\,\tilde{\epsilon}_t^{\lambda,C}. \tag{11}$$

Although this approach is model-agnostic and conceptually simple, its performance is often poor, since the above linear composition of scores is incorrect or doesn't correspond to the score of the diffusion forward distribution $\tilde{p}_t(x_t \mid c_1, \ldots, c_K)$ at any timestep $t > 0$ (Du et al., 2024).

In summary, both the correction-based approach and the composable diffusion idea can be interpreted as different ways of approximating Tweedie-means $\hat{x}_{tweedie}[\epsilon_t^{\lambda,c}]$ at time t.

3 CO3: CONTRASTING CONCEPTS TO COMPOSE

We aim to combine the strengths of correction-based approaches and composable diffusion. On one hand, correction-based methods are powerful and explicitly improve compositional alignment, but they are subject to the complex gradient of the base model. On the other hand, composable diffusion is fully model-agnostic, but suffers from poor performance because linear score composition is not consistent with the diffusion forward process.

To take advantage of both, assume that prompt C can be decomposed into constituent concepts $\{c_1, c_2, \ldots, c_K\}$. We propose an explicit *Concept Contrasting* corrector distribution based on our hypothesis on mode overlap discussed in Sec. 1. Specifically, we define the corrector distribution at each timestep t as:

$$\tilde{p}_t(x_t, C) \propto \frac{p_t(x_t \mid C)^{w_0}}{\prod_{k=1}^{K} p_t(x_t \mid c_k)^{w_k}}, \tag{12}$$

where $\{w_0, w_1, \ldots, w_K\}$ are composition weights. As discussed in Sec. 1, this corrector steers the generation toward regions where the distribution $p_t(x_t \mid C)$ is high, while simultaneously avoiding regions of overlap with the individual concept distributions $\{p_t(x_t \mid c_k)\}_k$. Observe that this encourages the model to generate samples that satisfy all concepts in C without over-emphasizing any single concept. We present the **CO3** corrector pseudo code in Algorithm 1.

To sample from the unnormalized probability distribution in Eq. 12, an well-known approach is to compose the concept distributions $\{p_t(x_t \mid c_k)\}_k$ in the space of score functions (Liu et al., 2022; Du et al., 2024). We break-away from this approach and compose the distribution in the Tweedie mean space; we show how this offers a more general framework for composition. In Tweedie-denoised space we define composition as:

$$\tilde{x}_{tweedie} = w_0\, \hat{x}_{tweedie}[\epsilon_t^{\lambda, C}] \;+\; w_1\, \hat{x}_{tweedie}[\epsilon_t^{\lambda, c_1}] \;+\; \ldots \;+\; w_K\, \hat{x}_{tweedie}[\epsilon_t^{\lambda, c_K}], \tag{13}$$

where $w_0 > 0$ and $w_1, \ldots, w_K < 0$ are concept weights. Note: $\hat{x}_{tweedie}[\epsilon_t^{\lambda, c_k}]$ is the Tweedie mean corresponding to the CFG composed noise prediction $\epsilon_t^{\lambda, c_k}$ for concept c_k with guidance weight λ.

How should composition weights be chosen? We analyze the effect of different weight assignments to Eq. 13, in particular, how the constraint on the concept weights w_i influences both the theoretical interpretation and the empirical behavior of the compositional Tweedie mean.

Algorithm 1 DDIM with **CO3** Corrector

Require: x_T, number of DDIM steps T, timestep threshold T_c, prompt C, set of concepts $\{c_k\}$

1: **for** $t = T$ down to 1 **do**
2: **if** $t > T_c$ **then**
3: $x_t^{(2)} \leftarrow$ **CO3**$(x_t, C, c_1, \ldots, c_k)$ ▷ Proposed correction
4: $x_t \leftarrow x_t^{(2)}$
5: **end if**
6: $x_t \leftarrow$ DDIM_step(x_t) ▷ Inverse DDIM
7: **end for**

A valid compositional Tweedie mean $\tilde{x}_{tweedie}$ in Eq. 13 must be one that can be expressed in the definition given in equation 5 as DDIM goes through a series of Tweedie means for image generation as described in section 2. Lemma 1 shows that this is guaranteed only when the weights satisfy the normalization condition $\sum_i w_i = 1$.

Lemma 1. *Let $\hat{x}_{tweedie}[\epsilon_t^{\lambda, c}] := x_t - \sigma_t\, \epsilon_t^{\lambda, c}$ be the tweedie mean from CFG composed noise $\tilde{\epsilon}_t^{\lambda} = \epsilon_t^{\phi} + \lambda(\epsilon_t^{C} - \epsilon_t^{\phi})$ for some λ. Let, $\tilde{\hat{x}}_{tweedie}$ be the composed tweedie-mean defined as $\tilde{\hat{x}}_{tweedie} = \sum_k w_k \hat{x}_{tweedie}[\epsilon_t^{\lambda, c_k}]$. Then,*

*a) **CO3**-corrector: $\tilde{\hat{x}}_{tweedie}$ can be expressed in the form of a tweedie-mean at time t, i.e. $\tilde{\hat{x}}_{tweedie} = x - \sigma_t\, \tilde{\epsilon}_t^{\tilde{\lambda}, C}$ if and only if $\sum_k w_k = 1$. Here $\tilde{\lambda} = \lambda$ and CFG composed noise $\tilde{\epsilon}_t^{\tilde{\lambda}, C} = \epsilon_t^{\phi} + \lambda(\sum_k w_k \epsilon_t^{c_k} - \epsilon_t^{\phi})$.*

*b) **CO3**-resampler: $\tilde{\hat{x}}_{tweedie} = -\lambda\, \sigma_t\, \sum_k w_k \epsilon^{c_k}$ is weighted noise if and only if $\sum_k w_k = 0$.*

Remarks: ❶ See Appendix section A.2 for proof. Lemma 1 a) shows that when $\sum_k w_k = 1$, denoted as ***CO3**-corrector*, the composed Tweedie-mean corresponds to the CFG composed noise $\tilde{\epsilon}^{\tilde{\lambda}, C} = \epsilon^{\phi} + \lambda(\sum_k w_k \epsilon^{c_k} - \epsilon^{\phi})$ with the same guidance scale λ. This preserves the relative guidance

strength between unconditional and conditional scores as proposed by classifier-free guidance (CFG) (see Eq. 4). This is in contrast to composable diffusion methods (Liu et al., 2022; Du et al., 2024) which use arbitrary guidance weights $\{\lambda_i\}$ for composing different concepts (see Eq. 10); hence this does not preserve the CFG form. We believe this explains, at least partly, why composable diffusion often produces out of manifold samples. See Sec. A.5 in Appendix for more discussion.

❷ Lemma 1 b) shows that when the weights sum to zero, denoted as *resampler*, composed Tweedie mean is independent of the current sample x_t and is only a function of the concept noises. Instead of correcting the current sample x_t, it replaces x_t with a weighted combination of the concept noises. In other words, when $\sum_k w_k = 0$, the composition is sampling from noise, and only valid at $t = T$ where $x_T \sim \mathcal{N}(0, 1)$.

To further characterize the behavior of the composition under the two conditions $\sum_k w_k = 1$ and $\sum_k w_k = 0$, we evaluate the performance of different weighting strategies across timesteps t of the diffusion chain. Figure 2 reveals complementary dynamics between the resampler and the corrector. Specifically, the resampler ($\sum_k w_k = 0$) is effective at high t values near T, but its effectiveness diminishes as the denoising process progresses toward lower t. Although resampling is not theoretically valid for $t < T$, we observe practical improvements until approximately $t \approx 0.9T$. In contrast, the corrector ($\sum_k w_k = 1$) exhibits gradual improvement with increasing timesteps and saturates around $t \approx 0.75T$. These findings invite a hybrid strategy into **CO3**: apply resampling during early timesteps ($t > T_R$), followed by correction until a threshold ($T_C < t < T_R$), beyond which further gains are marginal.

The pseudocode for the zero-sum-weight resampler (**CO3**-resampler) and the unit-sum-weight (**CO3**-corrector) corrector is provided in Algorithm 2 and Algorithm 3, respectively. Importantly, the same diffusion denoiser model can be reused to compute the individual concept scores under different input conditions c_k, thereby *eliminating the need for any costly backward passes to obtain gradients*. Furthermore, the framework operates without reliance on model-specific architectural details, making it *fully model-agnostic* and *gradient-free*.

Closeness-Aware Concept Weight Modulation: To make the **CO3** corrector more adaptive, we anchor the weight w_0 (w_0=1.0 in Algo. 2 and $w_0 = 2.0$ in Algo. 3), and assign weights to each concept based on how *close* the current noise prediction ϵ^C is, to the noise corresponding to that concept, ϵ^{c_k}. Intuitively, if a sample looks closer to concept c_k than to all others, we want to penalize c_k more strongly (i.e., give it a larger negative weight), while reducing the strength of the other concepts. This encourages the sampler to move away from the nearest mode, preventing collapse toward one dominant concept. Formally, let $d_k = \|\epsilon^C - \epsilon^{c_k}\|$ denote the distance between the current sample and the mode of concept c_k. We convert distances $\{d_k\}_{k=1}^K$ into *affinity scores* using an exponential kernel:

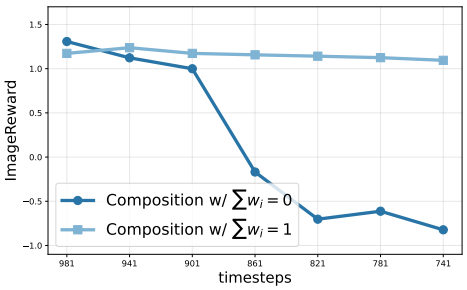

Figure 2: Characterization of Resampler and Corrector steps. Resampling is more powerful at high t while the Corrector improves slowly with more timesteps and saturates.

$$a_k = \exp(-\beta\, d_k), \qquad \beta > 0 \tag{14}$$

so that concepts closer to the current sample (smaller d_k) receive higher affinity. These affinity scores are then normalized to define the weights:

$$w_k = -\frac{a_k}{\sum_{j=1}^K a_j}, \tag{15}$$

which ensures each weight is negative and the total sum satisfies $\sum_{k=1}^K w_k = -1$. As a result, concepts that are closer to the current sample receive stronger negative weights, while farther concepts are down-weighted in proportion to their distance. This weight modulation scheme samples from the whole hyper-plane $\sum_{k=0}^K w_k = 1$ (for **CO3**-correction) or $\sum_{k=0}^K w_k = 0$ (for **CO3**-resampling) instead of a fixed weights for the entire course of generation.

Algorithm 2 CO3-resampler ($\sum w_k = 0$)

Require: x_t, denoiser ψ, $\bar{\alpha}_t$, $w_{0:K}$, guidance λ

for $p = 1$ to P **do**
  $\epsilon^{c_0} = \psi(x_t, C)$, $\epsilon^{c_k} = \psi(x_t, c_k)$,
    $k = 1, \ldots, K$
  $\hat{x}_{tweedie}[\epsilon_t^{\lambda, c_k}] = x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{\lambda, c_k}$
  $\tilde{\hat{x}}_{tweedie} = -\lambda\sqrt{1-\bar{\alpha}_t}\sum_{k=0}^{K} w_k \epsilon^{c_k}$
  ▷ Projection onto noise space
  $x_t^{(2)} = \tilde{\hat{x}}_{tweedie} + \sqrt{1-\bar{\alpha}_t}\,\epsilon_t^{\phi}$
  ▷ Uncond. manifold correction
  $x_t \leftarrow x_t^{(2)}$
end for
return x_t

Algorithm 3 CO3-corrector ($\sum w_k = 1$)

Require: x_t, denoiser ψ, $\bar{\alpha}_t$, $w_{0:K}$, guidance λ

for $p = 1$ to P **do**
  $\epsilon^{c_0} = \psi(x_t, C)$, $\epsilon^{c_k} = \psi(x_t, c_k)$,
    $k = 1, \ldots, K$
  $\tilde{\hat{x}}_{tweedie} = \sum_{k=0}^{K} w_k \hat{x}_{tweedie}[\epsilon_t^{\lambda, c_k}]$
  ▷ Tweedie mean composition
  $r = \frac{\|\hat{x}_{tweedie}^{c_0}\|_2}{\|\tilde{\hat{x}}_{tweedie}\|_2}$ ▷ mean normalization
  $\tilde{\hat{x}}_{tweedie} \leftarrow \tilde{\hat{x}}_{tweedie} * r$
  $x_t^{(2)} = \tilde{\hat{x}}_{tweedie} + \sqrt{1-\bar{\alpha}_t}\,\epsilon_t^{\phi}$
  $x_t \leftarrow x_t^{(2)}$
end for
return x_t

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation details: We implement our method on SDXL (Podell et al., 2023) using a 50-step DDIM sampler with guidance scale $\lambda = 5.0$. **CO3** is applied during the first 20% of denoising steps ($T_c = 0.8T$), with the initial three steps reserved for resampling. For weight modulation, we set $\beta = 0.8$. All experiments are performed on a single NVIDIA A6000 GPU (48GB). Additional hyperparameter details are provided in the Appendix section A.3.

Evaluation benchmark and metrics: We evaluate our method on the benchmark dataset from Chefer et al. (2023), which includes three categories of prompts: Animal–Animal, Animal–Object, and Object–Object. While the benchmark is limited to these cases, **CO3** naturally extends to more complex prompts involving arbitrary numbers of concepts. For evaluation, we adopt BLIP-VQA (Huang et al., 2023) and ImageReward (Xu et al., 2023), both widely used in recent works (taihang Hu et al., 2024; Hu et al., 2024) to measure the faithfulness of generated images to input prompts. Notably, ImageReward is a learned reward model trained on human preferences, capturing aspects such as prompt adherence, aesthetics, and overall quality. For both metrics, higher scores indicate better alignment.

Comparison methods: We compare **CO3** against state-of-the-art approaches, including optimization-based correction methods (Attend-and-Excite (Chefer et al., 2023), SynGen (Rassin et al., 2023), Divide-and-Bind (Liu et al., 2023), ToME (taihang Hu et al., 2024)), the composable generation method Composable Diffusion (Liu et al., 2022), and the noise-optimization method InitNO (Guo et al., 2024). All optimization-based approaches are architecture-dependent and require gradient computations at inference, with InitNO being model-agnostic but still incurring costly back-propagation. In contrast, only Composable Diffusion and our proposed **CO3** are purely sampling-based, model-agnostic, and completely gradient-free.

4.2 EXPERIMENTAL RESULTS

Quantitative Comparison:

Table 1 demonstrates the extent to which **CO3**, despite being model-agnostic and optimization-free, competes or outperforms optimization-based and model-agnostic baselines on ImageReward. Gains are substantive in the Animal–Animal and Animal–Object categories. On BLIP-VQA, **CO3** delivers the best performance in these two categories and improves over the base SDXL on the Object category, which contains prompts of the form: "a [color1] [object1] and a [color2] [object2]."

Unlike methods such as ToMe taihang Hu et al. (2024), SynGen (Rassin et al., 2023), Attend-and-Excite (Chefer et al., 2023), and Divide-and-Bind Liu et al. (2023), which rely on model-specific

	Properties			BLIP-VQA			ImageReward		
	training-free	gradient-free	model-agnostic	animals	animals-objects	objects	animals	animals-objects	objects
SD1.5(Rombach et al., 2022)	✓	✓	-	0.3239	0.5958	0.2730	-0.2733	0.4262	-0.5521
Attend-excite(Chefer et al., 2023)	✓	✗	✗	0.6980	0.7865	0.5155	0.8244	1.2380	0.8741
SynGenRassin et al. (2023)	✓	✗	✗	0.3348	0.7689	**0.6595**	-0.2445	0.9838	0.7315
Divide & Bind	✓	✗	✗	0.7201	0.8399	0.5887	0.8499	1.2516	0.8134
InitNO(Guo et al., 2024)	✓	✗	✓	0.7264	0.7998	0.5406	1.0082	1.3927	**1.1383**
SDXLPodell et al. (2023)	✓	✓	-	0.6950	0.8654	0.4926	0.7820	1.5574	0.6789
SynGen(SDXL)Rassin et al. (2023)	✓	✗	✗	0.6816	0.8578	0.4652	0.6998	1.5622	0.6441
ToMe(taihang Hu et al., 2024)	✓	✗	✗	0.6257	0.8808	0.6440	0.3895	1.5736	1.0118
Compose-diffusion(Liu et al., 2022)	✓	✓	✓	0.2846	0.5656	0.4529	-1.1399	-0.2068	-0.0955
CO3(Ours)	✓	✓	✓	**0.7441**	**0.8878**	0.5146	**1.2342**	**1.6744**	**1.0158**

Table 1: Quantitative comparison of different methods on compositional generation tasks. We evaluate the generated images using two metrics: BLIP-VQA and ImageReward. Top performing model is highlighted in **Black** and 2nd best in Blue. Higher the score the better.

architectures or explicit attention-based binding losses, **CO3** is architecture-independent, sampling-based, and free of explicit subject–attribute binding. Interestingly, by simply targeting "pure" modes under $p(x \mid C)$, **CO3** discovers layouts that better capture multi-object relations (Animals–Animals, Animal–Objects) and finds high-probability regions that preserve subject–attribute bindings (Objects). This performance gain empirically validates our hypothesis on mode overlap and also highlights **CO3**'s effectiveness in correcting the compositional limitations of SDXL.

Qualitative Comparison: Figure 3 compares **CO3** with several baselines. Columns 1–2 show results from Divide-Bind (Liu et al., 2023) and Attend-and-Excite (Chefer et al., 2023) (based on SDv1.5 (Rombach et al., 2022)), while the remaining columns compare against Composable Diffusion (Liu et al., 2022), SynGen (Rassin et al., 2023), and ToMe (taihang Hu et al., 2024). We illustrate common issues in compositional alignment—concept missing, attribute mixing, and object binding—across all categories. Divide-Bind and Attend-and-Excite enhance the attention layer of the base model, but often yield cropped or incomplete concepts (rows 1, 3). Composable Diffusion suffers from missing or merged concepts (rows 2, 3). SynGen and ToMe perform well on Object prompts of the form "a [color1] [object1] and a [color2] [object2]," but still exhibit overlapping or mixed concepts (rows 2, 3; columns 5, 6). In contrast, **CO3** achieves stronger object binding in the Animals and Animal–Objects categories while also improving results on Objects.

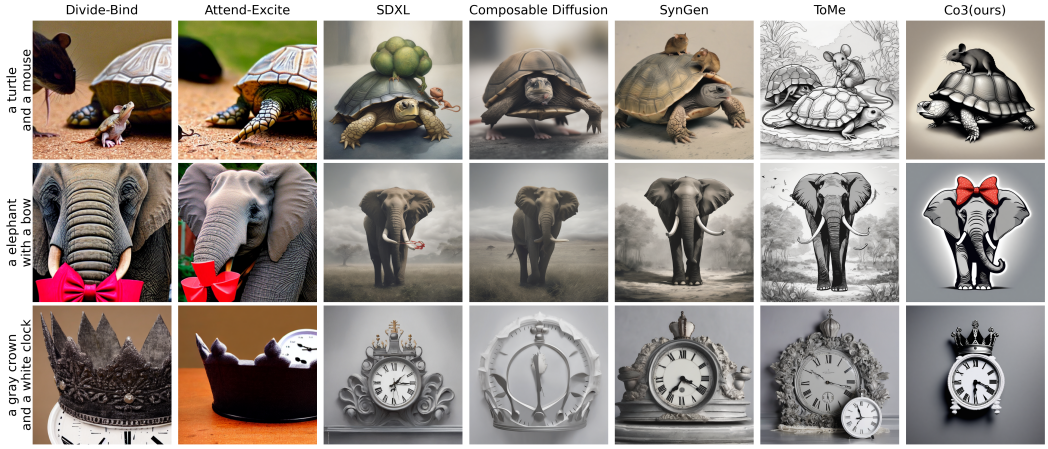

Figure 3: Qualitative comparison of different methods on simpler prompts.

Figure 4 extends qualitative results to complex prompts from the Complex category of Huang et al. (2023), comparing **CO3** with the top-performing SDXL models (Table 1). The first row shows prompts involving spatial relations (e.g., "on," "to the left of"), while the second row depicts more descriptive prompts with multiple interacting concepts. SynGen and ToMe successfully bind concepts to their subjects but often miss concepts or relations. By mitigating dominant-concept bias, **CO3** remains effective even for complex prompts with multiple relations.

Ablations Studies are preformed on the key components of **CO3**, starting from the base SDXL model (Table 2). Adding resampling in the first three DDIM steps substantially improves aver-

Figure 4: Qualitative comparison of **CO3** with competing methods on complex prompts.

age performance, particularly for the Animals and Animal–Objects categories. Since image layout is largely determined in the early timesteps, **CO3**-resampler iteratively explores better initial noise configurations that align with the prompt. Incorporating **CO3**-corrector over the next seven steps further improves prompts requiring fine-grained details such as color or texture, boosting performance on the Objects category in both ImageReward and BLIP-VQA. Additional gains come from the Closeness-Aware Concept Weight Modulation strategy, which consistently benefits all categories. We hypothesize that as the diffusion trajectory evolves, the latent state often drifts closer to one concept mode c_i; the modulation mechanism assigns c_i a more negative weight, repelling the latent and mitigating concept dominance (Tunanyan et al., 2023) commonly observed in stable-diffusion.

	BLIP-VQA			ImageReward			
Methods	animals	animals-objects	objects	animals	animals-objects	objects	Avg.
SDXL	0.6951	0.8654	0.4926	0.7820	1.5474	0.6789	0.8435
+ resampling	0.7351	0.8666	0.4528	1.0881	1.6755	0.8263	0.9407
+ Corrector	0.7177	0.8794	0.4796	1.0630	1.6429	0.8949	0.9463
+ weight-modulation	0.7441	0.8878	0.5146	1.2342	1.6744	1.0158	1.0118

Table 2: Ablations study conducted on the **CO3** method. Starting with the base-model SDXL we progressively add different components.

Comparison with Optimization-Free Approaches: To further demonstrate **CO3**'s model-agnostic nature and generalization to other base models, we evaluate it on PixArt-Σ (Chen et al., 2024). We compare against both the base model and the sampling-based, optimization-free Composable Diffusion (Liu et al., 2022). Unlike Composable Diffusion, which composes diffusion chains within the DDIM prediction step, **CO3** applies composition as a corrector. As shown in Fig. 5, Composable Diffusion suffers from missing or mixing concepts, while **CO3** consistently preserves all concepts with correct bindings in the output.

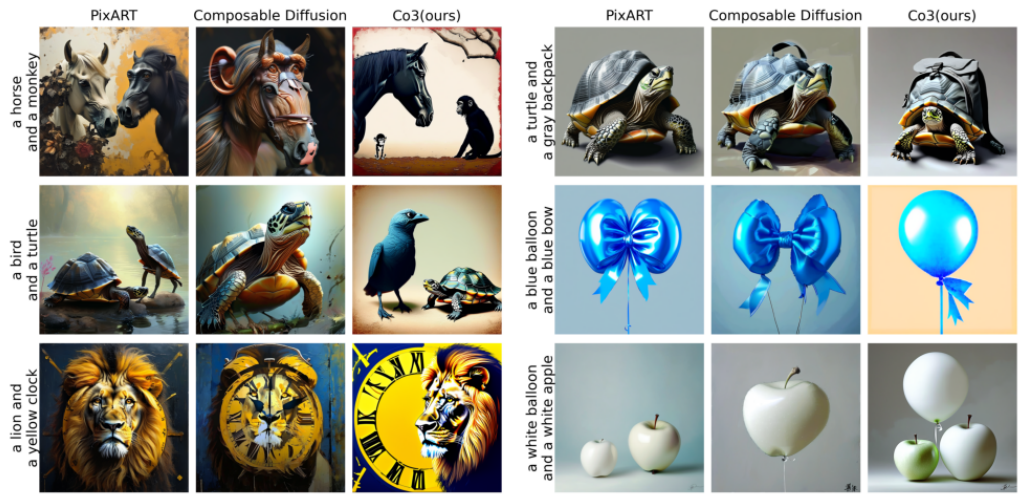

Figure 5: *Model Agnostic behavior*: Qualitative comparison of generation from PixART-Σ(Chen et al., 2024) base diffusion model, PixART-Σ + **CO3**, and PixART-Σ + Composable Diffusion.

5 Related Work

Optimization-based correction approaches: ***(1) Text-embedding optimization*** works like T2I-Zero variants (Tunanyan et al., 2023), and style or prompt inversion (Li et al., 2023a), improve semantic alignment by optimizing or restructuring the text representation during inference. These methods tweak or invert token embeddings so that the denoising trajectory better reflects entity–attribute bindings or multi-concept prompts.
(2) Attention-map optimization works (Zhang et al., 2024; Meral et al., 2024; Tumanyan et al., 2023; Hertz et al., 2022) directly manipulate the cross- and self-attention maps during sampling. These methods typically inject constraints or losses that (i) boost token–region correspondence for each entity, (ii) reduce overlap between different concepts, or (iii) preserve early layout information across timesteps. Attend-and-Excite (Chefer et al., 2023) increases the activation of object tokens. Divide-and-Bind (Liu et al., 2023) maximizes total variation to elicit distinct excitations for multiple objects and aligns attribute–entity attention maps, while A-STAR (Agarwal et al., 2023) further reduces cross-token overlap and preserves early attention signals. SynGen (Rassin et al., 2023) leverages syntactic parses to penalize mismatched attention overlaps, ensuring linguistic binding between entities and modifiers. InitNO (Guo et al., 2024) optimizes the initial noise so that sampling begins in more favorable regions that yield stronger, less-conflicted attention.

Composable generation works: view conditional diffusion models as energy/score functions, enabling algebraic composition. Composed-Diffusion (Liu et al., 2022) frames score-composition under CFG and demonstrates test-time generalization, but suffers from concept mixing and missing. Follow-ups propose training-free and model-agnostic methods like energy-parameterized diffusion and Metropolis/MCMC-corrected samplers that markedly improve multi-condition generation (Du et al., 2024), but their performance is often poor (Chefer et al., 2023; Feng et al., 2022). Kwon & Ye (2025) uses a composition similar to our resampler but not in the context of a corrector. Instead they use multiple DDIM forward-backward steps to sample initial noise.

Layout augmentation image generation: ***(1) Layout-to-image methods:*** (Xie et al., 2023; Phung et al., 2024; Kim et al., 2023; Zhao et al., 2023) use a strategy with explicit spatial priors—-boxes, masks, or region texts to bridge text and image: training-free controllers steer attention so objects materialize in designated regions. Other works add instance-level handles, for fine-grained placement and attributes across multiple entities (Wang et al., 2024). In parallel, fine-tuning approaches inject layout channels into the backbone (Li et al., 2023b; Mou et al., 2023; Zhang et al., 2023).
(2) LLM-augmented methods: Collectively, these works leverage LLM reasoning/representations to better bridge linguistic structure and the denoising trajectory by (i) decomposing complex prompts into regional sub-tasks (Yang et al., 2024; Hu et al., 2024) (ii) infer layouts from text (Qu et al., 2023), and (iii) act as stronger text encoders or timestep-aware adapters (Saharia et al., 2022b).

6 Conclusion & Future Work

We propose a gradient-free, model-agnostic image composition method that switches between noise-resampling and correction. Our *Concept Contrasting* corrector is based on the hypothesis that composition is degraded by "problematic" modes—those that overlap with individual concept modes, resulting in strong alignment with particular concepts while suppressing others. Our corrector employs a Closeness-Aware Weight Modulation scheme that emphasizes "pure" modes where all concepts coexist without any dominating ones. Crucially, we attribute shortcomings of past composition methods to the choice of composition weights and show that only when the weights sum to 1, the interpretation of Tweedie-mean correction and the CFG guidance strength are preserved. Results demonstrate that **CO3** outperforms baseline approaches for several diffusion model choices and serves as a *plug-and-play* module offering improved compositional generation performance.

While our corrector's closeness-aware weight modulation suppresses the concept mode overlap, the noise/score landscape is strongly biased by the training paradigms of diffusion models, i.e., the quantity and quality of the multi-concept bindings in the training set. This problem also manifests when presented with more unrealistic prompts, resulting in poor text/concept alignment. As shown in Fig. 7, while our work offers a significant boost to multi-concept compositional generation, clearly, there are failure cases that are not yet addressed. However, we believe that more advanced energy-based compositional samplers—that explicitly take into account the probabilities of the landscape during sampling—can further help in avoiding such cases. We hope to investigate these special cases in the future. We also hope to explore applications beyond composition using the **CO3** corrector.

7 Reproducibility Statement

All method implementations, inference pipelines are included in the supplementary material. This includes instructions for environment setup, dependencies, and reproducible random seeds. The datasets used in our experiments are publicly available. We provide detailed descriptions of the dataset in Sec. 4. Hyperparameters, inference schedules, and evaluation pipelines are described in Sec. 4, with further details in the Appendix A.3. Our project website is available at https://anomxyz26.github.io/co3-anom-web/ and implementation code is available at https://github.com/anomxyz26/co3-anom

References

Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2283–2293, 2023.

Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. Separate-and-enhance: Compositional finetuning for text-to-image diffusion models. In *SIGGRAPH (Conference Paper Track)*, pp. 103, 2024. URL https://doi.org/10.1145/3641519.3657527.

Hila Chefer, Omer Ratzon, Roni Paiss, and Lior Wolf. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques*, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ : $Weak-to-strongtrainingofdiffusiontransformerfor4ktext-to-imagegeneration$, 2024. URL.

Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DsEhqQtfAG.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc, 2024. URL https://arxiv.org/abs/2302.11552.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2022. arXiv:2212.05032.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. *arXiv preprint arXiv:2404.04650*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ACM SIGGRAPH Conference Proceedings (preprint)*, 2022. arXiv:2208.01626.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Deep Generative Models and Downstream Applications*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. URL https://arxiv.org/abs/2006.11239. v2.

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

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

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Kim_Dense_Text-to-Image_Generation_with_Attention_Modulation_ICCV_2023_paper.pdf.

Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Nayau9fwXU.

Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation, 2025. URL https://arxiv.org/abs/2410.05591.

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Li_GLIGEN_Open-Set_Grounded_Text-to-Image_Generation_CVPR_2023_paper.pdf.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision – ECCV 2022*, volume 13677 of *Lecture Notes in Computer Science*, pp. 325–343. Springer, 2022. 10.1007/978-3-031-19790-1_26.

Xiang Liu, Yongqi Li, Shizhe Zhou, Bo Li, Xiangtai Li, and Zhenzhong Ma. Divide and bind: Improving compositional text-to-image generation with concept binding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. 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 (CVPR)*, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. URL https://arxiv.org/abs/2302.08453.

Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Phung_Grounded_Text-to-Image_Synthesis_with_Attention_Refocusing_CVPR_2024_paper.pdf.

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

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In Asli Celikyilmaz and Tsung-Hsien Wen (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108, Online, July 2020. Association for Computational Linguistics. 10.18653/v1/2020.acl-demos.14. URL https://aclanthology.org/2020.acl-demos.14/.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. *arXiv preprint arXiv:2308.05095*, 2023. URL https://arxiv.org/abs/2308.05095.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022a. URL https://arxiv.org/abs/2204.06125.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b. URL https://papers.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=tRRWoa9e80.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1931, 2023.

Hayk Tunanyan, Xuan Li, Szymon Pirk, Leonidas Guibas, Alexei Efros, Tali Avraham, Shai Bagon, and Tali Dekel. Multi-concept t2i-zero: Tweaking only the text embeddings for zero-shot multi-concept image generation. *arXiv preprint arXiv:2310.07419*, 2023.

Xudong Wang, Sai Saketh Rambhatla, Rohit Girdhar, Ishan Misra, and Trevor Darrell. Instanced-iffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2402.03290.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2307.10816.

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

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2401.11708.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.pdf.

Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, and Kenji Kawaguchi. Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models. In *European Conference on Computer Vision*, pp. 70–86. Springer, 2024.

Peiang Zhao, Han Li, Ruiyang Jin, and S. Kevin Zhou. Loco: Locally constrained training-free layout-to-image synthesis. *arXiv preprint arXiv:2311.12342*, 2023. URL https://arxiv.org/abs/2311.12342.

A Appendix

A.1 Broader Impacts

CO3 improves compositional text-to-image generation by adjusting the sampling distribution of diffusion models. At the same time, it introduces certain risks. The method could be misused to create deceptive or misleading visuals, contributing to the spread of misinformation. When applied to depictions of public figures, it may compromise personal privacy. Moreover, the automatically produced content can raise concerns around copyright and intellectual property.

A.2 Composition in Tweedie-denoised space for a valid CFG

Classifier free guidance: In CFG, to sample from $p(x|c)$, we compose the conditonal and unconditonal predicted noise at each time step t as:

$$\epsilon_t^{\lambda,c} = \epsilon_t^{\phi} + \lambda(\epsilon_t^{c} - \epsilon_t^{\phi}) \tag{16}$$

where ϵ_t^c is the noise predicted for the distribution $p_t(x_t|c)$ at time t, i.e., $\nabla_{x_t} \log p_t(x_t|c) = -\frac{\epsilon_t^c}{\sqrt{1-\bar{\alpha}_t}}$. Then we use this composed noise to sample the next step as,

$$\hat{x}_{t,0}^{\lambda,c} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{\lambda,c}}{\bar{\alpha}_t}, \quad x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_{t,0}^{\lambda,c} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t^{\lambda,c} \tag{17}$$

Let, $\hat{x}_{tweedie}^{\lambda,c} := x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{\lambda,c}$ be the tweedie mean from CFG composed noise at time t. Then we can rewrite the above equation as,

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}\hat{x}_{tweedie}^{\lambda,c} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t^{\lambda,c} \tag{18}$$

Composition of CFG scores: (Liu et al., 2022) introduced the idea composable-diffusion where they compose scores from different diffusion models or conditional distributions to generate images from the composed distribution. They assume that, $p(x_0|C) = \prod p(x_0|c_i)$ and proposed the modified CFG composition for each t as,

$$\tilde{\epsilon}_t^{\lambda,C} = \epsilon_t^{\phi} + \lambda_i(\sum_i \epsilon_t^{c_i} - \epsilon_t^{\phi}) \tag{19}$$

The issue with above composition is that for arbitrary λ_i, the composed CFG noise doesn't satisfy the CFG equation in equation 16, i.e., $\tilde{\epsilon}_t^{\lambda,C} \neq \epsilon_t^{\lambda,C}$ for any λ in equation equation 16 where $\epsilon_t^{\lambda,C}$ is the noise predicted for the distribution $p_t(x_t|C)$ at time t.

To conretize this, we state the following simple result.

Lemma 2. *Let's define CFG as a function at time t as, $f_{CFG}(\epsilon;\lambda) = \epsilon^{\phi} + \lambda(\epsilon - \epsilon^{\phi})$. Then for any λ and $K > 1$, $\sum_k w_k f_{CFG} = f_{CFG}(\sum_k w_k \epsilon^k)$ only if $\sum_k w_k = 1$.*

Proof. Let, $\tilde{\epsilon}_t^{\lambda} = \sum_k w_k f_{CFG}$. Then,

$$\begin{aligned}
\tilde{\epsilon}_t^{\lambda} &= \sum_k w_k(\epsilon^{\phi} + \lambda(\epsilon^k - \epsilon^{\phi})) \\
&= \sum_k w_k\epsilon^{\phi} + \lambda(\sum_k w_k\epsilon^k - \sum_k w_k\epsilon^{\phi}) \\
&= \epsilon_{\phi}\sum_k w_k + \lambda(\sum_k w_k\epsilon^k - \epsilon^{\phi}\sum_k w_k) \\
&= \epsilon_{\phi} + \lambda(\sum_k w_k\epsilon^k - \epsilon^{\phi}) \quad \text{if } \sum_k w_k = 1 \\
&= f_{CFG}(\sum_k w_k\epsilon^k) \quad \text{if } \sum_k w_k = 1
\end{aligned}$$

If $\sum_k w_k \neq 1$, then $\tilde{\epsilon}_t^{\lambda} \neq f_{CFG}(\sum_k w_k\epsilon^k)$. □

Remark: One immediate implication of this lemma is that, whenever we sample from a composed distribution with the CFG composed score of the form $\tilde{\epsilon}^{\lambda} = w_1\epsilon^{\lambda,1} + w_2\epsilon^{\lambda,2} + \cdots + w_K\epsilon^{\lambda,K}$, then $\tilde{\epsilon}^{\lambda}$ corresponds to the CFG noise for the distribution $p(x|C) = \prod_i p(x \mid c_i)^{w_i}$ only when $\sum_i w_i = 1$.

Lemma 3. *Let $\hat{x}_{tweedie}[\epsilon_t^{\lambda,c}] := x_t - \sigma_t\ \epsilon_t^{\lambda,c}$ be the tweedie mean from CFG composed noise $\tilde{\epsilon_t}^{\lambda} = \epsilon_t^{\phi} + \lambda(\epsilon_t^{C} - \epsilon_t^{\phi})$ for some λ. Let, $\tilde{\hat{x}}_{tweedie}$ be the composed tweedie-mean defined as $\tilde{\hat{x}}_{tweedie} = \sum_k w_k \hat{x}_{tweedie}[\epsilon_t^{\lambda,c_k}]$. Then,*

*a) **CO3**-corrector: $\tilde{\hat{x}}_{tweedie}$ can be expressed in the form of a tweedie-mean at time t, i.e. $\tilde{\hat{x}}_{tweedie} = x - \sigma_t\ \tilde{\epsilon}_t^{\tilde{\lambda},C}$ if and only if $\sum_k w_k = 1$. Here $\tilde{\lambda} = \lambda$ and CFG composed noise $\tilde{\epsilon}_t^{\tilde{\lambda},C} = \epsilon_t^{\phi} + \lambda(\sum_k w_k \epsilon_t^{c_k} - \epsilon_t^{\phi})$.*

*b) **CO3**-resampler: $\tilde{\hat{x}}_{tweedie} = -\lambda\ \sigma_t\ \sum_k w_k \epsilon^{c_k}$ is weighted noise if and only if $\sum_k w_k = 0$.*

Proof. a) The proof can be obtained by simplifying the $\hat{x}^{comp}_{tweedie}$ expression and using direct application of Lemma 2. We have,

$$\begin{aligned}
\tilde{\hat{x}}_{tweedie} &= \sum_k w_k \hat{x}^{\lambda,c_k}_{tweedie} \\
&= \sum_k w_k (x_t - \sigma_t\ \epsilon_t^{\lambda,c_k}) \\
&= x_t \sum_k w_k - \sigma_t \sum_k w_k \epsilon_t^{\lambda,c_k} \\
&= x_t \sum_k w_k - \sigma_t \sum_k f_{CFG}(\epsilon_t^{c_k}) \quad (\text{ from Lemma 2}) \\
&= x_t - \sigma_t\ f_{CFG}(\sum_k w_k \epsilon^{c_k}) \quad \text{if } \sum_k w_k = 1
\end{aligned}$$

So, we have $\tilde{\hat{x}}_{tweedie} = x_t - \sigma_t\ \tilde{\epsilon_t}^{\lambda,C}$ where $\tilde{\epsilon_t}^{\lambda,C} = f_{CFG}(\sum_k w_k \epsilon_t^{c_k}) = \epsilon_t^{\phi} + \lambda(\sum_k w_k \epsilon_t^{c_k} - \epsilon_t^{\phi})$. If $\sum_k w_k \neq 1$, then $\tilde{\hat{x}}_{tweedie}$ cannot be expressed in the form $x_t - \sigma_t\ \tilde{\epsilon_t}^{\lambda,C}$ for any λ and $\tilde{\epsilon_t}^{\lambda,C}$.

b) From the 4-th equality of part a)

$$\begin{aligned}
\tilde{\hat{x}}_{tweedie} &= x_t \sum_k w_k - \sigma_t \sum_k w_k \epsilon_t^{\lambda,c_k} \\
&= -\sigma_t \lambda \sum_{k=1}^{K} w_k \epsilon_t^{k} \text{ (from Lemma 2)}
\end{aligned}$$

□

A.3 More Implementation Details

A.3.1 CO3 Implementation Details

For the results in Table 1 we use $T_c = 10$(number of time-steps to correct), $T_r = 3$(number of resampling steps) and $P = 5$(number of corrector iteration). For concept aware weight modulation, we use exponential kernel with $\beta = 0.8$ while anchoring w_0 at 1.0 for **CO3**-resampler and 2.0 for **CO3**-corrector.

We use Stanza (Qi et al., 2020) to parse the prompts. We parse the prompts to extract different noun chunks and filter each of them to remove articles and adjectives. The remaining proper noun is used as concept in **CO3**. For example, if C is "a black cat and a brown dog", we consider c_1 ="cat" and c_2 ="dog".

A.3.2 Baseline Methods

Attend-Excite (Chefer et al., 2023), Divide-Bind (Liu et al., 2023), InitNo (Guo et al., 2024), SynGen (Rassin et al., 2023) and Composable Diffusion (Liu et al., 2022) are methods based on SD1.5.

We run experiments on these models using their publicly available code. ToMe (taihang Hu et al., 2024) is SDXL based we use their publicly available code-base implementation. We also adapted Composable Diffusion and SynGen to SDXL.

A.4 Time Complexity Analysis

We conduct an experiment to analyze the time-complexity of each SDXL based methods. The Table 3 indicates the average time (in secs) taken by the model per sample. The experiments are conducted on NVIDIA A100 gpu.

Method	Inference Steps	Time Cost (sec)	Animals	Animals-Objects	Objects
SDXL	50	7.20	0.6950	0.8654	0.4926
Composable Diffusion	50	9.80	0.2846	0.5656	0.4529
SynGen(SDXL)	50	11.48	0.6816	0.8578	0.4652
ToMe	50	16.58	0.6257	0.8808	0.6440
CO3(ours)	50	19.9	0.7441	0.8878	0.5146

Table 3: Comparison of different methods across efficiency and compositionality metrics.

A.5 Comparison of compositions in score space

In this section we compare our composition framework with Composable-Diffusion. As already described in section 3, we propose composition in Tweedie-denoised space as

$$\tilde{x}_{tweedie} = w_0\, \hat{x}_{tweedie}[\epsilon_t^{\lambda,C}] \;+\; w_1\, \hat{x}_{tweedie}[\epsilon_t^{\lambda,c_1}] \;+\; \ldots \;+\; w_K\, \hat{x}_{tweedie}[\epsilon_t^{\lambda,c_K}] \tag{20}$$

$$\tag{21}$$

which leads to

$$\tilde{x}_{tweedie} = x - \sigma_t\, \tilde{\epsilon}_t^{\tilde{\lambda},C} \text{ iff } \sum w_{k=0}^{K} = 1 \tag{22}$$

$$\text{where } \tilde{\epsilon}_t^{\tilde{\lambda},C} = \epsilon_t^{\phi} + \lambda(\sum_k w_k \epsilon_t^{c_k} - \epsilon_t^{\phi}) \tag{23}$$

Contrast this with the Composable-Diffusion's noise/score composition:

$$\tilde{\epsilon}_{t,compdiff}^{\lambda,C} \;=\; \epsilon_t^{\phi} + \lambda_1\big(\epsilon_t^{c_1} - \epsilon_t^{\phi}\big) + \lambda_2\big(\epsilon_t^{c_2} - \epsilon_t^{\phi}\big) + \cdots + \lambda_K\big(\epsilon_t^{c_K} - \epsilon_t^{\phi}\big) \tag{24}$$

For arbitrary weights λ_k, this cannot be expressed in the from equation 23. This proves that $\tilde{\epsilon}_{t,compdiff}^{\lambda,C}$ doesn't lead to a valid Tweedie-mean.

A.6 More Visualizations

A.6.1 **CO3** Failure Cases

Despite correcting for the "problematic" modes, there still remain open challenges in image composition as shown in Figure 7. The score and, hence, the correction landscape is heavily influenced by the training schemes employed in diffusion model training, i.e., the quantity and quality of the multi-concept bindings in the training set. In addition, the usage of unrealistic prompts perhaps not encountered in training also results in poor text/concept alignment. We leave this investigation for the future.

A.7 Ablations on Hyperparameters

In this section, we analyze the contribution of the following five factors to the performance of our **CO3** corrector.

Notation recap: β is the exponential decay factor in the *affinity scores*; λ scales the *composed_score* for CFG; `num_resampling` is the number of resampling steps at the start of diffusion; `num_ts` is the number of early timesteps where the corrector is used; `num_steps` is the number of iterations per corrector application.

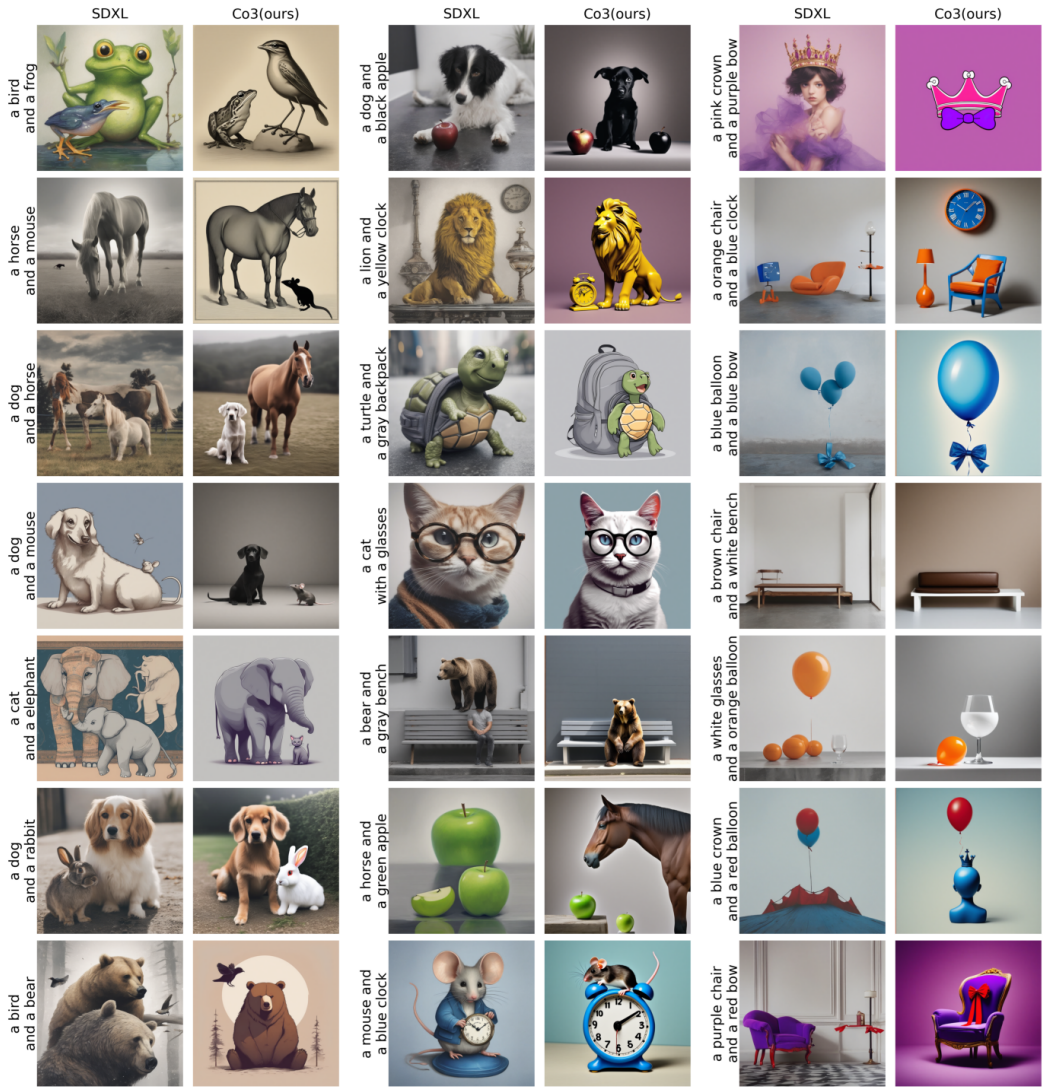

Figure 6: Qualitative results of SDXL base diffusion model, and SDXL + **CO3** for *animal-animal*,*animal-object*, and *object-object* prompt categories

Figure 7: Failure scenarios of PixART-Σ base diffusion model, PixART-Σ + **CO3**, and PixART-Σ + Composable Diffusion.

A.7.1 NUMBER OF RESAMPLING STEPS

Table 4 shows results for different resampling steps. We observe that using more resampling steps can blur concept separation early in the diffusion process rather than helping to disentangle concepts.

Table 4: **num_resampling** sweep. Within each group, only num_resampling changes; all other settings are identical.

Frozen settings	num_resampling	ImageReward (↑)			BLIP-VQA (↑)		
		Animals	Animals_Objects	Objects	Animals	Animals_Objects	Objects
$\beta = 0.8$, *num_ts=7*, $\lambda = 0.8$, *num_steps=5*	3	1.2218	1.7020	0.9405	0.7375	0.8811	0.4782
	4	1.1705	1.7031	0.9512	0.7259	0.8828	0.4696
$\beta = 0.3$, *num_ts=7*, $\lambda = 0.8$, *num_steps=5*	3	1.1484	1.6915	0.9005	0.7240	0.8810	0.4691
	4	1.1297	1.6948	0.8996	0.7145	0.8774	0.4697
$\beta = 1.1$, *num_ts=7*, $\lambda = 0.8$, *num_steps=5*	3	1.2566	1.7062	0.9916	0.7422	0.8813	0.4834
	4	1.1927	1.7164	0.9934	0.7292	0.8835	0.4811

A.7.2 CORRECTOR APPLICATION TIMESTEPS

In **CO3**, the corrector is applied during the first num_ts diffusion steps. Table 5 shows that larger num_ts generally yields better results, especially for *animals* and *objects*. This indicates the corrector remains beneficial beyond only the earliest timesteps.

Table 5: **num_ts** sweep with all other settings fixed: $\beta = 1.1$, num_resampling=3, num_steps=5, $\lambda = 0.8$. (Best values are bold)

num_ts	ImageReward (↑)			BLIP-VQA (↑)		
	Animals	Animals_Objects	Objects	Animals	Animals_Objects	Objects
4	1.1862	1.7136	0.8885	0.7430	0.8783	0.4630
6	1.2510	**1.7117**	0.9750	0.7432	0.8801	0.4812
7	1.2566	1.7062	0.9916	0.7422	0.8813	0.4834
10	**1.3149**	1.6816	**1.0095**	**0.7576**	**0.8832**	**0.5074**

A.7.3 CORRECTOR ITERATIONS

Table 6 reports the effect of increasing the number of corrector iterations. Despite setting num_ts=4, raising num_steps reduces performance substantially, suggesting that applying the corrector only at the very beginning is not sufficient; employing it until much later in the diffusion trajectory is more effective.

Table 6: **num_steps**. Within each group, only num_steps changes; all other settings are identical.

Fixed settings	num_steps	ImageReward (↑)			BLIP-VQA (↑)		
		Animals	Animals_Objects	Objects	Animals	Animals_Objects	Objects
(a) $\beta = 1.1$, *num_ts=4, num_resampling=2,* $\lambda = 0.9$	7	**1.2622**	**1.7129**	**0.9959**	**0.7542**	**0.8843**	**0.4707**
	10	1.2452	1.6994	0.9753	0.7448	0.8809	0.4706
	15	1.2151	1.6701	0.9437	0.7318	0.8781	0.4556
(b) $\beta = 1.1$, *num_ts=4, num_resampling=2,* $\lambda = 0.8$	7	**1.2755**	**1.7049**	**0.9424**	**0.7538**	0.8762	0.4611
	10	1.2503	1.6969	0.9191	0.7446	**0.8798**	**0.4617**
	15	1.2480	1.6658	0.9063	0.7356	0.8750	0.4577

A.7.4 EXPONENTIAL DECAY FACTOR β OF *affinity scores*

We vary β in equation 14 from 0.3 to 1.1. Tables 7 and 8 demonstrate that the performance generally increases with larger β, with the strongest gains in the *animals* and *objects* categories, while a few settings exhibit minor regressions.

Table 7: **Exponential decay factor** β with all other settings fixed: `num_resampling`=3, `num_steps`=5, $\lambda = 0.8$. We report two blocks: `num_ts`=6 and `num_ts`=7. (Best values for each block are bold)

num_ts	β	**ImageReward** (↑)			**BLIP-VQA** (↑)		
		Animals	Animals_Objects	Objects	Animals	Animals_Objects	Objects
6	0.3	1.1458	1.7016	0.8801	0.7262	0.8793	0.4613
	0.6	1.1945	1.7069	0.9010	0.7317	0.8786	0.4645
	0.7	1.2043	1.7109	0.9152	0.7310	0.8800	0.4706
	0.8	1.2157	1.7024	0.9156	0.7398	0.8776	0.4751
	0.9	1.2305	1.7146	0.9511	0.7400	0.8809	0.4766
	1.0	1.2282	**1.7183**	0.9596	0.7350	**0.8808**	0.4770
	1.1	**1.2510**	1.7117	**0.9750**	**0.7432**	0.8801	**0.4811**
7	0.3	1.1484	1.6915	0.9005	0.7240	0.8809	0.4691
	0.8	1.2218	1.7020	0.9405	0.7375	0.8811	0.4782
	1.1	**1.2566**	**1.7062**	**0.9916**	**0.7422**	**0.8813**	**0.4834**

Table 8: **Exponential decay factor** β with all other settings fixed: `num_resampling`=3, `num_steps`=5, $\lambda = 0.9$. We report two blocks: `num_ts`=6 and `num_ts`=7. (Best values for each block are bold.)

num_ts	β	**ImageReward** (↑)			**BLIP-VQA** (↑)		
		Animals	Animals_Objects	Objects	Animals	Animals_Objects	Objects
6	0.3	1.1437	1.7020	0.9518	0.7248	0.8816	0.4764
	0.7	1.2051	1.7130	0.9860	0.7407	0.8851	0.4788
	0.8	1.2037	1.7110	0.9916	0.7388	0.8832	**0.4894**
	0.9	1.2003	1.7159	0.9883	**0.7416**	**0.8842**	0.4891
	1.1	**1.2098**	**1.7180**	**0.9956**	0.7371	0.8841	0.4855
7	0.3	1.1273	1.6938	0.9490	0.7274	0.8848	0.4796
	0.8	1.2142	1.7000	0.9937	0.7363	0.8845	**0.4952**
	1.1	**1.2331**	**1.7046**	**1.0148**	**0.7496**	**0.8864**	0.4945