
TILT: Test-Time Reward Alignment via Distribution Tilting for Compositional Generation

Anonymous Authors¹

Abstract

Recent advances in powerful text-to-image generation models have made it increasingly important to develop test-time methods that modify the sampling trajectory to produce images more faithful to complex compositional prompts. We present **TILT**, a training-free framework for compositional text-to-image generation via test-time reward alignment. We interpret compositional failures as overlap modes between joint and single-concept distributions, and define a pure-mode reward that favors samples where all concepts are jointly present while remaining close to the pretrained model. This yields a KL-constrained objective with a closed-form tilted target distribution and principled guiding steps for diffusion sampling. Our framework also recovers CO3 as a special case, giving theoretical grounding to prior methods for compositional generation using heuristic correctors. Experiments on prompts from T2ICompBench show that our method improves compositional alignment while preserving image quality compared to previous baselines.

1. Introduction

State-of-the-art text-to-image (T2I) diffusion models (Rombach et al., 2022; Saharia et al., 2022b; Podell et al., 2023; Huang et al., 2023) can faithfully render complex and intricate prompts. However, these models frequently fails in generating faithful images when prompted with complex, compositional prompts. While pinpointing the exact cause is difficult, a plausible explanation lies in *imperfect compositional generalization*. When a prompt combines concepts in a complex and novel way, the model composes them based on what it has seen during training, which are often only the individual concepts or familiar subsets of them. Due to data

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

imbalance or imperfect training, the model may develop particularly higher affinity toward certain concepts over others. At inference time, these affinities cause concepts to compete for dominance in the output, with the stronger one consistently dominating over other concepts.

This phenomenon of *concept dominance* has been studied through attention-based failures in the denoising process. Attend-and-Excite (Chefer et al., 2023) identifies where Stable Diffusion fails to generate one or more subjects in the prompt, and intervenes by strengthening cross-attention activations for subject tokens (Chefer et al., 2023). Structured Diffusion (Feng et al., 2022) similarly improves attribute binding and multi-object composition by manipulating cross-attention representations using linguistic structure. These methods leverage cross-attention signals to improve compositional generation. However, they are architecture-dependent and require access to internal attention maps.

An alternative strategy is to improve compositional generation by modifying the *target distribution* at inference time, without touching the model’s weights, so as to either ensure the presence of all concepts or to actively avoid samples exhibiting concept dominance. Composable Diffusion (Liu et al., 2022) proposes train-free method to alter the sampling step by combining concept-conditioned diffusion scores. A subsequent work (Du et al., 2024) shows that naive score composition can fail by pushing the samples from the learned manifold and proposes correction based on Markov chain Monte Carlo-based method for compositional generation.

In another line of work, CO3 (Dutta et al., 2026) and TweedieMix (Kwon & Ye, 2025) pursues a related direction via Tweedie-mean composition. CO3 (Dutta et al., 2026) explains concept dominance as the result of *mode overlap* between the joint prompt distribution and the individual concept distributions. As a remedy, they propose a corrector mechanism that steers generation toward a "concept-contrasting" distribution, one that emphasizes "pure" joint modes where all concepts coexist with balanced visual presence and suppresses modes that align too closely with any single concept. Although the mode-overlap hypothesis is intuitive and its method is empirically effective, the correction step itself remains a heuristic: it is not derived from a prin-

055 cipled objective, and it is unclear what terminal distribution
 056 the correction mechanism ultimately targets or what it in-
 057 herently optimizes. Given the effectiveness of the approach,
 058 the concept-contrasting distribution motivates for a more
 059 mathematically principled approach with deeper analysis.

060 We break away from the heuristic corrector and instead
 061 frame sampling from pure modes of the joint concept distri-
 062 bution as a *reward alignment problem*. Our central premise
 063 is that modern T2I models have already learned strong priors
 064 capable of generating high-quality concept-specific samples.
 065 We show that, *these priors can be combined to formulate*
 066 *an appropriate reward function which is intrinsic to the*
 067 *model. When this reward is optimized at inference time, the*
 068 *model can produce images with the desired compositional*
 069 *structure*.

071 Rather than requiring external supervision from another
 072 foundation model or fine-tuning the generative model it-
 073 self, our approach exploits this fact by reward aligning the
 074 sampling process, at inference time, to extract those good
 075 samples from the model’s existing distribution. Defining a
 076 suitable terminal reward which guides sample toward pure
 077 modes, we rigorously derive the intermediate objectives
 078 that should guide the generation process at each diffusion
 079 step, producing principled guidance signals in place of the
 080 heuristic corrector from prior works.

081 Our framework naturally gives rise to two complementary
 082 update algorithms that trade off efficiency and fidelity. The
 083 first update (**TILT-S**) is computationally efficient and well
 084 suited to early high-noise denoising steps, while the sec-
 085 ond update (**TILT-C**) provides more accurate concept-wise
 086 guidance at later low-noise steps where fine-grained com-
 087 positional binding becomes important. This motivates a hybrid
 088 scheme that switches between the two across diffusion time.
 089 We also show that CO3 is recovered as a special case of our
 090 framework, providing theoretical grounding for its empirical
 091 success. Empirically, our hybrid method achieves compar-
 092 able or stronger generation quality than prior approaches on
 093 multiple compositional generation benchmarks.

094 We can summarize the contributions as follows.

- 095 • We formalize multi-concept compositional generation as
 096 an test-time reward alignment problem, with pure-mode
 097 sampling as the intrinsic reward, and rigorously derive
 098 guidance objectives from this formulation (§3).
- 099 • We show that prior works can be considered as a special
 100 case of our framework, providing a principled justification
 101 for its empirical effectiveness (§3).
- 102 • Our framework yields two complementary guidance al-
 103 gorithms; a hybrid combining both gives comparable or
 104 outperforming performance compared with prior methods
 105 on compositional generation benchmarks (§4).

2. Background

2.1. Classifier-Free Guidance and Variants

In diffusion-based Text-to-Image (T2I) generation (Rom-
 bach et al., 2022; Saharia et al., 2022a; Ramesh et al., 2022),
 given the noisy latent x_t at timestep t , a denoised estimate
 can be derived using Tweedie’s formula:

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

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. In the DDIM sampler (Song et al., 2021), under
 the noise-free condition, the subsequent step deterministi-
 cally evolves \hat{x}_0 to x_{t-1} :

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(x_t, t|c). \quad (2)$$

Here, the same predicted noise ϵ_θ is reused, eliminating the
 renoisification step present in stochastic samplers such as
 DDPM(Ho et al., 2020).

In practice, most T2I models adopt *classifier-free guidance*
 (CFG) (Ho & Salimans, 2022), where we use convex com-
 bination of the conditional and unconditional scores as the
 final score to use during inference:

$$\epsilon_t^{\lambda,c} = \lambda \epsilon_\theta(x_t, t|c) + (1 - \lambda) \epsilon_\theta(x_t, t|\emptyset), \quad (3)$$

Then the denoising and DDIM steps proceed as before, but
 using $\epsilon_t^{\lambda,c}$ in place of $\epsilon_\theta(x_t, c, t)$.

CFG improves prompt alignment, but using the guided pre-
 diction in both the Tweedie estimate and the DDIM up-
 date can move the trajectory off the data manifold. CFG++
 (Chung et al., 2024a) addresses this using smaller strength
 to estimate the Tweedie mean, but using unconditional noise
 prediction to re-noisify it. To be more specific, standard
 CFG forms $\epsilon_t^{\lambda,c}$ and uses it both to estimate \hat{x}_0 and to
 propagate the noise component. CFG++ keeps the guided
 Tweedie estimate, but replaces the renoising direction by
 the unconditional prediction:

$$x_{t-1}^{\text{CFG++}} = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0 \left[\epsilon_t^{\lambda,c} \right] + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(x_t, t|\emptyset). \quad (4)$$

Thus, CFG++ still interpolates between unconditional and
 conditional denoised estimates, but the transport from x_t to
 x_{t-1} follows the unconditional diffusion manifold.

2.2. Composable Diffusion

Generating samples that satisfy multiple conditions $\{c_i\}$
 can be formulated as sampling from the joint distribution

$$\tilde{p}_0(x_0 | c_1, \dots, c_K) \propto p(x_0) \prod_{k=1}^K p_0(c_k | x_0). \quad (5)$$

To achieve this, Liu et al. (2022) proposed Composable Diffusion, which directly composes the score function from different conditional diffusion models during sampling.

Specifically,

$$\tilde{\epsilon}_t^{\lambda, C} = \epsilon_t^\phi + \lambda_1(\epsilon_t^{c_1} - \epsilon_t^\phi) + \lambda_2(\epsilon_t^{c_2} - \epsilon_t^\phi) + \dots + \lambda_K(\epsilon_t^{c_K} - \epsilon_t^\phi) \quad (6)$$

where ϵ_t^ϕ denotes the unconditional score, and λ_k controls the classifier free guidance strength for concept c_k . Then, the next sample is predicted via the usual DDIM step with Tweedie formulation:

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}} \hat{x}_0[\tilde{\epsilon}_t^{\lambda, C}] + \sqrt{1 - \bar{\alpha}_{t-1}} \tilde{\epsilon}_t^{\lambda, C}. \quad (7)$$

Although this approach is model-agnostic and conceptually simple, it cannot accurately generate images from complex prompts. This is because there does not exist a score of the diffusion forward distribution $\tilde{p}_t(x_t | c_1, \dots, c_K)$, at any timestep $t > 0$ (Du et al., 2024), that coincide with heuristically defined linear combination of scores.

2.3. Compositional Corrector for Diffusion

In regard of improving the composition in Diffusion models further, CO3 (Dutta et al., 2026) proposes to utilize compositional corrector during sampling. For each sample of particular timestep x_t , CO3 updates the sample using convex combination of $\hat{x}_{0, C}$ and $\{\hat{x}_{0, c_i}\}$. TweedieMix (Kwon & Ye, 2025) is another similar work that proposes correction mechanism using Tweedie’s formula. Rather than directly interpolating denoised predictions, both works construct compositional corrections in the estimated clean-sample space by combining score estimates associated with the full prompt and individual concept prompts. Unlike optimization-based compositional methods requiring additional training or LoRA finetuning, these methods can operate in a fully training-free setting during inference.

3. Method

We establish that compositional failures arise from overlap modes, formulate pure-mode sampling as reward alignment, derive a closed-form solution, and instantiate it via DPS-style guidance where the Jacobian choice serves as a design knob. CO3 emerges as a special case, and the framework generalizes to any modality.

Given a pretrained conditional score $s_\theta(x_t, t | c)$, a compositional prompt $C = \{c_1, \dots, c_K\}$, and concept conditionals $p^\theta(x | C)$ (joint) and $p^\theta(x | c_i)$ (per-concept), we define $\hat{x}_0(x_t)$ as the Tweedie posterior mean (with superscripts indicating conditioning: $\hat{x}_0^C, \hat{x}_0^{c_i}$). A *pure mode* is a sample with high joint likelihood and balanced per-concept likelihood; *concept dominance* is the opposite failure mode.

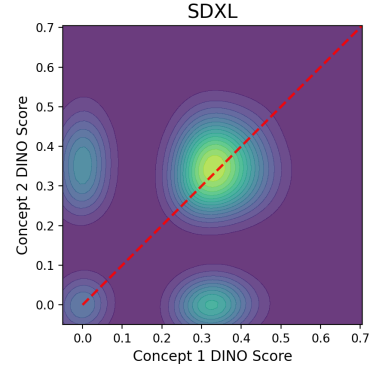


Figure 1. Histogram of alignment scores (DINOv2) between each sub-prompts c_1, c_2 and the generated images. The text prompts are in the form of "A [Animal] and a [Object]".

Empirical Evidence: Dominance emerges from Mode Overlap.

Figure 1 measures DINOv2 concept-quality scores for prompts of the form “a [Animal] and a [Object]” under SDXL. The empirical density concentrates near the axes $x=0$ and $y=0$, indicating that conditional samples disproportionately retain only one of the two concepts. Crucially, the modes near the axes are precisely the regions where the joint conditional density $p^\theta(x | C)$ overlaps with one of the marginals $p^\theta(x | c_i)$. We can summarize our analysis as follows:

Compositional failures in a pretrained T2I model are dominated by samples drawn from overlap modes. This is due to the joint conditional inadvertently coinciding with a single-concept conditional.

We want a sampling target whose mass concentrates on pure modes and avoids overlap modes. A natural construction reweights the joint by the inverse product of marginals:

$$\tilde{p}(x_0 | C) \propto \frac{p^\theta(x_0 | C)}{\prod_{i=1}^K p^\theta(x_0 | c_i)}. \quad (8)$$

The reweighting suppresses regions where any single $p^\theta(x | c_i)$ is large – exactly the overlap regions identified above – and preserves regions where mass under the joint is supported by all concepts simultaneously.

3.1. Pure-mode Sampling via Intrinsic Reward

Rather than approximate equation 8 directly, we pose the problem as test-time *reward alignment*: we seek the distribution p^* that maximizes a pure-mode reward while staying close to the pretrained joint conditional.

$$p^* = \arg \max_p \mathbb{E}_{x_0 \sim p} \left[\log \frac{p^\theta(x_0 | C)}{\prod_{i=1}^K p^\theta(x_0 | c_i)} \right] \quad (9)$$

s.t. $\text{KL}(p \| p^\theta(\cdot | C)) < \epsilon.$

The KL constraint anchors p^* to the support of the pretrained model, ensuring we do not drift onto out-of-distribution

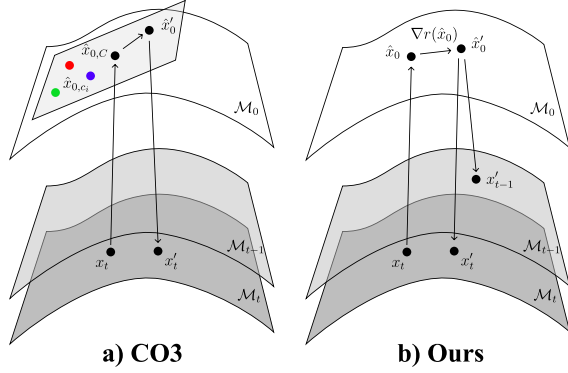


Figure 2. **Comparison of different test-time correction methods.** (a) **CO3** infers multiple denoised samples $\hat{x}_{0,C}$, $\{\hat{x}_{0,c_i}\}$, and then heuristically proposes new sample \hat{x}'_0 as a convex combination of them. (b) **TILT (Ours)** proposes new sample \hat{x}'_0 with the gradient of a reward function, using a more mathematically principled approach. We observe that this improves the empirical compositionality of the generated samples.

images.

Proposition 3.1 (Pointwise concept synergy and population-level Total Correlation). *For a fixed prompt C , applying Bayes’ rule to the reward equation 9 gives, modulo x -independent terms (Appendix A):*

(i) *Pointwise objective. The expected reward decomposes into the pointwise conditional total correlation,*

$$\mathbb{E}_{x \sim p^*}[R(x)] = \mathbb{E}_{x \sim p^\theta(\cdot|C)}[\iota_C(x)] + \text{const.}, \quad (10)$$

where $\iota_C(x) := \log \frac{p^\theta(C|x)}{\prod_i p^\theta(c_i|x)}$ is the pointwise conditional total correlation (pcTC), a per-prompt, per- x measure of how synergistically x explains the concepts under the model’s posterior.

(ii) *Population aggregate. Under a prompt distribution $C \sim p(C)$, the expected reward recovers the population-level conditional Total Correlation:*

$$\mathbb{E}_{C \sim p(C)} \mathbb{E}_{x \sim p_C^*}[R(x)] = \text{TC}(c_1; \dots; c_K | X) + \text{const.}, \quad (11)$$

where $\text{TC}(c_1; \dots; c_K | X) := \sum_{i=1}^K H(c_i | X) - H(c_1, \dots, c_K | X)$ is the conditional Total Correlation.

Interpretation. ① The sampler optimizes the pointwise primitive ι_C at each prompt to sample a suitable image. But the population-level reward equals the conditional Total Correlation of concepts given the sampled images (across the prompts). We therefore retain the operational benefit of a per-prompt signal while inheriting a clean population-level information-theoretic guarantee. ② If the model treats concepts as conditionally independent given x then $\iota_C \equiv 0$ and the reward collapses to a prior-likelihood term. The reward is informative *only* when the model couples concepts through shared visual structure, which is exactly the regime where compositional reasoning is non-trivial.

3.2. Relaxation to KL constraint

Relaxing the KL constraint with multiplier $\lambda > 0$ yields the soft-constrained loss

$$\mathcal{J}(p, \lambda) = -\mathbb{E}_{x_0 \sim p} \left[\log \frac{p^\theta(x_0|C)}{\prod_{i=1}^K p^\theta(x_0|c_i)} \right] + \lambda \text{KL}(p \| p^\theta(\cdot | C)), \quad (12)$$

whose minimizer is available in closed form. Setting $\beta := 1/\lambda$,

$$p^*(x_0) = \frac{1}{Z} p^\theta(x_0 | C) \exp(\beta R(x_0)), \quad (13)$$

where $R(x_0) := \log \frac{p^\theta(x_0|C)}{\prod_{i=1}^K p^\theta(x_0|c_i)}$ now becomes the reward function. Intuitively, we are targeting to generate samples from a reward-tilted distribution p^* .

Equation 13 defines p^* at $t=0$, whereas a diffusion sampler must operate at intermediate times $t > 0$. Optimizing the above reward function requires solving the PF-ODE to evaluate likelihood at x_0 and then backpropagating through it to update x_t —an extremely computationally extensive task. It’s equivalent to optimizing the quantity (Yeh et al., 2025): $\mathbb{E}_{p(x_0|x_t)}[R(x_0)]$, which depends on the terminal model likelihood and is not directly tractable in a diffusion sampler. The remaining question is how to construct a surrogate at $t > 0$ that induces samples from p^* .

3.3. Diffusion Sampling with Reward Tilting

To answer the above question, we take inspiration from the braod literature of inverse problems (Chung et al., 2022; 2024b; He et al., 2023). We introduce a binary observation $O = 1$ to denote the event of observing a sample with high reward (i.e. high $R(x_0)$). Then, likelihood $p(O=1 | x_0, C) \propto \exp(-\mathcal{L}(x_0))$ with $\mathcal{L}(x_0) := -R(x_0)$. Bayes’ rule on the noised state yields

$$\begin{aligned} \nabla_{x_t} \log p(x_t, C | O=1) &= \underbrace{\nabla_{x_t} \log p(x_t | C)}_{\text{prior score}} + \underbrace{\nabla_{x_t} \log p(O=1 | x_t, C)}_{\text{guidance}} \\ &\approx s_\theta(x_t, t | C) - \nabla_{x_t} \mathcal{L}(\hat{x}_0(x_t)). \end{aligned} \quad (14)$$

Using chain rule, this guidance term can be calculated with Jacobian-vector product:

$$\nabla_{x_t} \mathcal{L}(\hat{x}_0(x_t)) = \frac{\partial \hat{x}_0}{\partial x_t}^\top \cdot \nabla_{\hat{x}_0} \mathcal{L}(\hat{x}_0), \quad (15)$$

This observation reduces the design space to the choice of Jacobian approximation, which unifies the subsequent derivations.

We provide two instantiations of the proposed method.

TILT-S: Shared-Jacobian guidance. Use a single Jacobian $J_{x_t}(\hat{x}_0)$ —evaluated on the joint-conditional Tweedie

estimate – and apply the chain rule once to equation 14:

$$\begin{aligned} \nabla_{x_t} \log p(x_t, C \mid O=1) \\ \approx s_\theta(\hat{x}_0, t \mid C) + J_{x_t}(\hat{x}_0)^\top \left\{ \beta s_\theta(\hat{x}_0, 0, C) \right. \\ \left. - \frac{\beta}{K} \sum_{i=1}^K s_\theta(\hat{x}_0, 0, c_i) \right\}. \end{aligned} \quad (16)$$

The approximation in Eqn 15 makes the intractable likelihood calculation at $t = 0$ to simple model forward pass to evaluate s_θ . This variant requires only one backward pass per step and uses the joint-conditional geometry for all concepts, but it can lose fidelity when per-concept directions differ substantially.

TILT-C: Per-concept Jacobian guidance. Alternatively, if we employ chain rule and differentiate each term through its own Tweedie path, we get:

$$\begin{aligned} \nabla_{x_t} \log p^\theta(x_t, C \mid O = 1) \\ \approx s_\theta(x_t, t \mid C) + \beta J_{x_t}(\hat{x}_0^C)^\top s_\theta(\hat{x}_0^C, 0) \\ - \frac{\beta}{K} \sum_{i=1}^K J_{x_t}(\hat{x}_0^{c_i})^\top s_\theta(\hat{x}_0^{c_i}, 0). \end{aligned} \quad (17)$$

This variant provides a tighter approximation because each concept contributes through its own diffusion path, at the cost of $K+1$ backward passes per step.

TILT-H: Hybrid Algorithm. The two instantiations have complementary profiles: near $t = T$ (high noise, early reverse steps), all concept scores are close since they must maintain Gaussian structure, making **TILT-S** efficient; at low noise (late steps), per-concept directions diverge and **TILT-C** is necessary for fidelity. We combine them using a noise-level schedule indexed by a switching threshold $\tau \in (0, T)$.

Algorithm 1 TILT-H: Hybrid pure-mode sampling

Require: pretrained scores s_θ , multi-prompt C , stopping time-index M , schedule $\{t_n\}$, Strength of reward alignment β , Switching threshold τ

- 1: $x_T \sim \mathcal{N}(0, I)$
 - 2: **for** $n = N, N-1, \dots, M+1$ **do**
 - 3: $\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t} \epsilon_\theta(x_{t_n}, t_n \mid C)}{\sqrt{\alpha_t}}$
 - 4: **if** $t_n > \tau$ **then**
 - 5: $\hat{x}_0 \leftarrow \hat{x}_0 + \nabla_{x_t} \mathcal{L}(\hat{x}_0(x_t))$ (Eq. 16, **TILT-S**)
 - 6: **else**
 - 7: $\hat{x}_0 \leftarrow \hat{x}_0 + \nabla_{x_t} \mathcal{L}(\hat{x}_0(x_t))$ (Eq. 17, **TILT-C**)
 - 8: **end if**
 - 9: $x_{t_{n-1}} \leftarrow \text{DDIM}(\hat{x}_0; s_\theta(x_{t_n}, t_n \mid C))$
 - 10: **end for**
 - 11: **return** x_0
-

Comparison with CO3. Note that CO3 constructs a corrector update that is, in spirit, a step toward equation 8, but it is introduced as a heuristic and does not specify (i) what distribution it ultimately samples from, or (ii) what objective the corrector minimizes. Figure 2 illustrates the key difference between CO3 and **TILT**: while CO3 constructs a corrected sample via a heuristic convex combination of multiple denoised predictions, **TILT** directly optimizes a reward-guided correction direction, yielding a more principled test-time update. Appendix B actually shows that CO3’s update is recovered from **TILT-S** by setting the diffusion Jacobian to identity and freezing score evaluation at (x_t, t) . Thus, CO3 appears as a special case of the proposed framework.

4. Experiments

4.1. Experimental Setup

Implementation details. All experiments are conducted with the Stable Diffusion XL (SDXL) base model, without any additional training or fine-tuning. Our method modifies only the inference procedure. We use the DDIM scheduler with 50 denoising steps and generate images at 1024×1024 resolution. For each prompt, we automatically decompose the full text prompt into concept-level sub-prompts using noun parsers and simple text preprocessing, such as removing leading conjunctions.

During sampling, we compute the standard multi-concept conditional prediction from the full prompt and concept-level predictions from the extracted sub-prompts. The proposed correction is applied only during the first few denoising steps, where global composition is typically determined. Unless otherwise specified, we correct the first 5 denoising steps, use one latent correction step per corrected timestep, and use 10 correction iterations at the initial timestep. In **TILT**, we use CFG with guidance scale 5.0 or CFG++ with guidance scale 0.8.

Evaluation benchmark and metrics. We evaluate compositional text-to-image generation using prompts from T2I-CompBench, a benchmark designed to assess whether generated images correctly satisfy multiple compositional attributes described in text prompts. The benchmark contains four categories of compositional prompts: *Color*, *Shape*, *Texture*, and *Complex*. The first three categories evaluate relatively localized attribute binding, such as assigning the correct color or texture to an object, while the *Complex* category evaluates more challenging multi-concept reasoning involving multiple objects, relations, and attributes simultaneously.

Following prior work, we evaluate generated images using four automatic metrics. **ImageReward** (Xu et al., 2023) provides a learned human preference score that captures

Table 1. Quantitative comparison on multi-concept prompts from T2ICompbench. We report ImageReward, CLIP, DINO, and BLIP-VQA scores. All experiments are done using SDXL (Podell et al., 2023) base model.

Method	ImageReward \uparrow				CLIP \uparrow				DINO \uparrow				BLIP-VQA \uparrow			
	Color	Shape	Texture	Complex	Color	Shape	Texture	Complex	Color	Shape	Texture	Complex	Color	Shape	Texture	Complex
CFG (Ho & Salimans, 2022)	0.6235	0.2748	0.4394	0.3073	0.3333	0.3131	0.3220	0.3107	0.2012	0.1677	0.2073	0.1203	0.5661	0.4806	0.5414	0.4165
Comp-Diff (Liu et al., 2022)	0.2322	0.0187	0.2429	0.0658	0.3253	0.3084	0.3205	0.3080	0.2207	0.1824	0.2283	0.1333	0.4548	0.4419	0.5314	0.3848
R2F (Park et al., 2024)	0.6179	0.2588	0.4333	0.3561	0.3322	0.3131	0.3214	0.3109	0.2016	0.1713	0.2108	0.1217	0.5815	0.4837	0.5422	0.4423
CFG++ (Chung et al., 2024a)	0.7642	0.3567	0.6053	<u>0.4467</u>	0.3377	<u>0.3212</u>	0.3275	<u>0.3152</u>	0.2227	0.1913	0.2346	0.1349	<u>0.6247</u>	0.5122	0.5809	<u>0.4536</u>
CO3 (Dutta et al., 2026)	0.9648	<u>0.4245</u>	0.4927	0.4406	0.3424	0.3195	0.3180	0.3125	0.2599	0.2137	<u>0.2470</u>	0.1508	0.6326	<u>0.5041</u>	0.5476	0.4761
TILT (Ours)	<u>0.8569</u>	0.4338	<u>0.5929</u>	0.4804	<u>0.3416</u>	0.3226	<u>0.3265</u>	0.3157	<u>0.2570</u>	<u>0.2106</u>	0.2493	<u>0.1479</u>	0.5770	0.5026	<u>0.5665</u>	0.4497



Figure 3. Qualitative comparison of text-to-image compositional generation methods on T2ICompBench prompts. The prompts, ordered by row, are drawn from the color, shape, texture, and complex categories. Our method shows improved text alignment with better compositional consistency compared to prior baselines.

overall image quality and prompt alignment. CLIP (Radford et al., 2021) and DINO (Oquab et al., 2023) measures the image-text semantic alignment and visual consistency. Furthermore, BLIP-VQA (Li et al., 2022) measures compositional correctness by querying generated images with attribute-specific questions derived from the prompt. All results are averaged over four random seeds.

Comparison methods and baselines. We compare our method against several guidance and compositional generation baselines built upon Stable Diffusion XL (SDXL) which are train-free, gradient-free and model-agnostic. Specifically, some of the important baselines: (1) CFG (Classifier-Free Guidance), the standard guidance method widely used in diffusion-based text-to-image generation; (2) Compos-

able Diffusion, which composes multiple concept-specific score functions to improve compositional alignment in text-to-image generation; (3) R2F, a compositional generation method designed to improve multi-concept fidelity and attribute binding; (4) CFG++, an improved variant of classifier-free guidance designed to provide more stable and accurate guidance behavior; and (5) CO3, a recent compositional generation framework that enhances compositional alignment with correction mechanism. All methods are evaluated under the same SDXL backbone and benchmark settings.

4.2. Quantitative Results

In Table 1, we report results on T2ICompBench, which evaluates more challenging multi-concept compositional prompts spanning color, shape, texture, and complex relational categories. Our method consistently achieves strong performance across both BLIP-VQA and ImageReward metrics. In particular, our method achieves the best ImageReward score on the Shape and Complex categories while remaining competitive with CO3 and CFG++ on BLIP-VQA. Notably, the improvement on the Complex category suggests that the proposed correction is particularly effective for prompts requiring simultaneous satisfaction of multiple attributes and object relationships. While some baselines achieve high BLIP-VQA scores by aggressively enforcing compositional constraints, they often exhibit reduced perceptual quality or unstable image structure. Our method instead provides a more balanced trade-off between compositional correctness and visual realism, indicating that test-time score correction can effectively improve compositional consistency without sacrificing the generative prior learned by SDXL.

4.3. Qualitative Comparison

Figure 3 shows qualitative results on T2ICompBench prompts covering color, shape, texture, and complex compositional categories. Across these examples, prior baselines often satisfy only part of the prompt, such as generating the correct object category while missing attribute binding, object count, material, or spatial relation. In contrast, our method better preserves the individual concepts and their associated attributes, producing images that more faithfully reflect the requested color, shape, texture, and relational constraints. These results suggest that our method using test-time correction improves compositional consistency while retaining the visual quality in T2I generation using SDXL.

5. Discussion

We presented TILT, a principled framework for compositional text-to-image generation based on test-time reward alignment. Rather than treating compositional correction as a heuristic manipulation of diffusion trajectories, we formulate a mathematically principled pure-mode compositional sampling as optimizing an intrinsic reward. This perspective leads to a closed-form target distribution and naturally yields guidance rules derived through diffusion posterior sampling. Within this framework, we show that CO3 emerges as a special case under specific approximations, thereby providing theoretical grounding for prior empirical observations. Experimentally, our method achieves strong performance across multiple compositional generation benchmarks while preserving perceptual quality and

remaining entirely training-free and model-agnostic.

Limitation. Although TILT provides a principled test-time reward alignment framework, it currently has some limitations. First, the proposed guidance relies on gradient-based updates through the reward objective, which introduces additional computational cost compared to gradient-free correction methods. This cost becomes more pronounced for the per-concept Jacobian variant, which requires multiple backward passes per denoising step. Second, optimizing model-likelihood gradients can be less stable than optimizing standard external rewards, especially due to the Jacobian-vector term in the guidance derivation. This instability may adversely affect the formation of individual concepts, which is reflected in relatively lower BLIP-VQA scores in some categories despite strong ImageReward performance.

Future Work. For future work, our formulation is modality-agnostic and depends only on conditional score estimation and factorizable conditioning variables. This suggests that pure-mode reward alignment may extend naturally beyond text-to-image generation to other compositional generative settings, including text-to-audio synthesis, molecular generation, and multi-attribute editing. We believe this perspective opens a promising direction toward general test-time alignment objectives for controllable generation across modalities, where compositional consistency can be enforced through intrinsic reward structure rather than task-specific supervision or retraining.

Impact Statement

This paper presents work whose goal is to advance the field of text-to-image generation. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Chefer, H., Ratzon, O., Paiss, R., and Wolf, L. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques*, 2023.
- Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.
- Chung, H., Kim, J., Park, G. Y., Nam, H., and Ye, J. C. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024a.
- Chung, H., Lee, S., and Ye, J. C. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *The Twelfth International Conference on Learning Rep-*

- resentations, 2024b. URL <https://openreview.net/forum?id=DsEhqQtfgAG>.
- Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc, 2024. URL <https://arxiv.org/abs/2302.11552>.
- Dutta, D., Chen, J., Rajagopalan, R., Wei, Y.-L., and Choudhury, R. R. Steer away from mode collisions: Improving composition in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.
- Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2022. arXiv:2212.05032.
- He, Y., Murata, N., Lai, C.-H., Takida, Y., Uesaka, T., Kim, D., Liao, W.-H., Mitsufuji, Y., Kolter, J. Z., Salakhutdinov, R., and Ermon, S. Manifold preserving guided diffusion, 2023. URL <https://arxiv.org/abs/2311.16424>.
- Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Deep Generative Models and Downstream Applications*, 2022.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. URL <https://arxiv.org/abs/2006.11239>. v2.
- Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.
- Kwon, G. and Ye, J. C. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation, 2025. URL <https://arxiv.org/abs/2410.05591>.
- Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.
- Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. 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. doi: 10.1007/978-3-031-19790-1__26.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Park, D., Kim, S., Moon, T., Kim, M., Lee, K., and Cho, J. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. *arXiv preprint arXiv:2410.22376*, 2024.
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Seyed Ghasemipour, S. K., Karagol Ayan, B., Mahdavi, S. S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.
- Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

440 Watanabe, S. Information theoretical analysis of multivari-
441 ate correlation. *IBM J. Res. Dev.*, 4(1):66–82, January
442 1960. ISSN 0018-8646. doi: 10.1147/rd.41.0066. URL
443 <https://doi.org/10.1147/rd.41.0066>.

444
445 Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang,
446 J., and Dong, Y. Imagereward: Learning and evaluat-
447 ing human preferences for text-to-image generation. *Ad-
448 vances in Neural Information Processing Systems*, 36:
449 15903–15935, 2023.

450 Yeh, P.-H., Lee, K.-H., and Chen, J.-C. Training-free diffu-
451 sion model alignment with sampling demons, 2025. URL
452 <https://arxiv.org/abs/2410.05760>.

453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494

A. Pointwise reward & Interpretation

This section derives what the pure-mode reward

$$R(x) = \log \frac{p^\theta(x | C)}{\prod_{i=1}^K p^\theta(x | c_i)} \quad (18)$$

actually quantifies. We follow the InfoNCE-style template: identify the *pointwise* (per-prompt, per- x) information-theoretic primitive that the sampler acts on (§A.1–§A.3), then show that aggregating this primitive across a prompt distribution recovers the standard population-level conditional Total Correlation (§A.4).

A.1. Decomposition via Bayes

Applying Bayes’ rule to each conditional in equation 18,

$$\begin{aligned} \log p^\theta(x | C) &= \log p^\theta(C | x) + \log p^\theta(x) - \log p^\theta(C), \\ \log p^\theta(x | c_i) &= \log p^\theta(c_i | x) + \log p^\theta(x) - \log p^\theta(c_i), \end{aligned} \quad (19)$$

and substituting into equation 18,

$$\begin{aligned} R(x) &= \log \underbrace{\frac{p^\theta(C | x)}{\prod_{i=1}^K p^\theta(c_i | x)}}_{\iota_C(x)} \\ &\quad + (1 - K) \log p^\theta(x) + \log \underbrace{\frac{\prod_{i=1}^K p^\theta(c_i)}{p^\theta(C)}}_{\kappa(C)}. \end{aligned} \quad (20)$$

The constant $\kappa(C)$ is independent of x .

A.2. The pointwise primitive: pcTC

We refer to

$$\iota_C(x) := \log \frac{p^\theta(C | x)}{\prod_{i=1}^K p^\theta(c_i | x)} \quad (21)$$

as the **pointwise conditional Total Correlation** at the prompt tuple C given x . This is the per-realization analogue of conditional Total Correlation, in the same way pointwise mutual information is the per-realization analogue of mutual information (Watanabe, 1960).

For a fixed prompt C , $\iota_C(x)$ quantifies how much more likely the concept tuple C is jointly under the model’s posterior at x than it would be if the per-concept posteriors at x were independent factors.

A.3. Per-prompt expected reward

The constrained optimum p^* solves

$$\max_p \mathbb{E}_{x \sim p}[R(x)] \quad \text{s.t.} \quad \text{KL}(p \| p^\theta(\cdot | C)) < \epsilon, \quad (22)$$

so for small ϵ , $p^*(x) \approx p^\theta(x | C)$ and we can take expectations under the latter up to $O(\epsilon)$:

$$\begin{aligned} \mathbb{E}_{x \sim p^*}[R(x)] &= \mathbb{E}_{x \sim p^\theta(\cdot | C)}[\iota_C(x)] \\ &\quad - (K - 1) \mathbb{E}_{x \sim p^\theta(\cdot | C)}[\log p^\theta(x)] \\ &\quad + \kappa(C) + O(\epsilon). \end{aligned} \quad (23)$$

Of the three terms, only the first is informative for sampling: the second is a C -dependent cross-entropy (constant in any optimization that varies x for fixed C), and the third is a pure constant. Hence, up to terms that do not steer x ,

$$\mathbb{E}_{x \sim p^*}[R(x)] = \mathbb{E}_{x \sim p^\theta(\cdot | C)}[\iota_C(x)] + \text{const}(C) + O(\epsilon). \quad (24)$$

This is the *per-prompt* statement in Proposition 3.1: the sampler pushes the probability mass toward x where pcTC is high.

A.4. Population-level interpretation: from pcTC to conditional TC

Compositional generation is evaluated on benchmarks containing many prompts. Aggregating equation 24 across a benchmark with empirical prompt distribution $p(C)$ gives a clean information-theoretic identity. Recall the conditional Total Correlation of the concept tuple given x ,

$$\text{TC}(c_1; \dots; c_K | X) = \mathbb{E}_{p(x)}[\mathbb{E}_{p(C|x)}[\iota_C(x)]] \quad (25)$$

Using $p(x)p(C | x) = p(x, C) = p(C)p(x | C)$, this rewrites equivalently as

$$\text{TC}(c_1; \dots; c_K | X) = \mathbb{E}_{C \sim p(C)}[\mathbb{E}_{x \sim p^\theta(\cdot | C)}[\iota_C(x)]] \quad (26)$$

Comparing to equation 24,

$$\mathbb{E}_{C \sim p(C)} \mathbb{E}_{x \sim p_C^*}[R(x)] = \text{TC}(c_1; \dots; c_K | X) + \text{const}. \quad (27)$$

Interpretation. ① Our per-prompt reward equation 10 aggregates into the population-level functional equation 11 when averaged across a benchmark of prompts. The pointwise primitive $\iota_C(x)$ is a local signal that can be evaluated at each x_t during diffusion; benchmarks evaluate the population summary $\text{TC}(\cdot | X)$ that this primitive aggregates to. ② TC vanishes when concepts are conditionally independent given X , and is maximized when the joint distribution concentrates on configurations where all concepts are jointly determined by X – precisely the regime where compositional reasoning is non-trivial.

B. CO3 as a special case of TILT-S

CO3’s corrector also composes scaled Tweedie-means. Concretely, with weights $w_0 = 1 + \beta$ and $w_k = -\beta/K$,

550 CO3 forms a weighted Tweedie-mean composition $\hat{x}^{\text{tw}} =$
 551 $w_0 \hat{x}_t^{\text{tw}}[\epsilon_t^C] + \sum_k w_k \hat{x}_t^{\text{tw}}[\epsilon_t^{c_k}]$ and re-noises with ϵ_t^ϕ , where
 552 $\hat{x}_t^{\text{tw}}[\epsilon] = x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon$ and $\epsilon_t^c := \epsilon_\theta(x_t, t, c)$.

553 The CO3 corrector update is recovered from TILT-S equa-
 554 tion 16 by simultaneously imposing two approximations:
 555

- 556 1. **(Identity Jacobian)** $J_{x_t}(\hat{x}_0) \approx I$ – discards the
 557 diffusion-time geometry of Tweedie’s posterior;
- 558 2. **(Time-frozen score)** $s_\theta(\hat{x}_0, 0 | c) \approx s_\theta(x_t, t | c)$ – eval-
 559 uates conditional and per-concept scores at the current
 560 noised state instead of at the Tweedie mean.
 561

562 Under (i) and (ii), the guidance term in equation 16 reduces
 563 to CO3’s correction direction (up to a scalar absorbed into
 564 β).
 565

566 This connection shows that CO3 is a specific approximation
 567 of the same objective. CO3 (i) discards the diffusion-time
 568 Jacobian and (ii) does not reroute scores through Tweedie’s
 569 posterior; both approximations are highly inaccurate at high
 570 noise levels, where the Tweedie mean is far from x_t and the
 571 Jacobian deviates strongly from I .
 572
 573
 574
 575
 576
 577
 578
 579
 580
 581
 582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593
 594
 595
 596
 597
 598
 599
 600
 601
 602
 603
 604