TILT: MODEL-INTRINSIC REWARD ALIGNMENT FOR COMPOSITIONAL DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Consider conditional generation $p(x \mid C = \{c_1, c_2, \dots c_k\})$ where C is a prompt composed of multiple concepts c_i. Diffusion models often struggle with compositional prompts, producing samples in which some concepts dominate while others are missing or weakly represented. Prior work attributes these failures to *mode collision*, where single-concept modes of $p(x \mid c_i)$ overlap with modes of the joint $p(x \mid C)$. To seek out collision-free modes of $p(x \mid C)$, or *pure modes*, corrector-based approaches have attempted to suppress collisions at intermediate diffusion times. However, local corrections are often heuristic and do not necessarily steer the generation to a *pure mode* in the final data space. Derived from a principled formulation, we present TILT (**T**est-time model-**I**ntrinsic reward a**L**ignment via **T**ilting), a training-free framework that poses eventual pure mode sampling as a reward for intermediate-time alignment. This reward offers valuable advantages: (1) it is intrinsic to the model, hence external reward models need not be trained by modality-specific datasets, (2) it yields a closed-form target under a variational approximation, which makes it realizable through standard diffusion sampling, and (3) it is interpretable, hence amenable to preference-based modifications. Project page: `https://tinyurl.com/yr5j3bwe`

1 INTRODUCTION

Diffusion models have been surprisingly successful across multiple domains, including image (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2023b; Huang et al., 2023b), audio (Liu et al., 2023; Copet et al., 2024), video (Blattmann et al., 2023; Ho et al., 2022), molecules, and proteins generation (Abramson et al., 2024; Xu et al., 2022; Watson et al., 2023; Corso et al., 2023). These are largely driven by *conditional sampling*, which lets users generate samples that satisfy a given condition, including text-to-image and text-to-audio generation. To generate samples that satisfy multiple conditions $c_{1:k}$, a common approach is to compose them into a single condition C and give it to the model as input. For text-to-image, say, C=`astronaut on moon`, with c_1=`astronaut`, c_2=`on`, and c_3=`moon`. The model is then expected to generate an image that satisfies all these conditions jointly.

Several studies have observed that diffusion models often suffer from compositional failures when conditioned on multiple concepts; the model confidently drops some concepts or fails to produce a coherent sample that satisfies all of them, *e.g.*, missing objects in text-to-image generation. Such failures are broadly described as *concept dominance*: during generation, concepts compete with each other and some dominate the output, so resulting samples miss or only weakly reflect the others.

Many approaches exist to mitigate the concept-dominance problem: model surgery approaches manipulate attention maps or model activations taihang Hu et al. (2024); Chefer et al. (2023); Rassin et al. (2023); guidance approaches steer sampling trajectories with an external reward models Tang et al. (2024); Novack et al. (2024), and correction-based approaches such as Dutta et al. (2026a); Kwon & Ye (2025) correct trajectories with a modified score function.

Among these, the recent corrector-based approaches have shown promising results. Dutta et al. (2026a) diagnose concept-dominance as *mode collisions* in the model's learned distribution. Under this hypothesis, the model's distribution creates overlaps (or *collisions*) with the modes of individual concept distributions during training. During inference, the denoiser's trajectory arrives into one of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

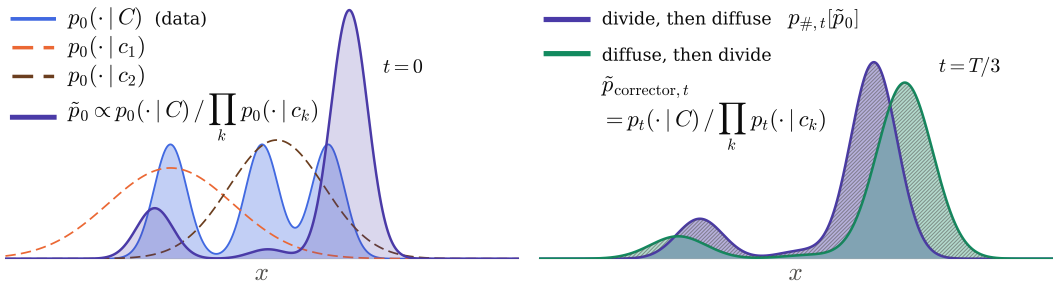


Figure 1: **The correct intermediate-time distribution. Left:** in data space ($t = 0$), the joint $p_0(\cdot \mid C)$ (blue) shares modes with individual concept distributions $p_0(\cdot \mid c_1)$ (orange) and $p_0(\cdot \mid c_2)$ (brown). The distribution $\tilde{p}_0$ (purple) describes the *pure modes* – the modes of the joint that no single concept explains. **Right:** at $t = T/3$, dark-purple is the diffusion pushforward of $\tilde{p}_0$ (*divide, then diffuse*), while green is the corrector distribution from Eq. 1, which follows *diffuse, then divide*. The two distributions diverge – a corrector defined at $t > 0$ therefore steers the sampler towards some other distribution, not the noised "pure mode" target.

these overlapping modes and creates concept-dominated samples. To avoid this, they aim to sample from the *pure modes* of the target distribution by designing a corrector of the form:

$$p_{t,\text{corrector}}(x) \propto \frac{p_t(x \mid C)}{\prod_{i=1}^{n} p_t(x \mid c_i)} \tag{1}$$

This corrector is used at each intermediate time t during denoising. The idea is to preserve modes of $p_t(x \mid C)$ so all the concepts contribute jointly, but suppress regions dominated by a single $p_t(x \mid c_i)$ in the denominator. This ratio is expected to reduce overlap, resulting in *pure modes*.

Although insightful, these approaches are heuristic. They construct the corrector to mitigate overlap in the intermediate distributions (at $t > 0$) while the actual overlap exists in the data-space (at $t = 0$). As Fig. 1 shows, noising the true target distribution $\tilde{p}_0$ (purple) to time, say $T/3$, is not the same as noising the individual distributions to $T/3$ and then diving them to form the corrector $\tilde{p}_t$ (green). The gap between the purple and green distributions grows significantly with higher number of dimensions, indicating that corrector-based approaches steer their sampler to the wrong target.

We depart from correction-based approaches and instead ask: *"How should intermediate generation be steered so that the denoised samples at $t = 0$ are drawn from the pure modes?"* To answer this question, we formulate pure-mode sampling as the terminal reward and derive an intermediate steering direction that maximizes this reward.

Observe that this reward is defined through the model's likelihoods and a diffusion model would not provide it directly[1]. Our key observation is that the reward does not need these likelihoods individually. The reward is a ratio of likelihoods, so under the variational approximation of the diffusion likelihood, the intractable terms cancel and the reward reduces to a contrast of denoising errors. We show how the new reward is interpretable, and can be mapped to a *Diffusion classifier* score that samples from *pure modes* by trying to increase the classification score for the joint condition C, while reducing the scores for the individual concept conditions c_i. Finally, since the reward is purely distribution-based, it's also modality-agnostic and does not rely on training an external reward model. We show results in both image and audio generation benchmarks demonstrating that such a reward is effective for compositional generation and viable in both domains. Our contributions are:

❶ We derive a tractable reward for intermediate steps of the denoising process that provably optimizes the end-goal of sampling *pure modes*. This reward is intrinsic to the model, hence generalizes across modalities (obviating the need for data-trained reward models); it is easily interpretable.

❷ We show why the reward works: maximizing the reward is equivalent to maximizing the total correlation between a sample and the input conditions. Intuitively, the reward favors samples that satisfy the synergy or coherence among the concepts.

[1]A naive approach could track the likelihood all the way to a future clean sample, and backpropagation through the entire sampling trajectory, which would be computationally prohibitive.

❸ Results show how the reward consistently avoids mode overlap or concept dominance by trying to select samples that increase the classification score for condition C while reducing the scores for the individual concepts $c_{1:k}$.

2 BACKGROUND

■ **Diffusion Sampling and Classifier-Free Guidance.** Diffusion models (Ho et al., 2020; Song et al., 2021b) learn to reverse a noising process that maps data $x_0 \sim p_0$ to noise $x_T \sim p_T$. Conditional models approximate the score $\nabla_{x_t} \log p_t(x_t \mid c)$, commonly parameterized by a noise prediction $\epsilon_\theta(x_t, t \mid c)$. Given x_t, the Tweedie estimate of the clean sample is

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

A deterministic DDIM step then updates $x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(x_t, t \mid c)$ (Song et al., 2021a). For text-conditioned generation, classifier-free guidance (CFG) (Ho & Salimans, 2022) combines conditional and unconditional noise predictions as

$$\epsilon_t^{\lambda,c} = \lambda\,\epsilon_\theta(x_t, t \mid c) + (1-\lambda)\,\epsilon_\theta(x_t, t \mid \varnothing), \tag{3}$$

and substitutes $\epsilon_t^{\lambda,c}$ into the sampling update. CFG++ (Chung et al., 2025) instead uses the guided prediction to estimate $\hat{x}_0$ while retaining the unconditional prediction for the noise direction:

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

We build on this clean-space view of diffusion sampling, as our method modifies the Tweedie estimate $\hat{x}_0$ using a model-intrinsic reward before propagating the next sampling step.

■ **Diffusion Steering with Rewards.** Diffusion sampling can also be steered by an external reward $r(x_0)$, e.g., a human-preference or aesthetic score (Kim et al., 2025). Such rewards are typically defined only on clean samples x_0. Given a pre-trained model with distribution p_{pre}, the goal is a target distribution that maximizes expected reward while staying close to p_{pre}:

$$p_{\text{tar}} = \arg\max_p \left[\mathbb{E}_{x_0\sim p}[r(x_0)]\right] \quad \text{s.t.} \quad D_{\text{KL}}(p \,\|\, p_{\text{pre}}) < \epsilon. \tag{5}$$

Introducing a Lagrange multiplier $\alpha > 0$ for the constraint, equation 5 can equivalently be written in its unconstrained, penalized form:

$$p_{\text{tar}} = \arg\max_p \left[\mathbb{E}_{x_0\sim p}[r(x_0)] - \alpha D_{\text{KL}}(p \,\|\, p_{\text{pre}})\right]. \tag{6}$$

Taking the functional derivative of this objective with respect to p and enforcing the normalization constraint $\int p(x_0)dx_0 = 1$ yields the optimality condition $\log p_{\text{tar}}(x_0) = \log p_{\text{pre}}(x_0) + r(x_0)/\alpha + \text{const}$. Thus, the regularized objective admits the closed-form solution as follows, where Z is the normalizing constant ensuring that p_{tar} integrates to one.

$$p_{\text{tar}}(x_0) = \frac{1}{Z} p_{\text{pre}}(x_0) \exp\left(\frac{r(x_0)}{\alpha}\right), \qquad Z = \int p_{\text{pre}}(x_0) \exp\left(\frac{r(x_0)}{\alpha}\right) dx_0 \tag{7}$$

3 DIFFUSION STEERING WITH INTRINSIC REWARD

We first motivate with a simple experiment that compositional failures arise from overlap modes. Then we formulate pure-mode sampling as reward alignment and show theoretical justification as to why the model-intrinsic reward is a readily available and useful quantity for alignment. We end with the derived reward guidance algorithm.

3.1 PURE MODE AS AN INTRINSIC REWARD

■ **Observation on compositional generation.** We start with an experiment by giving prompts of the form "`a [Animal] and a [Object]`" to SDXL (Podell et al., 2023b) and generated samples over multiple seeds. Figure 2 measures DINOv2 concept-quality scores for generated samples from these compositional prompts. A significant portion of 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 $p^\theta(x_0 \mid C)$ overlaps with one of the marginals $p^\theta(x_0 \mid c_i)$. This substantiates the hypothesis that compositional failures are closely linked to mode overlap.

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:

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

The reweighting suppresses regions where any single $p^\theta(x \mid c_i)$ is large and preserves regions with concept synergy where mass under the joint is supported by all concepts in a balanced way. Notably, the above distribution lies in data space or over x_0. As discussed in 1, previous works Dutta et al. (2026a); Kwon & Ye (2025) target sampling from the above distribution by defining a heuristic corrector of the same form at all $t > 0$.

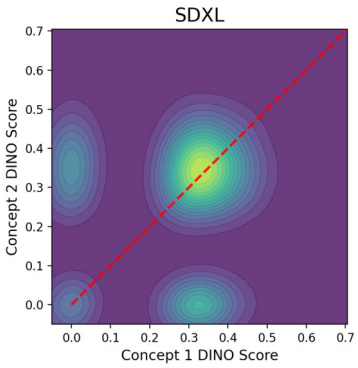


Figure 2: Density plot of DINOv2 scores between each sub-prompt c_1, c_2 and the generated image using SDXL. This visualizes the failure in compositional generation.

■ **Formulation.** We break away from the heuristic corrector approach and we pose the "pure mode" sampling as test-time *reward alignment* problem at $t = 0$. Following equation 5, we setup our optimization as:

$$p^* \;=\; \arg\max_{p} \mathbb{E}_{x_0 \sim p}\Big[\log \tfrac{p^\theta(x_0|C)}{\prod_{i=1}^{K} p^\theta(x_0|c_i)^{w_i}}\Big] \quad \text{s.t.} \quad \mathrm{D_{KL}}\big(p \,\|\, p^\theta(\cdot \mid C)\big) < \epsilon. \tag{9}$$

where $w_i \in \mathbb{R}^+$ and $\sum_i w_i = 1$. Note that, the reward $R_w(x_0) := \log \frac{p^\theta(x_0|C)}{\prod_{i=1}^{K} p^\theta(x_0|c_i)^{w_i}}$ is defined intrinsically from the pretrained model itself. The Proposition below theoretically justifies why the optimization in equation 9 leads to a distribution that emphasizes pure modes where all concepts synergize.

Proposition 1 (Pure-mode reward and concept synergy, informal)**.** *(i) For a fixed prompt C, applying Bayes' rule to the reward R_w in equation 9 gives, up to a constant independent of x_0,*

$$R_w(x_0) \;=\; \underbrace{\log \frac{p^\theta(C \mid x_0)}{\prod_i p^\theta(c_i \mid x_0)}}_{\text{point-wise conditional Total Correlation (pcTC) } \iota_C(x_0)} \;+\; \underbrace{\sum_i (1 - w_i) \log p^\theta(c_i \mid x_0)}_{\text{per-concept evidence } \pi_w(x_0)} \;+\; \text{const.} \tag{10}$$

Hence solving equation 9 amounts to maximizing the expected pcTC plus per-concept evidence while staying close to $p^\theta(\cdot \mid C)$, and the solution p^* raises this quantity relative to the pretrained model.

(ii) At the population level (across prompts and samples),

$$\mathbb{E}_C\, \mathbb{E}_{x_0}\big[R_w(x_0)\big] \;=\; \underbrace{\mathrm{TC}(c_1; \ldots; c_K \mid X_0)}_{\text{conditional Total Correlation}} \;-\; \underbrace{\sum_i (1 - w_i)\, H(c_i \mid X_0)}_{\text{per-concept uncertainty}} \;+ \text{const}, \tag{11}$$

where $\mathrm{TC}(c_1; \ldots; c_K \mid X_0) = \sum_i H(c_i \mid X_0) - H(c_1, \ldots, c_K \mid X_0)$ denotes the conditional Total Correlation (Watanabe, 1960), i.e., how strongly the concepts explaining a generated sample depend on each other, and $H(c_i \mid X_0)$ denotes the uncertainty about concept c_i that remains after seeing the sample (Appendix A.2.1).

■ **Remarks.** ❶ *Synergy.* ι_C is the pointwise conditional total correlation (pcTC), the per-sample analogue of conditional Total Correlation (Watanabe, 1960). For $K = 2$ it equals pointwise mutual information between c_1 and c_2 given x_0. It is high when the model's posterior at x_0 explains the concepts *jointly* rather than independently, and it vanishes when the posterior factorizes over concepts. ❷ *Presence.* π_w rewards the evidence x_0 provides for each concept on its own, so synergy cannot be gained by depicting every concept only weakly. ❸ *Pure modes.* Together, $\iota_C + \pi_w = \sum_i w_i \log \frac{p^\theta(C|x_0)}{p^\theta(c_i|x_0)}$, which is negative when a single concept c_i satisfies x_0 better than the full prompt C. These are the overlap modes of Fig. 2, which p^* down-weights. ❹ *Population view.* At the population level, the pure-mode objective favors high Total Correlation among the concepts while penalizing concepts that cannot be identified from the sample.

3.2 SAMPLING USING PURE MODE GUIDANCE

As discussed in section 2, the problem in equation 9 admits the closed form solution:

$$p^*(x_0) \;=\; \tfrac{1}{Z}\, p^\theta(x_0 \mid C)\, \exp\big(\beta\, R_w(x_0)\big) \tag{12}$$

i.e., $\log p^*(x_0) \;=\; \log p^\theta(x_0 \mid C) + \beta\, R_w(x_0) - \log Z$, where $R_w(x_0) = \log \frac{p^\theta(x_0|C)}{\prod_i p^\theta(x_0|c_i)}$.

■ **Reward alignment as posterior sampling.** We introduce a binary optimality variable $O \in \{0, 1\}$ with

$$p(O = 1 \mid x_0, C) \propto \exp\big(\beta R_w(x_0)\big), \tag{13}$$

such that samples with higher intrinsic reward are exponentially more likely to be optimal. Applying Bayes' rule with prior $p^\theta(\cdot \mid C)$ gives

$$p(x_0 \mid C, O = 1) \propto p^\theta(x_0 \mid C) \exp\big(\beta R_w(x_0)\big) \propto p^*(x_0). \tag{14}$$

Thus, sampling from the reward-tilted distribution p^* is equivalent to posterior sampling conditioned on optimality. We omit "$O = 1$" hereafter for brevity. Reverse diffusion requires the posterior score at every diffusion time:

$$\nabla_{x_t} \log p(x_t \mid C, O) = \underbrace{\nabla_{x_t} \log p(x_t \mid C)}_{\text{prior score}} + \underbrace{\nabla_{x_t} \log p(O \mid x_t, C)}_{\text{reward guidance}}. \tag{15}$$

The latter term is generally intractable since

$$p(O \mid x_t, C) = \mathbb{E}_{p(x_0|x_t,C)} \left[p(O \mid x_0, C) \right]. \tag{16}$$

Following posterior-guided diffusion (Chung et al., 2023), we approximate this expectation using the Tweedie's formula for posterior mean $\hat{x}_0(x_t)$:

$$p(O \mid x_t, C) \approx p(O \mid \hat{x}_0(x_t), C). \tag{17}$$

Substituting the optimality likelihood gives

$$\nabla_{x_t} \log p(O \mid x_t, C) \approx \beta \nabla_{x_t} \hat{R}_w(x_t), \qquad \hat{R}_w(x_t) \coloneqq R_w(\hat{x}_0(x_t)). \tag{18}$$

■ **Approximating the model-intrinsic reward.** Direct evaluation of $R_w(\hat{x}_0)$ requires the terminal likelihoods $p^\theta(\hat{x}_0 \mid C)$ and $p^\theta(\hat{x}_0 \mid c_i)$. Our key observation is that the reward can be written as a sum of log-likelihood ratios,

$$R_w(\hat{x}_0) = \sum_{i=1}^{K} w_i \log \frac{p^\theta(\hat{x}_0 \mid C)}{p^\theta(\hat{x}_0 \mid c_i)}. \tag{19}$$

Using the variational form of the diffusion log-likelihood, $-\log p^\theta(x_0 \mid c) = \int_0^T \gamma(t) \mathbb{E}_\epsilon \left[\|\epsilon - \epsilon_\theta(x_t, t \mid c)\|^2 \right] \mathrm{d}t + \text{const}$, we can approximate the likelihood ratio using the differences of Diffusion Classifier Energies (Li et al., 2023a):

$$\hat{R}_w(\hat{x}_0) = -\frac{1}{K} \sum_{i=1}^{K} w_i \left[E_C(\hat{x}_0) - E_{c_i}(\hat{x}_0) \right], \tag{20}$$

where $E_c(x_0) = \mathbb{E}_{\tau,\epsilon} \left[\gamma(\tau) \|\epsilon - \epsilon_\theta(x_\tau, \tau \mid c)\|^2 \right]$. Here, $x_\tau = \sqrt{\bar{\alpha}_\tau} x_0 + \sqrt{1 - \bar{\alpha}_\tau}\epsilon$. We use the same sampled (τ, ϵ) across the joint and constituent conditions when estimating their energy differences. We want to minimize the energy E_C while maximizing energies E_{c_i}. The full derivation and interpretation are provided in Appendix A.2.2.

■ **Reward-corrected Tweedie sampling.** Equation 18 suggests differentiating the reward through the Tweedie estimate, $\nabla_{x_t} \hat{R}_w(x_t) = J_{\hat{x}_0}(x_t)^\top \nabla_{\hat{x}_0} \hat{R}_w(\hat{x}_0)$, where $J_{\hat{x}_0}(x_t)$ denotes the Jacobian of the Tweedie prediction with respect to x_t. Rather than explicitly propagating the reward gradient

Algorithm 1 TILT: Diffusion sampling using model-intrinsic reward

Require: pretrained diffusion model ϵ_θ, multi-prompt $C = \{c_i\}_{i=1}^K$, number of diffusion steps N, stopping index M, schedule $\{t_n\}$, reward strength β.

$x_T \sim \mathcal{N}(0, I)$

for $n = N, N-1, \ldots, M+1$ **do**

$\hat{x}_0 \leftarrow \dfrac{x_{t_n} - \sqrt{1-\bar{\alpha}_{t_n}}\,\epsilon_\theta(x_{t_n}, t_n \mid C)}{\sqrt{\bar{\alpha}_{t_n}}}$ ▷ Tweedie estimate

$\hat{R}_t \leftarrow \dfrac{1}{K}\sum_{i=1}^K w_i \left[E_C(\hat{x}_0) - E_{c_i}(\hat{x}_0)\right]$ ▷ Energy-based reward, Eq. 20

$\tilde{x}_0 \leftarrow \hat{x}_0 + \beta \nabla_{\hat{x}_0} \hat{R}_t$ ▷ Reward correction, Eq. 21

$x_{t_{n-1}} \leftarrow \mathrm{DDIM}(\tilde{x}_0; \epsilon_\theta(x_{t_n}, t_n \mid C))$

end for

return x_0

through this Jacobian, we use a simple Jacobian-free clean-space corrector. At each denoising step, we compute $\nabla_{\hat{x}_0}\hat{R}_w(\hat{x}_0)$ and directly update the predicted clean sample as

$$\tilde{x}_0 \leftarrow \hat{x}_0 + \beta \nabla_{\hat{x}_0}\hat{R}_w(\hat{x}_0) \tag{21}$$

where β is the reward strength. The corrected prediction $\tilde{x}_0$ is then used in place of $\hat{x}_0$ in the subsequent DDIM update, following He et al. (2024). Note that this does not require external reward model or fine-tuning of the model.

4 Experiments

■ **Implementation Details.** For experiments in image space, we use SDXL (Podell et al., 2023b) as a base model and without any additional training or fine-tuning. Our method modifies only the inference procedure. We use the DDIM scheduler with CFG (scale 5.0) with 50 denoising steps. In text2audio space, we use AudioLDM (Liu et al., 2023) as a base model with DDIM with CFG (scale 3.5) as the base sampler with 50 steps. The proposed correction is applied on early denoising steps. Unless otherwise specified, we correct the first 5 denoising steps, use 2 latent correction step per corrected timestep, and use 5 correction iterations at the initial timestep. More implementation details in Appendix A.3.

■ **Evaluation benchmarks and metrics.** For text-to-image generation, we use **GenEval** (Ghosh et al., 2023) for simple prompts, an object-centric benchmark covering co-occurrence, count, color, and spatial relations, and **T2I-CompBench** (Huang et al., 2023a) for open-world prompts, evaluating its *Color*, *Shape*, *Texture*, and *Complex* subsets.

Table 1: Quantitative results on GenEval prompts.

Method	T2V ↑	HPSv3 ↑
CFG	0.3934	5.5029
CFG++	0.5073	5.9806
R2F	0.4799	5.8176
SuperDiff	0.1589	0.082
CO3	0.5014	4.3026
DOS	0.5025	5.5744
TILT	0.5192	6.1219

For images, we report **T2V** (Lin et al., 2024) for text–image alignment and **HPSv3** (Ma et al., 2025) for learned human preference. For audio, we report **CLAP** (Elizalde et al., 2023) for text–audio alignment; **FAD** (Kilgour et al., 2019) and **FD** (Liu et al., 2023), which measure the distance between generated and reference audio features extracted with VGGish and PANNs, respectively; and **KL** divergence between their predicted audio-event distributions. Higher is better for T2V, HPSv3, and CLAP; lower is better for FAD, FD, and KL. CLIP (Radford et al., 2021) is used only as an external reward in baseline comparisons. To test whether our reward generalizes beyond images, we also evaluate text-to-audio generation on **AudioCaps** (Kim et al., 2019), a dataset of audio clips paired with human-written captions.

■ **Comparison baselines.** For text-to-image generation, we compare against **CFG** (Ho & Salimans, 2022), **CFG++** (Chung et al., 2025), and prior compositional generation methods, including

Table 2: Quantitative results on T2I-CompBench prompts.

	Color		Shape		Texture		Complex	
Method	T2V ↑	HPSv3 ↑	T2V ↑	HPSv3 ↑	T2V ↑	HPSv3 ↑	T2V ↑	HPSv3 ↑
CFG	0.5183	5.1216	0.3565	6.1377	0.5529	6.5021	0.6238	7.5702
CFG++	0.5668	5.7358	0.3839	6.9679	0.5839	7.1200	0.6307	8.1785
R2F	0.5328	5.5146	0.3452	6.5545	0.5573	6.7431	0.6234	8.0638
CO3	0.5938	4.4736	0.3895	6.1982	0.4252	6.2801	0.6313	6.6193
DOS	0.5028	4.8979	0.3865	6.0388	0.6199	6.4046	0.6603	6.4046
SuperDiff	0.2420	-0.2098	0.1660	2.6294	0.2528	0.4163	0.3865	2.3976
TILT (Ours)	0.5688	5.8002	0.4019	7.0523	0.5932	7.2136	0.6350	8.2156

SuperDiff (Skreta et al., 2025b), **R2F** (Park et al., 2025), **CO3** (Dutta et al., 2026a), and **DOS** (Byun et al., 2026), using the same SDXL (Podell et al., 2023a) backbone. For text-to-audio generation, we use **AudioLDM** (Liu et al., 2023) as the base generator and compare against the reward-steering method **SCORE** (Jung et al., 2025). Unless otherwise specified, all comparisons use the same underlying pretrained model and evaluation setting within each modality (see Sec. 5 for more details).

4.1 Comparison on Text-to-Image Generation

■ **Simple Prompts.** Table 1 shows the results for TILT on GenEval prompts. All the methods use SDXL as a base model and doesn't use or depend on any external signal for generation. All the baselines here depend only on the base model and don't utilize any external reward signal like TILT. In both metrics TILT outperforms the baselines. Especially, our method gains significantly over the baselines in HPSv3 metric which aligns with human preferences. Notably, it also outperforms CO3 which is a corrector-based approach that uses a corrector distribution that structurally similar to equation 6. The result validates our claim that "pure mode" as a data-space reward is more effective that intermediate-heuristic correctors.

■ **Complex Prompts.** Table 2 shows TILT's performance on more complex scenarios using the T2ICompbench prompts. Our method consistently outperforms or matches the performance of the baselines across most of the evaluated metrics. Similar to simpler prompts in Table 1, TILT also demonstrates superior performance in capturing the nuances of complex compositions and validates that our proposed intrinsic reward equally effective in the more complex and challenging scenarios. Our method is the best performing in HPSv3 metric that mimics human preferences.

■ **Qualitative Comparison.** Figure 3 compares TILT qualitatively with other baselines. The first two rows use GenEval prompts. SDXL, CFG++ and CO3 struggled with simpler prompts of the form `"a [concept1] and a [concept2]"` (1st row). With a slightly complicated modifier "right of" (2nd row), most methods were unable to generate the desired image. Bottom four rows show results on more complicated scenarios from T2ICompBench. 3rd and 4th rows contain *shape* and *texture* information while 5th row has *color* modifiers. Most methods either drop concepts (row 4) or mix the attributes between concepts (row 5). Last row contains *complex* prompts that combine multiple concepts with relations between them. TILT retains better fidelity even in these challenging cases.

Table 3: Comparison with extrinsic-reward steering on GenEval.

Method	BLIP-VQA ↑	T2V ↑	HPSv3 ↑
SDXL	0.339	0.483	5.50
FKS-CLIP	0.450	0.563	5.89
FKS-TILT	0.446	0.530	5.91

4.2 Comparison with methods using extrinsic rewards

■ **Text-to-Image Generation.** We compare against FKS (Singhal et al., 2025), a sequential Monte Carlo method that steers diffusion with an external CLIP reward. It runs 4 particles, scores them with CLIP at intermediate steps, and resamples to keep the most promising ones. To isolate the effect of the reward, we keep the FKS sampler fixed and replace CLIP with our intrinsic reward. Table 3 shows that FKS-TILT performs on par with FKS-CLIP and surpasses it on HPSv3. This comparison favors CLIP: it is an external text-image alignment model trained on a large-scale corpus of paired text and images, so it brings supervision from outside the generator. TILT, by contrast, is computed

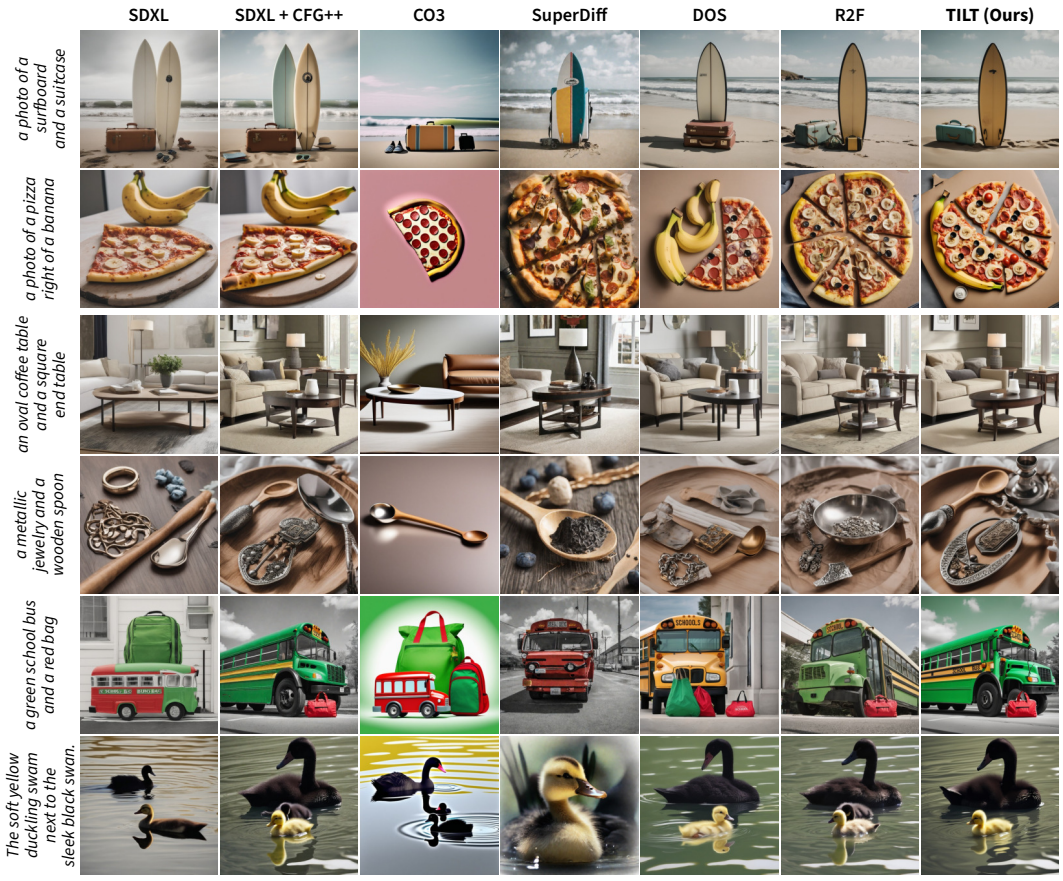


Figure 3: **Qualitative comparison on GenEval and T2I-CompBench.** The top two rows use prompts from GenEval and the bottom four rows use prompts from T2I-CompBench (shape, texture, color, and complex subsets); each row's prompt is shown on the left. All methods use the SDXL backbone and share the same random seed within each row.

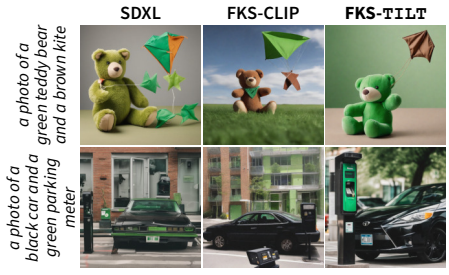


Figure 4: **Qualitative comparison with extrinsic reward on GenEval.** Each row shows SDXL and FKS, where the final sample is the best of $K=4$ particles under the CLIP reward (FKS-CLIP) or under our reward (FKS-TILT).

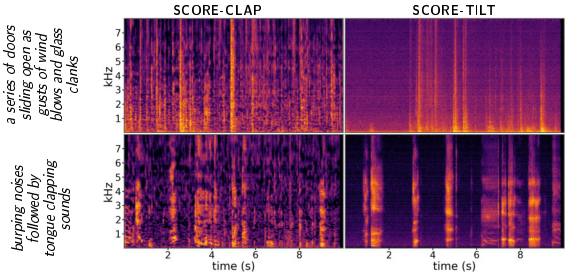


Figure 5: **Qualitative comparison with an extrinsic reward on AudioCaps.** Mel spectrograms of samples from Score steered by the CLAP reward (SCORE-CLAP) and by our reward (SCORE-TILT), both on AudioLDM; each row's prompt is shown on the left.

from the base diffusion model and uses no external data, annotations, or auxiliary networks. That a reward requiring no outside supervision is competitive with a large externally trained reward model shows that the base model already contains much of the signal needed to correct compositional failures, and that TILT makes this signal usable without collecting data or training an external model. Figure 4 illustrates the qualitative comparison. More visualizations in Appendix A.4.

Table 4: Quantitative results on audio generation using prompts from AudioCaps.

Method	CLAP ↑	FAD ↓	KL ↓	FD ↓
AudioLDM	0.4508	4.9867	0.7782	3.0924
AudioLDM-TILT	0.5350	3.6415	0.6213	2.7377
AudioLDM-TILT (best of 8)	0.6110	3.0848	0.6176	1.9624
SCORE-TILT (8 particles)	0.5205	3.8210	0.5890	1.8640
SCORE-CLAP (8 particles)	0.6258	3.1874	0.5852	1.7321

■ **Text-to-Audio Generation.** Because TILT is defined entirely through the base model's own conditional distributions, it makes no assumption about the data modality. We therefore apply it unchanged to text-to-audio generation on AudioLDM (Liu et al., 2023) and compare against SCORE (Jung et al., 2025), which steers 8 particles with an external CLAP reward, scoring them at intermediate diffusion steps and mutating them towards higher-reward samples. As shown in Table 4, AudioLDM-TILT improves the base model on every metric (rows 1–2). Within the Score framework, replacing CLAP with TILT (SCORE-TILT) remains competitive with SCORE-CLAP on KL and FD. The comparison again favors the extrinsic reward as CLAP is trained on external audio-text pairs. Finally, to match Score's budget of 8 particles, we evaluate AudioLDM-TILT with best-of-8, which achieves a lower FAD than SCORE-CLAP and comparable scores elsewhere. Figure 5 compares spectrograms on two prompts and shows how SCORE-TILT captures the sound events better than SCORE-CLAP. More visualizations in Appendix A.5.

5 Related Work

■ **Trajectory-based compositional generation.** A growing body of work improves multi-concept generation by modifying pretrained diffusion models at inference time. Composable Diffusion (Liu et al., 2022) combines conditional scores, with MCMC-based methods improving sampling from the composed distribution (Du et al., 2024). SuperDiff (Skreta et al., 2025b) enables principled products and mixtures, while R2F (Park et al., 2025), TweedieMix (Kwon & Ye, 2025), and DOS (Byun et al., 2026) use prompt- or trajectory-level interventions. More recently, CO3 (Dutta et al., 2026b) suppresses modes shared by joint and single-concept distributions using a concept-contrasting corrector. In contrast, we define the desired pure-mode target directly in data space and sample from it via reward steering. ■ **Attention and layout-based compositional control.** Another line of work manipulates internal attention or imposes explicit spatial structure. Attention-based approaches modify token-level activations to reduce missing objects and attribute misbinding (Chefer et al., 2023; Feng et al., 2023; Phung et al., 2024), while layout-conditioned methods add bounding boxes, masks, region prompts, or conditioning modules for spatial control (Xie et al., 2023; Kim et al., 2023; Zhao et al., 2023; Wang et al., 2024; Li et al., 2023b; Mou et al., 2023; Zhang et al., 2023). Our method requires neither attention manipulation nor explicit layouts. ■ **Reward steering and model-intrinsic rewards.** Inference-time reward guidance steers pretrained diffusion models without retraining. DITTO (Novack et al., 2024) differentiates through sampling, Feynman–Kac methods use particle-based steering (Singhal et al., 2025; Skreta et al., 2025a), and SCORE (Jung et al., 2025) combines multiple rewards, typically externally specified. Model-intrinsic approaches instead derive rewards from the generator itself: SAIL (He et al., 2026) uses self-generated preferences, while SRUM (Jin et al., 2026) and Self-SpectraReward (Huang et al., 2026) use the model to evaluate generated samples. In contrast, we derive rewards directly from conditional likelihoods of a fixed pretrained diffusion model for training-free test-time steering.

6 Limitations and Future Work

In our current formulation evaluating the reward and its gradient requires additional passes through the diffusion model, which makes our sampler more computationally expensive than standard guidance. We believe much of this overhead is not fundamental: cheaper and more accurate approximations of the reward should reduce it substantially, and we view this as a natural next step. Future work could also explore applicaitons beyond images and audio, such as molecules and proteins which are also compositional in nature. More broadly, we hope this work points to a promising direction: steering generative models with rewards drawn from the models themselves. Such rewards need no external data, annotations, or modality-specific reward models, so the same principle can serve images, audio, and potentially other modalities alike. We are optimistic that this line of research will lead to reward steering methods that are both data-efficient and modality-agnostic.

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. Implementation code can be found via our project website: https://tinyurl.com/yr5j3bwe

AI Use Statement

Generative AI tools were used to polish manuscript writing.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Dongnam Byun, Jungwon Park, Jungmin Ko, Changin Choi, and Wonjong Rhee. Dos: Directional object separation in text embeddings for multi-object image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2490–2497, 2026.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2301.13826.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations*, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024. URL https://arxiv.org/abs/2306.05284.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations*, 2023.

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.

Debottam Dutta, Jianchong Chen, Rajalaxmi Rajagopalan, Yu-Lin Wei, and Romit Roy Choudhury. Steer away from mode collisions: Improving composition in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Debottam Dutta, Jianchong Chen, Rajalaxmi Rajagopalan, Yu-Lin Wei, and Romit Roy Choudhury. Steer away from mode collisions: Improving composition in diffusion models. In *International Conference on Learning Representations*, 2026b.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

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 *International Conference on Learning Representations*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023.

Xiaoxuan He, Siming Fu, Wanli Li, Zhiyuan Li, Dacheng Yin, Kang Rong, Fengyun Rao, and Bo Zhang. Sail: Self-amplified iterative learning for diffusion model alignment with minimal human feedback. In *International Conference on Learning Representations*, 2026.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *International Conference on Learning Representations*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL `https://arxiv.org/abs/2207.12598`.

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.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. URL `https://arxiv.org/abs/2210.02303`.

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, 2023a.

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, 2023b.

Runhui Huang, Qihui Zhang, Zhe Liu, Yu Gao, Jie Wu, and Hengshuang Zhao. Read it back: Pretrained mllms are zero-shot reward models for text-to-image generation. *arXiv preprint arXiv:2607.11886*, 2026.

Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. Srum: Fine-grained self-rewarding for unified multimodal models. In *European Conference on Computer Vision*, 2026.

Jaemin Jung, Jaehun Kim, Inkyu Shin, and Joon Son Chung. SCORE: Scaling audio generation using standardized composite rewards. *arXiv preprint arXiv:2509.19831*, 2025.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, pp. 2350–2354, 2019.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of NAACL-HLT*, pp. 119–132, 2019.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

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. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation, 2025. URL https://arxiv.org/abs/2410.05591.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023a. URL https://arxiv.org/abs/2303.16203.

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.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023.

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. doi: 10.1007/978-3-031-19790-1_26.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

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.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. DITTO: Diffusion inference-time t-optimization for music generation. In *International Conference on Machine Learning*, 2024.

Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. In *International Conference on Learning Representations*, 2025.

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*, 2023a.

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*, 2023b.

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. doi: 10.18653/v1/2020.acl-demos.14. URL https://aclanthology.org/2020.acl-demos.14/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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, 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)*, 2022.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *International Conference on Machine Learning*, 2025.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alan Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. In *International Conference on Machine Learning*, 2025a.

Marta Skreta, Lazar Atanackovic, Joey Bose, Alexander Tong, and Kirill Neklyudov. The superposition of diffusion models using the itô density estimator. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=2o58Mbqkd2.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

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.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation, 2024. URL https://arxiv.org/abs/2402.03290.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4 (1):66–82, January 1960. ISSN 0018-8646. doi: 10.1147/rd.41.0066. URL https://doi.org/10.1147/rd.41.0066.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL http://dx.doi.org/10.1038/s41586-023-06415-8.

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.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

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.

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

CONTENTS

A.1 Impact Statement 15
A.2 Derivations 15
A.2.1 Derivation of Proposition 1 15
A.2.2 Derivation of the simplified reward 17
A.3 More implementation details 17
A.4 Additional Vision Qualitative results 18
A.4.1 More visual results for external reward 18
A.4.2 failure cases for TILT 18
A.4.3 failure cases for Comparison with external reward 18
A.5 Audio Qualitative results 18
A.6 Ablations & Additional Visual Numerical Results 18

A.1 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.

A.2 DERIVATIONS

A.2.1 DERIVATION OF PROPOSITION 1

Proposition 1 (Pure-mode reward and concept synergy)**.** *Let $w_i > 0$ with $\sum_i w_i = 1$, and let $R_w(x_0) \coloneqq \log \frac{p^\theta(x_0|C)}{\prod_i p^\theta(x_0|c_i)^{w_i}}$ be the reward in equation 9, with maximizer p^*_C (written p^* when C is fixed). Assume the conditionals $p^\theta(x_0 \mid C)$ and $p^\theta(x_0 \mid c_i)$ are induced by a single joint $p^\theta(x_0, c_1, \ldots, c_K)$ over samples and the concept tuple $C = (c_1, \ldots, c_K)$ (a* consistent implicit joint*). Then:*

(i) Pointwise objective. For a fixed prompt C, applying Bayes' rule to R_w gives, up to a constant independent of x_0,

$$R_w(x_0) \;=\; \underbrace{\log \frac{p^\theta(C \mid x_0)}{\prod_i p^\theta(c_i \mid x_0)}}_{\text{concept synergy (pcTC) } \iota_C(x_0)} \;+\; \underbrace{\sum_i (1 - w_i) \log p^\theta(c_i \mid x_0)}_{\text{per-concept evidence } \pi_w(x_0)} \;+ \text{const}(C), \tag{22}$$

where ι_C is the pointwise conditional total correlation *(pcTC). Hence solving equation 9 amounts to maximizing the expected pcTC plus per-concept evidence while staying close to $p^\theta(\cdot \mid C)$, and the solution p^* attains at least the pretrained model's value of this quantity:*

$$\mathbb{E}_{x_0 \sim p^*}\big[\iota_C(x_0) + \pi_w(x_0)\big] \;\geq\; \mathbb{E}_{x_0 \sim p^\theta(\cdot|C)}\big[\iota_C(x_0) + \pi_w(x_0)\big]. \tag{23}$$

(ii) Population level. Let prompts be drawn from the prompt marginal $p(C)$ of the consistent implicit joint, with concepts taking values in a finite vocabulary and the weights w shared across prompts. At the population level (across prompts and samples),

$$\mathbb{E}_{C \sim p(C)}\, \mathbb{E}_{x_0 \sim p^\theta(\cdot|C)}\big[R_w(x_0)\big] \;=\; \underbrace{\mathrm{TC}(c_1; \ldots; c_K \mid X_0)}_{\text{conditional Total Correlation}} \;-\; \underbrace{\sum_i (1 - w_i)\, H(c_i \mid X_0)}_{\text{per-concept uncertainty}} \;+ \text{const}, \tag{24}$$

where $\mathrm{TC}(c_1;\ldots;c_K \mid X_0) \coloneqq \sum_i H(c_i \mid X_0) - H(c_1,\ldots,c_K \mid X_0)$ *is the conditional Total Correlation (Watanabe, 1960), with entropies taken under the joint* $p(C)\,p^\theta(x_0 \mid C)$ *read in the recognition direction: given a sample, which prompts explain it?*

Proof of (i). Write $q \coloneqq p^\theta(\cdot \mid C)$.

Since both conditionals are induced by the consistent implicit joint, Bayes' rule gives

$$\begin{aligned}\log p^\theta(x_0 \mid C) &= \log p^\theta(C \mid x_0) + \log p^\theta(x_0) - \log p^\theta(C),\\ \log p^\theta(x_0 \mid c_i) &= \log p^\theta(c_i \mid x_0) + \log p^\theta(x_0) - \log p^\theta(c_i).\end{aligned} \tag{25}$$

Substituting into R_w, the unconditional term appears with coefficient $1 - \sum_i w_i = 0$ and drops out:

$$R_w(x_0) = \log p^\theta(C \mid x_0) - \sum_i w_i \log p^\theta(c_i \mid x_0) + \kappa_w(C), \qquad \kappa_w(C) \coloneqq \sum_i w_i \log p^\theta(c_i) - \log p^\theta(C). \tag{26}$$

Adding and subtracting $\sum_i \log p^\theta(c_i \mid x_0)$ and regrouping,

$$\begin{aligned}R_w(x_0) - \kappa_w(C) &= \Big[\log p^\theta(C \mid x_0) - \sum_i \log p^\theta(c_i \mid x_0)\Big] + \Big[\sum_i \log p^\theta(c_i \mid x_0) - \sum_i w_i \log p^\theta(c_i \mid x_0)\Big]\\ &= \underbrace{\log \frac{p^\theta(C \mid x_0)}{\prod_i p^\theta(c_i \mid x_0)}}_{\text{concept synergy (pcTC) } \iota_C(x_0)} + \underbrace{\sum_i (1 - w_i) \log p^\theta(c_i \mid x_0)}_{\text{per-concept evidence } \pi_w(x_0)},\end{aligned} \tag{27}$$

which is equation 22 with $\mathrm{const}(C) = \kappa_w(C)$.

$\iota_C(x_0)$ 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, it measures how much more likely the model's posterior at x_0 finds the concepts *jointly* than it would if they were explained independently. Equivalently, since $\sum_i w_i = 1$,

$$\iota_C(x_0) + \pi_w(x_0) \;=\; \sum_i w_i \log \frac{p^\theta(C \mid x_0)}{p^\theta(c_i \mid x_0)}, \tag{28}$$

which is strongly negative exactly when some single concept c_i explains x_0 much better than the composed prompt C. These are the overlap modes of Fig. 2.

Because equation 27 holds pointwise, for every p,

$$\mathbb{E}_{x_0 \sim p}[R_w(x_0)] \;=\; \mathbb{E}_{x_0 \sim p}\big[\iota_C(x_0) + \pi_w(x_0)\big] + \kappa_w(C). \tag{29}$$

The two objectives differ by a constant, so maximizing either over the KL ball in equation 9 gives the same maximizer p^*. The base model q lies in the ball, since $\mathrm{KL}(q \,\|\, q) = 0$. Hence $\mathbb{E}_{p^*}[R_w] \geq \mathbb{E}_q[R_w]$, and subtracting $\kappa_w(C)$ gives equation 23. Finally, substituting equation 27 into the closed form equation 12, $\kappa_w(C)$ is absorbed into the normalizer:

$$p^*(x_0) \;\propto\; q(x_0)\, \exp\big(\beta\,[\iota_C(x_0) + \pi_w(x_0)]\big). \tag{30}$$

So p^* reweights the base conditional exponentially toward samples with high pcTC and high per-concept evidence, and away from the overlap modes. □

Proof of (ii).

We average equation 27 over prompts $C \sim p(C)$ and samples $x_0 \sim p^\theta(\cdot \mid C)$. Since $p(C)\,p^\theta(x_0 \mid C) = p^\theta(x_0)\,p^\theta(C \mid x_0)$, this is an average over samples and the prompts that explain them (the recognition direction), so each log-posterior averages to a negative conditional entropy:

$$\mathbb{E}\big[\log p^\theta(C \mid x_0)\big] = -H(c_1,\ldots,c_K \mid X_0), \qquad \mathbb{E}\big[\log p^\theta(c_i \mid x_0)\big] = -H(c_i \mid X_0). \tag{31}$$

Substituting into ι_C and π_w and adding back the constant,

$$\mathbb{E}_C\, \mathbb{E}_{x_0 \sim p^\theta(\cdot \mid C)}\big[R_w(x_0)\big] \;=\; \underbrace{\sum_i H(c_i \mid X_0) - H(c_1,\ldots,c_K \mid X_0)}_{\mathrm{TC}(c_1;\ldots;c_K \mid X_0)} \;-\; \sum_i (1 - w_i)\, H(c_i \mid X_0) + \mathrm{const}, \tag{32}$$

which is equation 24, with $\mathrm{const} = \mathbb{E}_C[\kappa_w(C)]$. □

A.2.2 Derivation of the simplified reward

We derive equation 20, writing $\hat{x}_0$ for the Tweedie estimate $\hat{x}_0(x_t)$ and R for the weighted reward R_w of Proposition 1.

Step 1: Reward as likelihood ratios. Since $\sum_i w_i = 1$, we can write $\log p^\theta(\hat{x}_0 \mid C) = \sum_i w_i \log p^\theta(\hat{x}_0 \mid C)$, so

$$R_w(\hat{x}_0) \;=\; \log \frac{p^\theta(\hat{x}_0 \mid C)}{\prod_i p^\theta(\hat{x}_0 \mid c_i)^{w_i}} \;=\; \sum_i w_i \Big[\log p^\theta(\hat{x}_0 \mid C) - \log p^\theta(\hat{x}_0 \mid c_i) \Big]. \tag{33}$$

Each term compares how well the full prompt C and the single concept c_i explain $\hat{x}_0$.

Step 2: Likelihoods from denoising errors. As in diffusion classifiers (Li et al., 2023a), we approximate the conditional log-likelihood by the diffusion ELBO (Ho et al., 2020), written as an expectation over timesteps $\tau \sim \mathcal{U}[0, T]$ and noise $\epsilon \sim \mathcal{N}(0, I)$:

$$-\log p^\theta(\hat{x}_0 \mid c) \;\approx\; \mathbb{E}_{\tau,\epsilon}\Big[\gamma(\tau)\, \|\epsilon - \epsilon_\theta(x_\tau, \tau, c)\|^2\Big] + \text{const} \;=\; -E_c(\hat{x}_0) + \text{const}, \tag{34}$$

where $x_\tau = \sqrt{\bar{\alpha}_\tau}\, \hat{x}_0 + \sqrt{1 - \bar{\alpha}_\tau}\, \epsilon$ re-noises $\hat{x}_0$, $\gamma(\tau)$ is the ELBO's time weighting, $E_c(\hat{x}_0) := -\mathbb{E}_{\tau,\epsilon}\big[\gamma(\tau)\, \|\epsilon - \epsilon_\theta(x_\tau, \tau, c)\|^2\big]$ is the diffusion classifier score, and the constant does not depend on the condition c.

Step 3: Substituting. Substituting into equation 33, the constants cancel in each log-ratio:

$$R_w(\hat{x}_0) \;\approx\; \sum_i w_i \big(E_C(\hat{x}_0) - E_{c_i}(\hat{x}_0)\big), \tag{35}$$

which is equation 20; any positive prefactor only rescales the reward and is absorbed into the reward strength β in equation 18. □

Intuition. $E_c(\hat{x}_0)$ is high when the model denoises $\hat{x}_0$ well when told c. The reward is therefore large when the full prompt C explains $\hat{x}_0$ better than each single concept c_i does, and small in overlap modes, where some c_i alone explains $\hat{x}_0$ as well as C. By Bayes' rule, $E_C - E_{c_i} \approx \log \frac{p^\theta(C|\hat{x}_0)}{p^\theta(c_i|\hat{x}_0)} + \text{const}$, so the simplified reward is the objective of Proposition 1, computed from denoising errors.

A.3 More implementation details

For the results in Table 2 we run reward correction for the first 5 time steps where we use TILT-S only at the first timestep followed by TILT-C on the next 4 steps. At each of these timesteps, we run 5 steps of optimization with the exception of 10 steps at the initial timestep. We use $\beta = 0.2$.

For implementation of the energy loss, we use a adaptive weighting strategy. We use $w_i = softmax(\mathbb{E}_{\tau,\epsilon}||\epsilon - \epsilon^\theta(c_i, \tau)||),\ i = 1, \ldots, K$ where we sample 4 pairs of (ϵ, τ) and use for all concepts c_i. This is the reduced variance version for the Diffusion Classifier score described in (Li et al., 2023a).

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 TILT. For example, if C is "a black cat and a brown dog", we consider c_1 ="cat" and c_2 ="dog".

■ **Vision Baselines** For implementation of the baselines we use their official code given in CO3 (Dutta et al., 2026a), R2F (Park et al., 2025), FKS (**?**), DOS (Byun et al., 2026), superDiff (Skreta et al., 2025b). For FKS-TILT experiements, we used TILT reward as a drop in replacement for the standard CLIP reward.

■ **Audio Baselines** For audio experiments, we the code provided in AudioLDM (Liu et al., 2023) for the base model. SCORE is reimplemented from the details given in the paper (Jung et al., 2025). For all audio experiments, we use sampling rate of 16 kHz, DDIM with 50 steps, CFG scale of 3.5.

A.4 Additional Vision Qualitative results

Figure 6 shows additional prompts on which TILT succeeds, using prompts not shown in Figure 3. TILT renders every concept with its intended attribute, while the baselines often drop one concept (e.g., the person in "a photo of a person and a sink") or bind the attribute to the wrong object (e.g., the purple train in "a photo of a red train and a purple bear").

A.4.1 More visual results for external reward

Figure 7 gives more prompt where TILT reward has a better performance compared the external CLIP reward, validating our previous statement of TILT.

A.4.2 failure cases for TILT

We also present our failure cases in Fig.8 together with other baselines. Three modes of failures recur.
Concept Copying. One concept is replaced by another instance. For "a photo of green bus and a purple microwave". TILT generates a green bus and a purple bus.
Concept Dropping. One concept is missing completely. For prompt " a photo of a bus below a toothbrush", the toothbrush is missing.
Fusion and attribute swapping. For "a photo of a horse and a giraffe", the largest animal has a horse's body with a giraffe's coat, and the other animals are giraffes. For "a brown banana and a green bear", both objects appear but their colors are swapped.

A.4.3 failure cases for Comparison with external reward

Failure cases. In Figure 9, we present some mismatching attribute binding and wrong spatial relation.

A.5 Audio Qualitative results

We add more qualitative results for the audio generation in Figure 10.

A.6 Ablations & Additional Visual Numerical Results

Tables 5–9 change one variable of our implementation at a time on GenEval.



Figure 6: **Additional success cases of `TILT`.** The top two rows use prompts from GenEval (two objects and color attribute) and the bottom seven rows use prompts from T2I-CompBench (texture and complex subsets); each row's prompt is shown on the left. All methods use the SDXL backbone and share the same random seed within each row.



Figure 7: **Additional success cases of FKS with our reward on GenEval.** each row's prompt is shown on the left. With the CLIP reward, FKS keeps samples that miss or merge an object (e.g., the person with an apple for a head, or the spoon in place of the knife) or bind a color to the wrong object (e.g., the green umbrella), while our reward keeps samples that render every object with its intended color.



Figure 8: **Failure cases of `TILT`.** The top three rows use prompts from GenEval (color attribute, position, and two objects) and the bottom four rows use prompts from T2I-CompBench (color, shape, texture, and complex subsets); each row's prompt is shown on the left. `TILT` fails on these prompts for all three evaluation seeds. Typical failure modes include replacing one concept with a copy of the other, dropping a concept entirely, fusing two concepts into a hybrid and swapping attributes.



Figure 9: **Failure cases of FKS(TILT)on GenEval.** Each row's prompt is shown on the left. FKS with our reward fails on these prompts for both evaluation seeds. The top three rows are color-attribute prompts, where the selected sample binds a color to the wrong object (e.g., the swapped colors of the cake and the chair); the bottom three rows are position prompts, where the selected sample drops an object (the parking meter) or places the objects in the wrong relation (the cell phone on the chair and the tie on the bat).

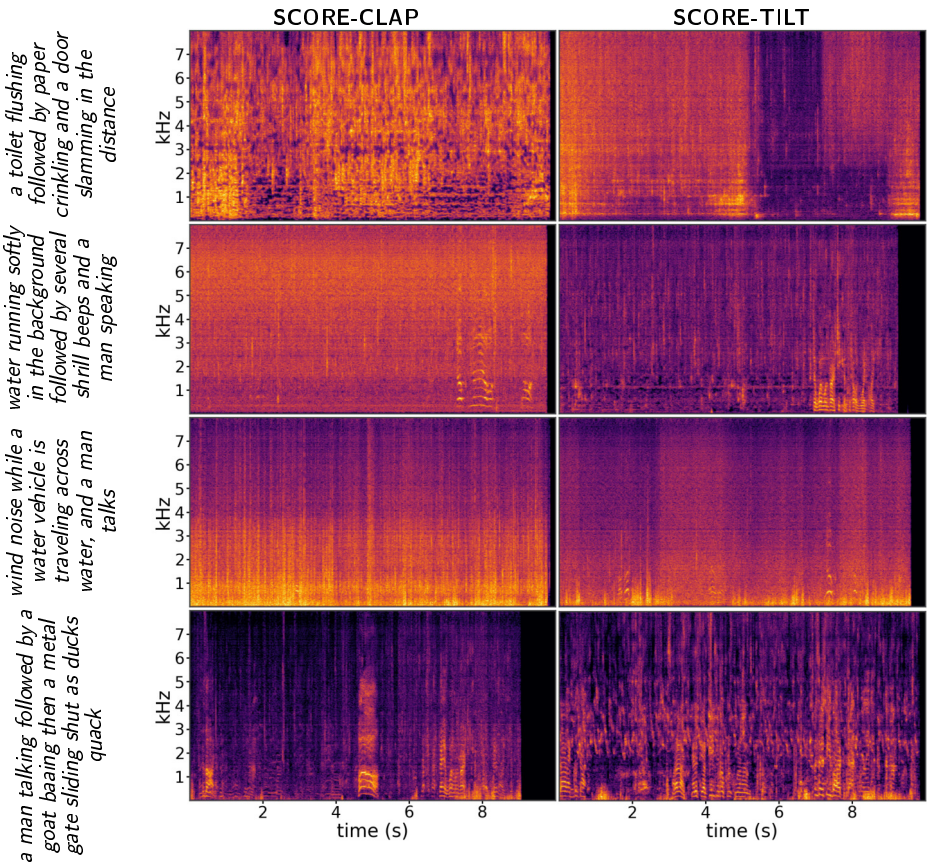


Figure 10: **Additional qualitative comparison on AudioCaps.** Mel spectrograms of samples from Score steered by the CLAP reward (SCORE-CLAP) and by our reward (SCORE-TILT), both on AudioLDM; each row's prompt is shown on the left.

Table 5: Effect of `energy_num_samples`, the number of (τ, ϵ) draws used to estimate the energy reward, on GenEval. As number of energy samples increases, the variance of the estimate decreases, leading to more stable performance.

`energy_num_samples`	BLIP-VQA ↑	T2V ↑	HPSv3 ↑
Base: default			
2	0.3942	0.5192	6.122
4	0.3900 (−0.0041)	0.4952 (−0.0240†)	6.049 (−0.072)
Base: default with `energy_adaptive_weights`, `energy_adaptive_zeta=3`			
2	0.3898	0.5178	6.120
4	0.3890 (−0.0009)	0.5067 (−0.0111)	6.069 (−0.051)
Base: default with `energy_adaptive_weights`, `energy_adaptive_zeta=5`			
2	0.3932	0.5194	6.104
4	0.3911 (−0.0021)	0.5023 (−0.0171)	6.082 (−0.022)
Base: default with `energy_adaptive_weights`, `energy_adaptive_zeta=7`			
2	0.3890	0.5212	6.134
4	0.3899 (+0.0009)	0.5098 (−0.0114)	6.093 (−0.042)
Base: default with `eta=0.1`, `nts_to_init_correct=0`			
2	0.3945	0.5110	6.137
8	0.3897 (−0.0047)	0.5062 (−0.0048)	6.103 (−0.034)
Base: default with `guidance_scale=0.8`, `eta=0.2`, `nts_to_init_correct=1`, `num_latent_corrector_steps=2`, `x0_hat_score_source=multi`‡			
2	0.3766	0.5093	6.005
4	0.3742 (−0.0024)	0.4959 (−0.0134)	6.022 (+0.018)

Table 6: Effect of `energy_adaptive_weights` defined in section A.3, that replaces the uniform per-concept weights of the energy reward by a softmin over the normalized score-difference norms, on GenEval.

`energy_adaptive_zeta`	BLIP-VQA ↑	T2V ↑	HPSv3 ↑
Base: default			
off	0.3942	0.5192	6.122
3	0.3898 (−0.0043)	0.5178 (−0.0014)	6.120 (−0.002)
5	0.3932 (−0.0010)	0.5194 (+0.0002)	6.104 (−0.018)
7	0.3890 (−0.0051)	0.5212 (+0.0020)	6.134 (+0.013)
Base: default with `energy_num_samples=4`			
off	0.3900	0.4952	6.049
3	0.3890 (−0.0011)	0.5067 (+0.0114)	6.069 (+0.019)
5	0.3911 (+0.0011)	0.5023 (+0.0070)	6.082 (+0.033)
7	0.3899 (−0.0001)	0.5098 (+0.0146†)	6.093 (+0.043)
Base: default with `guidance_scale=0.8`, `eta=0.5`, `energy_num_samples=4`			
off	0.3678	0.5036	6.010
3	0.3724 (+0.0046)	0.5019 (−0.0017)	6.004 (−0.006)
5	0.3710 (+0.0031)	0.4954 (−0.0082)	5.995 (−0.015)
7	0.3687 (+0.0008)	0.4865 (−0.0172†)	6.006 (−0.004)

Table 7: Effect of `x0_hat_score_source`, the score used to compute $\hat{x}_0$ on GenEval.

`x0_hat_score_source`	BLIP-VQA ↑	T2V ↑	HPSv3 ↑
Base: default			
ϵ_t^λ	0.3942	0.5192	6.122
ϵ_t §	0.3937 (−0.0004)	0.5180 (−0.0012)	6.109 (−0.013)

Table 8: Effect of `num_ts_to_correct`, the number of corrected timesteps, on GenEval

`num_ts_to_correct`	BLIP-VQA ↑	T2V ↑	HPSv3 ↑
Base: default			
5	0.3942	0.5192	6.122
2[§]	0.3981 (+0.0040)	0.5087 (−0.0106)	6.140 (+0.019)
Base: default with `eta=0.1`, `nts_to_init_correct=0`			
5	0.3945	0.5110	6.137
10	0.3956 (+0.0011)	0.5120 (+0.0010)	6.126 (−0.011)

Table 9: Effect of `num_latent_corrector_steps`, the number of latent-corrector steps per corrected timestep (the first `nts_to_init_correct` timesteps use `init_latent_corrector_steps` instead), on GenEval.

`num_latent_corrector_steps`	BLIP-VQA ↑	T2V ↑	HPSv3 ↑
Base: default with `guidance_scale=0.8`, `eta=0.2`, `energy_num_samples=4`, `nts_to_init_correct=1`, `x0_hat_score_source=multi`[‡]			
1	0.3725	0.4932	5.988
2	0.3742 (+0.0017)	0.4959 (+0.0027)	6.022 (+0.035)