Free^{2}Guide: Training-Free Text-to-Video Alignment using Image LVLM

Jaemin Kim, Bryan Sangwoo Kim, Jong Chul Ye
Kim Jaechul Graduate School of AI, KAIST
{kjm981995, bryanswkim, jong.ye}@kaist.ac.kr

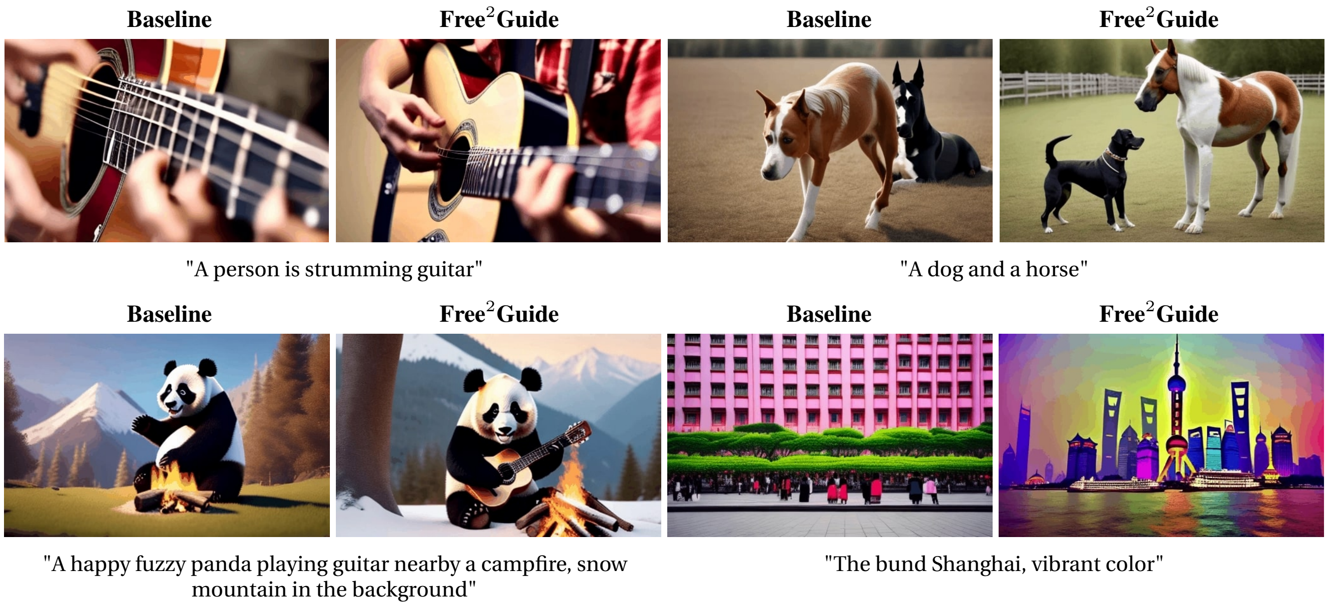

Figure 1. Representative video results using **Free^{2}Guide**, a novel framework that enables training-**Free**, gradient-**Free** video **Guid**ance leveraging a Large Vision-Language Model. *Each image shows the first frame of a video.*

Abstract

*Diffusion models have achieved impressive results in generative tasks for text-to-video (T2V) synthesis. However, achieving accurate text alignment in T2V generation remains challenging due to the complex temporal dependencies across frames. Existing reinforcement learning (RL)-based approaches to enhance text alignment often require differentiable reward functions trained for videos, hindering their scalability and applicability. In this paper, we propose **Free^{2}Guide**, a novel gradient-free and training-free framework for aligning generated videos with text prompts. Specifically, leveraging principles from path integral control, Free^{2}Guide approximates guidance for diffusion models using non-differentiable reward functions, thereby enabling the integration of powerful black-box Large Vision-Language Models (LVLMs) as reward models. To enable image-trained LVLMs to assess text-to-video alignment, we leverage stitching between video frames and use system prompts to capture sequential attributions. Our framework supports the flexible ensembling of multiple reward models to synergistically enhance alignment without significant computational overhead. Experimental results confirm that Free^{2}Guide using image-trained LVLMs significantly improves text-to-video alignment, thereby enhancing the overall video quality. Our results and code are available at our project page [1].*

1. Introduction

Diffusion models [21, 33, 34, 36] have emerged as powerful and versatile tools for generative modeling, achieving state-of-the-art results in tasks that require fine-grained control over content generation, such as text-to-image (T2I) [33] and text-to-video (T2V) generation [7, 15]. However, achieving perfect alignment with text conditions remains a significant challenge [12]. This issue becomes even more challenging in the video domain, where maintaining text-relevant content across frames requires handling complex temporal

[1] https://free2guide.github.io/

dependencies, often resulting in misalignment between generated frames and the given text prompt.

In the image domain, reinforcement learning (RL)-based methods have been introduced to address challenges in text-guided T2I generation by using reward models to estimate human preferences within diffusion models [2, 10, 47, 48]. Previous works mainly focus on either directly fine-tuning the diffusion model with gradients derived from a reward function [6, 30, 31] or employing an RL-based policy gradient approach [2, 10]. While these fine-tuning methods can effectively improve sample alignment, they have notable limitations: the former requires a differentiable reward function, while the latter is typically limited to only few prompts.

Directly adapting these text alignment approaches for the video domain presents two main challenges. First, they often require a dedicated video-specific reward function or additional training on curated video datasets. Collecting large-scale, aligned text-video datasets is far more complex than gathering image data, and developing reward functions tailored to video tasks is similarly difficult. Second, even with trained reward models for the video domain, additional challenges such as substantial memory demands for backpropagation emerge, which grow proportionally as model scale increases (*i.e.*, scaling laws) [19].

An alternative approach involves using differential reward models during inference time to guide diffusion models without fine-tuning model parameters [42]. However, guidance-based methods still require a differentiable reward function, which excludes non-differentiable options like state-of-the-art visual-language model APIs or human preference-based metrics. To address this, recent studies have explored stochastic optimization to guide diffusion models during the sampling process using non-differentiable objective functions in music generation [17], and concurrent research extends this idea within the image domain [50, 51]. However, such methods cannot be directly applied to video diffusion models due to the complex temporal dependencies involved.

To address these issues, here we introduce **Free^{2}Guide**—a novel text-to-video alignment method by leveraging the temporal understanding capabilities of Large Vision-Language Models (LVLMs). Specifically, Free^{2}Guide aligns text prompts in video generation without requiring gradients from the reward function. More specifically, drawing on principles from path integral control, Free^{2}Guide approximates guidance to align generated videos with text prompts, regardless of the reward function's differentiability. Another important contribution of this paper is a technique to adapt image-based LVLMs for temporal understanding. In particular, we concatenate video frames in a structured grid layout, and design prompts that explicitly indicate sequence order and reasoning to help LVLMs evaluate videos more comprehensively. By doing so, Free^{2}Guide enables the use of powerful black-box vision-language models as reward models, improving text-video alignment, as illustrated in Fig. 1. Finally, our framework allows for the flexible combination of reward models by eliminating the need for computationally intensive fine-tuning and backpropagation. As such, we explore several combinatorial approaches to collaborate LVLMs with existing large-scale image-based models. Extensive experiments show that our methods improve text alignment and the quality of generated videos.

Our contributions are summarized as follows:

- We introduce **Free^{2}Guide**, a novel framework for aligning generated videos with text prompts without requiring gradients from the reward function. To the best of our knowledge, Free^{2}Guide is the first gradient-free guidance approach for text-to-video generation that requires no additional training.
- We adapt non-differentiable image-based LVLM APIs to enhance text-video alignment by leveraging stitching and prompt design to capture video-specific attributes.
- We develop an effective ensemble approach that integrates large-scale image-based models to improve video generation guidance.

2. Related Work

Text-to-Video diffusion model Text-to-Video diffusion models (e.g., LaVie [43], VideoCrafter [3, 4]) employ diffusion processes to generate coherent video sequences from textual prompts [13, 16, 27]. However, a notable limitation is that video diffusion models often struggle to generate videos that align accurately with the given text prompts, specifically in terms of spatial relationships (e.g., "A on B") and the representation of temporal style (e.g., "zooming in").

Diffusion model with LVLM feedback While several approaches have been proposed to improve the diffusion generation process with Large Language Models (LLMs) [11, 25, 46, 52], there has been limited exploration of methods leveraging Large Vision Language Models (LVLMs) that can also handle image domains. Recent works explore the integration of LVLMs as a feedback mechanism to image diffusion models to enhance control and guide diffusion processes. For instance, RPG [49] utilizes an LVLM as a planner to manipulate cross-attention layers in the diffusion model, while Demon [50] demonstrates that LVLMs can guide diffusion in alignment with a given persona. In contrast, our approach leverages LVLMs' ability to comprehend stitched images, utilizing this capability to enhance frame-to-frame dynamic understanding and applying it within the video domain to improve text-video alignment.

Human Preference Alignment via Reward Models Aligning with human preferences has improved generative quality in diffusion models through fine-tuning diffusion model using reward model gradients (DRaFT [6], AlignProp [30]) or policy gradients (DDPO [2], DPOK [10]). On the other hand,

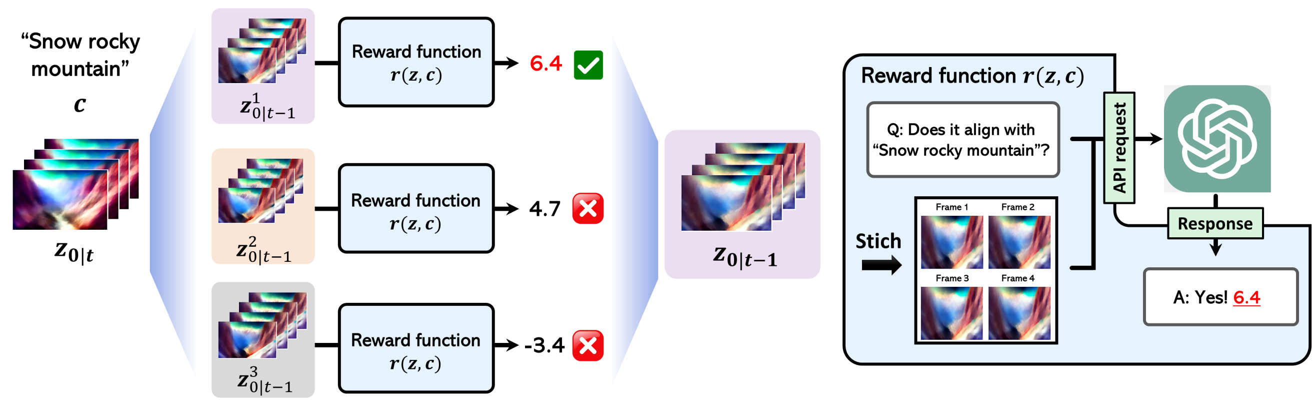

Figure 2. Overall pipeline of training-free gradient-free **Free²Guide**. Free²Guide leverages LVLMs' ability to comprehend stitched images, utilizing this capability to enhance frame-to-frame dynamic understanding and applying it within the video domain to improve text-video alignment. It also enables an effective ensemble approach that integrates large-scale image-based models to improve video generation guidance.

DOODL [42] and Demon [50] guide the denoising process to achieve text alignment without training diffusion models. Note, however, that the previously mentioned methods all focus on the image domain. Recent work VADER [31] fine-tunes a pre-trained video diffusion model using gradients of reward models for aesthetic and text-aligned generation. While this approach shows promising results for using video reward models, it demands substantial memory and does not utilize LVLMs. We address these limitations by proposing a text-video alignment method that approximates image reward gradients without fine-tuning.

Zeroth-order gradient approximation Zeroth-order gradients, or gradient-free approaches, approximate gradients of non-differentiable functions by evaluating multiple points [26, 28]. In diffusion-based inverse problems, methods like EnKF [51] and SCG [17] leverage gradient-free approximations to guide sampling based on non-differentiable or black-box forward models. However, there is a lack of research specifically focused on gradient-free approaches to guide sampling for video diffusion models. In video diffusion models, approximating a black-box reward model with a zeroth-order gradient is advantageous, as gradients of the reward are unavailable and the high-dimensional space of video data imposes memory limitations.

3. Preliminaries

3.1. Video Latent Diffusion Model

Video Latent Diffusion Models (VLDMs) learn a stochastic process by iteratively denoising random noise generated by the forward diffusion process [7]

$$q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{z}_0, \bar{\alpha}_t \mathbf{I}), \tag{1}$$

where $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x})$ is the latent encoding of the clean video with encoder $\mathcal{E}$ and $\bar{\alpha}_t$ is a noise scheduling coefficient at timestep t. The VLDM estimates the noise in $\boldsymbol{z}_t$ by minimizing the following objective:

$$\mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{\epsilon},t,\boldsymbol{c}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})\|^2\right], \tag{2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and $\boldsymbol{c}$ represents the conditioning input.

To retrieve a clean latent representation, we use a reverse-time Stochastic Differential Equation (SDE) sampling process:

$$\begin{aligned} d\boldsymbol{z}_t &= \bar{\boldsymbol{f}}(\boldsymbol{z}_t)dt + g(\boldsymbol{z}_t)\,d\bar{\mathbf{w}} \\ &= \left[\boldsymbol{f}(\boldsymbol{z}_t) - g(\boldsymbol{z}_t)^2\nabla_{\boldsymbol{z}_t}\log p(\boldsymbol{z}_t)\right]dt + g(\boldsymbol{z}_t)\,d\bar{\mathbf{w}}, \end{aligned} \tag{3}$$

where $\boldsymbol{f}$ and $\bar{\boldsymbol{f}}$ are the drift term for the forward SDE and reverse SDE, respectively, g is the diffusion coefficient, and $\bar{\mathbf{w}}$ represents a reverse time Wiener process. The initial point for reverse SDE is sampled from a normal Gaussian distribution. By discretizing the reverse SDE with an appropriate noise schedule, the VLDM retrieves a clean latent representation based on the DDIM [35] trajectory,

$$\begin{aligned} \sigma_t &:= \eta\sqrt{\left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\right)\left(1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}\right)} \\ \boldsymbol{z}_{0|t} &= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})\right) \\ \boldsymbol{z}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{z}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}) + \sigma_t\boldsymbol{\epsilon}, \end{aligned} \tag{4}$$

where σ_t controls the stochasticity of sampling, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and $\boldsymbol{z}_{0|t} = \mathbb{E}[\boldsymbol{z}_0|\boldsymbol{z}_t]$ denotes the posterior mean or denoised version of $\boldsymbol{z}_t$, computed by Tweedie's formula [8]. To transform the latent representation back to the video domain, a decoder $\mathcal{D}$ is used to decode the latent.

3.2. Guidance in Diffusion Model

Given the reverse SDE in Eq. (3), our goal is to obtain the optimal control $\boldsymbol{u}(\boldsymbol{z}_t)$:

$$d\boldsymbol{z}_t = \left[\bar{f}(\boldsymbol{z}_t) + \boldsymbol{u}(\boldsymbol{z}_t)\right] dt + g(\boldsymbol{z}_t)\, d\bar{\mathbf{w}}, \tag{5}$$

which directs the sampling process toward target distribution $p(\boldsymbol{z}_t|y)$, where y represent a desired condition, such as label, class or text prompt [45]. In classifier guidance [29], if an auxiliary classifier is available to estimate the likelihood $p(y|\boldsymbol{z}_t)$, the control term can be defined as

$$\boldsymbol{u}(\boldsymbol{z}_t) = -g(\boldsymbol{z}_t)^2 w \nabla_{\boldsymbol{z}_t} \log p(y|\boldsymbol{z}_t), \tag{6}$$

where w is a scaling factor that adjusts the strength of the guidance. This control term follows from applying the Bayes rule to express $p(\boldsymbol{z}_t|y) \propto p(\boldsymbol{z}_t|y)p(y|\boldsymbol{z}_t)^w$.

One might consider adapting classifier guidance by treating the reward model as a classifier. However, this approach presents two challenges: the reward model is not trained on noisy latent representations $\boldsymbol{z}_t$ and requires differentiability. To alleviate these limitations, we utilize a path integral control approach with zeroth-order gradient approximation, as described in the following Section 3.3.

3.3. Path Integral Control

Considering the diffusion model as an entropy regularized Markov Decision Process (MDP), we can conceptualize reverse SDE in the Reinforcement Learning (RL) framework [2, 10, 40] with the state $\boldsymbol{s}_t$ and the action $\boldsymbol{a}_t$ corresponding to the input $\boldsymbol{z}_t$. In this formula, the optimal policy p^* maximizes the following objective:

$$\mathbb{E}_p[\boldsymbol{r}(\boldsymbol{z}_0) - \alpha \sum_{\tau=T}^{1} D_{KL}(p(\boldsymbol{z}_{\tau-1}|\boldsymbol{z}_\tau)||p_\theta(\boldsymbol{z}_{\tau-1}|\boldsymbol{z}_\tau))], \tag{7}$$

where α is a coefficient of KL divergence with original policy p_θ defined by diffusion model. Let $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \sigma_t^2 \boldsymbol{I})$ be a reverse transition distribution in the SDE for the diffusion model and $p_\theta(\boldsymbol{z}_{0:t}) := p_\theta(\boldsymbol{z}_t)\Pi_{\tau=1}^{t} p(\boldsymbol{z}_{\tau-1}|\boldsymbol{z}_\tau)$. We can define a value function as

$$\begin{aligned}\exp\left(\frac{\boldsymbol{v}(\boldsymbol{z}_t)}{\alpha}\right) &= \int \exp\left(\frac{\boldsymbol{v}(\boldsymbol{z}_{t-1})}{\alpha}\right) p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) d\boldsymbol{z}_{t-1} \\ &= \mathbb{E}_{p_\theta(\boldsymbol{z}_{0:t})}\left[\exp\left(\frac{\boldsymbol{r}(\boldsymbol{z}_0)}{\alpha}\right)|\boldsymbol{z}_t\right],\end{aligned} \tag{8}$$

satisfying $\boldsymbol{v}(\boldsymbol{z}_0) = \boldsymbol{r}(\boldsymbol{z}_0)$ is a reward function [40].

The optimal control to address the entropy-regularized MDP system can be obtained by solving the Hamilton-Jacobi-Bellman (HJB) equation as follows [17, 41]:

$$\boldsymbol{u}(\boldsymbol{z}_t) = -\frac{\sigma_t^2 \nabla_{\boldsymbol{z}_t} \boldsymbol{v}(\boldsymbol{z}_t)}{\alpha}. \tag{9}$$

However, this term requires the gradient of the value function. To bypass the gradient requirements, one can use path integral control, which is an approach to estimate the optimal control (or guidance) based on the principles of stochastic optimal control [20, 39, 41]. In [17], the optimal control can be approximated as

$$\boldsymbol{u}(\boldsymbol{z}_t) \simeq -\frac{\mathbb{E}\left[\exp\left(\frac{\boldsymbol{r}(\boldsymbol{z}_0)}{\alpha}\right)(\boldsymbol{z}_{t-1} - \boldsymbol{\mu}_t)|\boldsymbol{z}_t\right]}{\mathbb{E}\left[\exp\left(\frac{\boldsymbol{r}(\boldsymbol{z}_0)}{\alpha}\right)|\boldsymbol{z}_t\right]}. \tag{10}$$

While SCG [17] utilizes this optimal control with diffusion models to solve inverse problems in image domain, we aim to use LVLMs to guide videos toward improved text alignment.

4. Free²Guide

In this section, we introduce Free²Guide, a framework that uses a non-differentiable reward model to guide video generation during the sampling process. In Sec. 4.1, we discuss how to apply image-based reward models, including LVLM, for text-video alignment. Sec. 4.2 outlines methods for ensembling multiple reward models to achieve synergistic effects. Finally, we interpret the diffusion model as an entropy-regularized MDP and describe its practical implementation (Sec. 4.3).

4.1. Video Guidance leveraging Image LVLMs

Motivation By leveraging the path integral control approach discussed in Sec. 3.3, we can guide the reverse process without relying on the gradient of the reward function. If the reward model $\boldsymbol{r}$ in Eq. (10) assesses the alignment of the generated video with the text prompt, it can help steer the video output to enhance fidelity to the prompt. However, due to the complexity of videos compared to static images, there are limited large-scale models specifically trained for video and text alignment. We analyze the impact of video-based reward models on video guidance and find that their effectiveness is limited (see Appendix D.1).

Applying these image-based reward models directly for video guidance, of course, presents challenges. Image-based models are not designed to process time-dependent features, such as motion, flow, and dynamics, so specific adaptations are required for these models to assess text-video alignment. As shown in Algorithm 1, we calculate the reward for a video by summing frame-by-frame rewards from the image-based model. This approach enables alignment with spatial information within individual video frames but still lacks guidance on temporal dynamics.

Image-based LVLMs as a Video Reward Model Although LVLMs are trained on static image-text data, their extensive pretraining across diverse visual contexts enables them to implicitly capture elements of motion. As shown in

Table 1 of [38], treating video as an image grid in LVLMs strongly correlates with human evaluation. Furthermore, results from [22, 24] demonstrate that image-based LVLMs achieve performance comparable to video-specific LLMs in video QA, validating our approach.

Accordingly, to adapt LVLMs for evaluating multiple frames simultaneously, we employ a method called *stitching*, which combines key frames into a single composite image (see Fig. 2). Specifically, we first sample key frames from the video and arrange them in a structured grid layout, labeling each frame with its index to indicate its position in the sequence. This approach allows LVLMs to process temporal information by leveraging spatial relationship between frames.

Then, to help LVLMs understand frame order within the composite image, we provide explicit sequence instructions through a system prompt. This efficient adaptation enables LVLMs to recognize frame order by referencing frame numbers rather than processing them linearly. We incorporate Zero-shot Chain-of-Thought [23] in the system prompt to enhance reasoning ability and mitigate hallucinations. In the user prompt, we instruct the LVLM to consider every key frame individually and evaluate the alignment score between the composite image and the text prompt on a scale of 1 to 9. The full system instructions and query templates are detailed in Appendix A.

4.2. Ensembling Reward Functions

Unlike gradient-based guidance, our method significantly reduces memory requirements by avoiding the computationally intensive backpropagation process. This enables us to concurrently employ multiple rewards for sampling guidance, potentially leading to synergistic benefits with large-scale image models. We explore ensemble methods that allow LVLMs to incorporate temporal information, thereby supporting more effective guidance for video alignment when combined with large-scale image models. Note that Demon [50], a concurrent work that also proposed ensemble rewards, failed to show the synergy effect of ensemble and did not have to handle temporal information.

Given the n videos $\{V_i\}_{i=1}^n$, we propose three ensembling methods to combine multiple reward models: Weighted Sum, Normalized Sum, and Consensus.

- **Weighted Sum**: This method combines the outputs by computing a fixed weighted sum, allowing us to control the influence of each reward model.

$$\text{Reward}_{\text{ens}}(V_i, \boldsymbol{r}_1, \boldsymbol{r}_2) = \beta \boldsymbol{r}_1(V_i) + (1-\beta)\boldsymbol{r}_2(V_i), \quad (11)$$

where $\beta \in [0,1]$ is a constant weight factor that balances the contributions of reward models $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$.

- **Normalized Sum**: To ensure a balanced contribution of each reward models, we first normalize each reward's output to the range $[0,1]$, then sum these normalized values to get the final ensemble reward.

$$\text{Reward}_{\text{ens}}(V_i, \boldsymbol{r}_1, \boldsymbol{r}_2) = \sum_{\boldsymbol{r}} \frac{\boldsymbol{r}(V_i) - \min(\boldsymbol{r}(V_i))}{\max(\boldsymbol{r}(V_i)) - \min(\boldsymbol{r}(V_i))}, \quad (12)$$

where $\max(\boldsymbol{r}), \min(\boldsymbol{r})$ represents the maximum and minimum score from n reward outputs.

- **Consensus**: Inspired by the Borda count [9], each reward model ranks the videos from best to worst, assigning $\text{points}_{\boldsymbol{r}}$ based on their rank. The top-ranked video receives the maximum points, down to 1 point for the lowest rank. The total reward for each video V_i is the sum of points

Algorithm 1 Reward Model $\boldsymbol{r}(\mathcal{D}(\boldsymbol{z}_{0|t}), \boldsymbol{c})$

Require: Reward function $\boldsymbol{r}$, condition $\boldsymbol{c}$, prompt $\boldsymbol{p}$, decoded frames $\boldsymbol{x}_{0|t} := \mathcal{D}(\boldsymbol{z}_{0|t})$, and key frames $k \subset [1, N]$

if $\boldsymbol{r}$ is CLIP **then**
  reward $\leftarrow \sum_{i \in k} \mathtt{sim}(\boldsymbol{r}(\boldsymbol{x}^i_{0|t}), \boldsymbol{r}(\boldsymbol{c}))$
else if $\boldsymbol{r}$ is ImageReward **then**
  reward $\leftarrow \sum_{i \in k} \boldsymbol{r}(\boldsymbol{x}^i_{0|t}, \boldsymbol{c})$
else if $\boldsymbol{r}$ is LVLM **then**
  reward $\leftarrow \boldsymbol{r}(\mathtt{concat}_{i \in k}(\boldsymbol{x}^i_{0|t}), \boldsymbol{c}, \boldsymbol{p})$
end if
return reward

Algorithm 2 Free^{2}Guide

Require: Video diffusion model $\boldsymbol{\epsilon}_\theta$, reward function $\boldsymbol{r}$, decoder $\mathcal{D}$, noise scheduling parameter $\{\bar{\alpha}_t\}_{t=1}^T, \{\sigma_t\}_{t=1}^T$

for $t = T$ **to** 1 **do**
  $\boldsymbol{z}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_{t-1}}}\left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t)\right)$
  $\hat{\boldsymbol{z}}_{t-1} \leftarrow \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t)$
  $\boldsymbol{\epsilon}^1, \cdots, \boldsymbol{\epsilon}^n \sim \mathcal{N}(0, \mathbf{I})$
  $\boldsymbol{z}^i_{t-1} \leftarrow \hat{\boldsymbol{z}}_{t-1} + \sigma_t \boldsymbol{\epsilon}^i$
  $\boldsymbol{z}^i_{0|t-1} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_{t-1}}}\left(\boldsymbol{z}^i_{t-1} - \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}^i_{t-1})\right)$
  $\boldsymbol{r}_1 \leftarrow$LVLM
  if Ensemble **then**
    $\boldsymbol{r}_2 \in \{$CLIP, ImageReward$\}$
    $j \leftarrow \text{argmax}_i \ \text{Reward}_{\text{ens}}(\mathcal{D}(\boldsymbol{z}^i_{0|t-1}), \boldsymbol{r}_1, \boldsymbol{r}_2)$ From Sec. 4.2.
  else
    $j \leftarrow \text{argmax}_i \ \boldsymbol{r}_1(\mathcal{D}(\boldsymbol{z}^i_{0|t-1}), \boldsymbol{c})$
  end if
  $\boldsymbol{z}_{t-1} \leftarrow \boldsymbol{z}^j_{t-1}$
end for
return $\boldsymbol{z}_0$

from both reward model.

$$\text{Reward}_{\text{ens}}(V_i, \boldsymbol{r}_1, \boldsymbol{r}_2) = \text{points}_{\boldsymbol{r}_2}(V_i) + \text{points}_{\boldsymbol{r}_1}(V_i). \tag{13}$$

4.3. Guidance using Path Integral Control

To guide the reverse sampling process without computing the gradient of the reward function, we utilize the framework outlined in Eq. (10). However, the expectation of the reward function in Eq. (10) demands extensive network function evaluations (NFE) by solving complex differential equations, such as PF-ODE [36]. Inspired by [17], we instead apply the DPS [5] approach to approximate Eq. (8) by using the posterior mean of $\boldsymbol{z}_t$, as defined in Eq. (4). Following DPS, we can set $p(\boldsymbol{z}_{0:t}) = \delta(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z}_0|\boldsymbol{z}_t])$ using Direc delta distribution δ in which case Eq. (10) becomes:

$$\boldsymbol{u}(\boldsymbol{z}_t) \simeq -\frac{\mathbb{E}_{p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)}\left[\exp\left(\frac{\boldsymbol{r}(\boldsymbol{z}_{0|t-1})}{\alpha}\right)(\boldsymbol{z}_{t-1} - \boldsymbol{\mu}_t)\right]}{\exp\left(\frac{\boldsymbol{r}(\boldsymbol{z}_{0|t})}{\alpha}\right)}. \tag{14}$$

To approximate this expectation using the Monte Carlo method, we sample n different $\boldsymbol{z}_{t-1}$ through the reverse SDE as outlined in Eq. (4). Then we assume $\alpha \to 0$ to obtain optimal control. Under this assumption, Eq. (3) becomes equivalent to selecting the $\boldsymbol{z}_{t-1}$ that maximizes the reward of $\boldsymbol{z}_{0|t-1}$ [17]. While [17] arbitrarily weighted the reward function and assumed the weight to be zero, we interpret this as relaxing the entropy-regularization term in Eq. (7) by defining the diffusion process as an entropy-regularized MDP. In practical terms, this approach eliminates careful parameter exploration by selecting $\boldsymbol{z}_{t-1}$ with the largest reward.

By following this adjusted sampling strategy as described in Algorithm 2, Free^{2}Guide can efficiently steer video generation towards better alignment with the reward signals.

5. Experiments

Baselines and Sampling Strategy. We use open-source text-to-video diffusion models, LaVie [43] and VideoCrafter2 [4], as baseline models. The generated videos contain 16 frames with a resolution of 320×512. We employ LVLM as `GPT-4o-2024-08-06` [1] using OpenAI APIs. We employ two large-scale models CLIP [32] and ImageReward [48], to validate that LVLM's capability to account for temporal dynamics can enhance text-video alignment when used alongside large-scale image reward models. In CLIP, we can assess alignment by measuring cosine similarity between text and image embeddings. On the other hand, we can use ImageReward output as an reward since it predicts human preference for image-text pairs. For adaptation to the video domain, we extract key frames from each denoised video and sum the reward for each frame to evaluate overall alignment, as outlined in Algorithm 1.

We employ stochastic DDIM sampling with $\eta = 1$ in Eq. (4) for a total of $T = 50$ steps and apply classifier-free guidance [14] using a guidance scale of $w = 7.5$ for LaVie and $w = 12$ for VideoCrafter2. The number of samples at each guidance step is set to $n = 5$ for LaVie and $n = 10$ for VideoCrafter2. Guidance is applied during the early sampling steps, specifically within $t \in [T, T-5]$. In the weighted sum ensemble, we assign a weight of $\beta = 0.75$ to the LVLM reward.

Text Alignment Evaluation. We conduct quantitative evaluation using VBench [18], a benchmark designed to evaluate the alignment of text-to-video (T2V) models with respect to a text prompt. Our evaluation protocol measures text alignment across six dimensions: Appearance Style, Temporal Style, Human Action, Multiple Objects, Spatial Relationship and Overall Consistency. For a fair comparison, we use standardized prompts for each metric, ensuring consistent conditions across different models.

General Video Quality Evaluation. In addition to text alignment, we evaluate the general quality of generated videos independently of text prompts using six metrics in VBench: Subject Consistency, Background Consistency, Motion Smoothness, Dynamic Degree, Aesthetic Quality, and Imaging Quality.

Video-specific Attributes. Since VBench prompts involve limited movement, we conducted additional experiments using T2V-CompBench [38] to analyze video-specific motion and dynamics. We measure Dynamic Attribution Binding, which evaluates how well models handle state changes (*e.g.* shape and texture) and color variations over time.

5.1. Results

In this section, we present both qualitative and quantitative results to demonstrate the effectiveness of our method. The top four rows of Fig. 3 shows visual comparisons between the baseline and reward models. We observe that leveraging the GPT-4o model to assess text-video alignment improves alignment with respect to temporal dynamics (*e.g.* "tilt down") and semantic representation (*e.g.* "A and B"). These results indicate that LVLM can account for temporal information by processing multiple sub-frames of video simultaneously, with strong performance in spatial understanding.

Building on LVLMs' capability to account for temporal dynamics, we validate the feasibility of ensembling techniques that integrate guidance from large-scale image mod-

Method	Avg.
LaVie + CLIP	0.5712
+ $\text{GPT}_{\text{Weighted Sum}}$	**0.5738**
+ $\text{GPT}_{\text{Normalized Sum}}$	0.5734
+ $\text{GPT}_{\text{Consensus}}$	0.5679

Method	Avg.
LaVie + ImageReward	0.5676
+ $\text{GPT}_{\text{Weighted Sum}}$	**0.5726**
+ $\text{GPT}_{\text{Normalized Sum}}$	0.5715
+ $\text{GPT}_{\text{Consensus}}$	0.5692

Table 1. Qualitative comparison between ensemble methods.

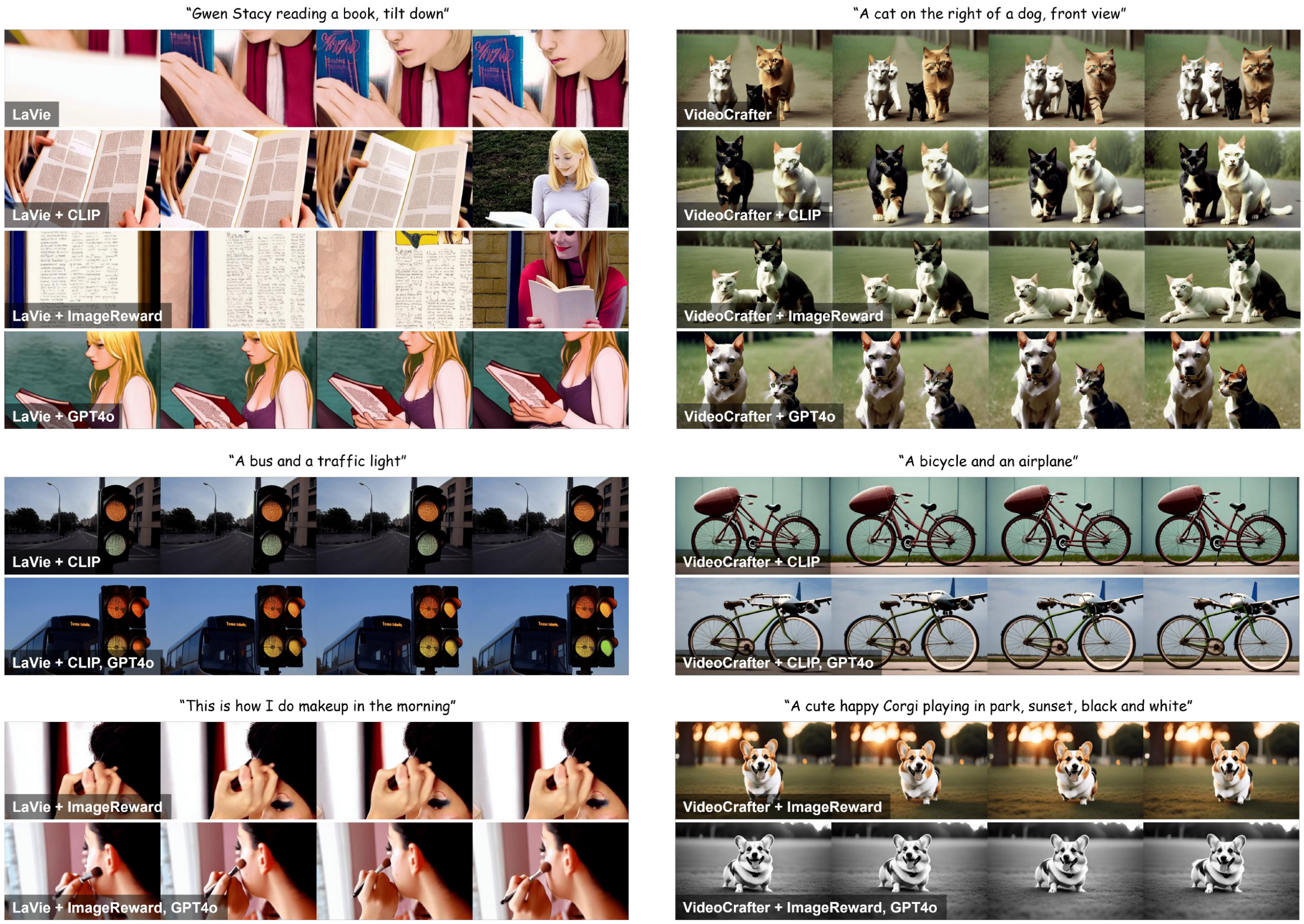

Figure 3. Qualitative results of our method. Comparison with LaVie on the left and VideoCrafter2 on the right.

Method	Appearance Style	Temporal Style	Human Action	Multiple Objects	Spatial Relationship	Overall Consistency	**Avg.**
LaVie [43]	0.2312	0.2502	0.9300	0.2027	0.3496	0.2694	0.3722
+ GPT	0.2366 (+2.3%)	0.2508 (+0.2%)	0.9300 (-0.0%)	0.2546 (+25.6%)	0.3531 (+1.0%)	0.2709 (+0.6%)	0.3827
+ CLIP	0.2370 (+2.5%)	0.2490 (-0.5%)	0.9400 (+1.1%)	0.2607 (+28.6%)	0.3074 (-12.1%)	0.2738 (+1.6%)	0.3780
++ GPT	0.2350 (+1.6%)	0.2487 (-0.6%)	1.000 (+7.5%)	0.2447 (+20.7%)	0.3180 (-9.0%)	0.2742 (+1.7%)	**0.3868**
+ ImageReward	0.2360 (+2.1%)	0.2483 (-0.8%)	0.9300 (-0.0%)	0.2637 (+30.1%)	0.2614 (-25.2%)	0.2728 (+1.2%)	0.3687
++ GPT	0.2373 (+2.6%)	0.2497 (-0.2%)	0.9400 (+1.1%)	0.2462 (+21.4%)	0.3014 (-13.8%)	0.2772 (+2.9%)	0.3753
VideoCrafter2 [4]	0.2490	0.2567	0.9300	0.3880	0.3760	0.2778	0.4129
+ GPT	0.2504 (+0.6%)	0.2568 (+0.0%)	0.9500 (+2.2%)	0.4878 (+25.7%)	0.4225 (+12.4%)	0.2872 (+3.4%)	0.4425
+ CLIP	0.2542 (+2.1%)	0.2621 (+2.1%)	0.9300 (-0.0%)	0.4261 (+9.8%)	0.2923 (-22.3%)	0.2802 (+0.9%)	0.4075
++ GPT	0.2490 (+0.0%)	0.2612 (+1.8%)	0.9600 (+3.2%)	0.4474 (+15.3%)	0.3361 (-10.6%)	0.2837 (+2.1%)	0.4229
+ ImageReward	0.2513 (+0.9%)	0.2574 (+0.3%)	0.9700 (+4.3%)	0.4733 (+22.0%)	0.4264 (+13.4%)	0.2826 (+1.7%)	0.4435
++ GPT	0.2533 (+1.7%)	0.2607 (+1.6%)	0.9400 (+1.1%)	0.5160 (+33.0%)	0.4371 (+16.3%)	0.2828 (+1.8%)	**0.4483**

Table 2. Quantitative evaluation on text alignment. Higher numbers indicate better alignment with the text prompt. The numbers in parentheses denote the performance difference from the baseline.

els to improve text-video alignment. This approach enables LVLMs to process temporal information, enhancing the quality of guidance. In Table 1, we explore the most effective ensemble method by comparing average scores on text alignment and general video quality evaluation from VBench. We find that assigning more weight to LVLM outperformed the alternative of balancing model contributions equally in the ensemble, indicating that the role of LVLM is significant. Thus, we adopt the weighted sum ensemble as the default setting. The bottom four rows of Fig. 3 also illustrate qualita-

Method	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Avg.
LaVie [43]	0.9450	0.9689	0.9718	0.4799	0.5687	0.6611	0.7659
+ GPT	0.9470	0.9693	0.9742	0.4725	0.5726	0.6615	0.7662
+ CLIP	0.9495	0.9712	0.9735	0.4560	0.5727	0.6637	0.7644
++ GPT	0.9622	0.9781	0.9804	0.3703	0.5951	0.6795	0.7609
+ IR	0.9443	0.9681	0.9732	0.4872	0.5664	0.6605	0.7666
++ GPT	0.9758	0.9813	0.9832	0.5165	0.5662	0.6530	**0.7699**
VC2 [4]	0.9658	0.9748	0.9818	0.3846	0.5860	0.6772	**0.7617**
+ GPT	0.9746	0.9800	0.9827	0.2949	0.5977	0.6924	0.7537
+ CLIP	0.9762	0.9816	0.9839	0.2491	0.6037	0.6886	0.7472
++ GPT	0.9770	0.9823	0.9838	0.2399	0.6042	0.6878	0.7458
+ IR	0.9739	0.9801	0.9828	0.2711	0.5994	0.6857	0.7488
++ GPT	0.9758	0.9813	0.9832	0.2564	0.6039	0.6877	0.7480

Table 3. Comparison of the general quality of the generated video independent of the text prompt. Higher numbers indicate better video quality. 'VC2' is VideoCrafter2 and 'IR' is ImageReward.

tive results for ensembling, showing that combining GPT-4o with other image reward models accurately resolves issues related to dynamics or multiple objects that standalone reward models struggle to properly identify, while maintaining overall structure.

For more detailed evaluations, we compare the quantitative results in Table 2 to assess text-video alignment. Analysis of the average evaluation scores reveals that incorporating LVLM consistently outperforms configurations that exclude it. Specifically, we observe the most significant improvement in handling Spatial Relationship across baselines. Since CLIP has a limited zero-shot spatial reasoning capability [37], the text alignment performance decreases in Spatial Relationship when using CLIP alone. However, ensembling with LVLM offers additional cues that help CLIP to better account for spatial semantics, leading to performance improvements. Furthermore, incorporating LVLM enhances Human action, Overall Consistency in overall case and Temporal Style, except when using CLIP as the reward model. Since LVLM can understand temporal nuances by processing multiple frames at once, it improves performance by supporting the alignment of temporal movement.

Additionally, we compare general video quality in Table 3. We confirm that even without explicit guidance for consistency or motion, alignment with text prompts improves most quality metrics except for Dynamic Degree. This metric often trades off with consistency but can be improved by ensembling GPT-4o with ImageReward in the LaVie model. This suggests that ImageReward compensates for the performance drop in Dynamic Degree that GPT-4o alone does not address, resulting in the best performance.

Method	Dynamic Attribution (↑)
LaVie	0.01242
+ GPT	0.01360
VC2	0.00663
+ GPT	0.00770

Table 4. Results for T2V-CompBench.

Figure 4. Example of T2V-CompBench.

As shown in Table 4, leveraging LVLM improves performance in Dynamic Attribution Binding. Figure 4 illustrates an example video where the water gradually fills up over time in response to a given prompt when utilizing LVLM, whereas the baseline model fails to capture this progression.

Method	NFEs	Avg.
Baseline	100	0.5815
Best-of-N	100	0.5802
Ours	100	**0.5981**

Table 5. Fixed NFE comparison on VBench.

Method	CLIP (↑)	IR (↑)	GPT (↑)
VC2	30.39	-0.10	7.09
+GPT	30.90	0.23	7.28
+CLIP	**30.96**	0.14	7.11
++GPT	30.95	0.20	7.07
+IR	30.92	0.22	7.28
++GPT	**30.96**	**0.28**	**7.33**

Table 6. Reward robustness.

5.2. Analysis

Computational efficiency To evaluate the computational efficiency of our method, we conduct experiments under a fixed NFE budget of 100 using VideoCrafter2, as shown in Table 5. The Baseline uses a single 100-step inference path, while Best-of-N selects the highest LVLM reward from two 50-step paths. Our approach uses 50 steps, with six samples in the first 10 steps, while the remaining 40 steps follow the baseline procedure. Notably, simply selecting from multiple final outputs is ineffective, as it does not influence the denoising process. In contrast, our method actively guides generation throughout sampling, leading to improved text alignment and coherence that cannot be achieved through post-hoc selection.

Robustness of Rewards We verify that our method achieves robust performance without overfitting to any particular reward, avoiding reward hacking, a common issue in RL literature. Table 6 compares the rewards for the final video outputs generated by each method. Video guidance ensembled with LVLM generally achieves higher metrics, exhibiting a trend similar to the text alignment results in Table 2. These findings indicate that the ensemble approach is not over-optimized for a particular reward, enhancing robustness across diverse evaluation criteria. Additional ablation studies can be found in Appendix C.

6. Conclusion

In this paper, we introduced Free^{2}Guide, a novel gradient-free framework to enhance text-video alignment in diffusion-based generative models without relying on reward gradients. By approximating the gradient of the reward function, Free^{2}Guide effectively integrates non-differentiable reward models, including powerful black-box LVLMs, to steer the video generation process towards better alignment. Our experiments demonstrate that Free^{2}Guide consistently improves alignment with text prompts and general video quality. By enabling ensembling with LVLM, our method benefits from synergistic effects, further enhancing performance.

7. Acknowledgments

This work was supported by the National Research Foundation of Korea under Grant RS-2024-00336454, and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2024-00457882, AI Research Hub Project; No. RS-2025-02304967, AI Star Fellowship (KAIST))

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2, 4

[3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 2

[4] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 2, 6, 7, 8

[5] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. 6

[6] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 2

[7] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 1, 3

[8] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 3

[9] Peter Emerson. The original borda count and partial voting. *Social Choice and Welfare*, 40(2):353–358, 2013. 5

[10] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[11] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[12] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 1

[13] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[17] Yujia Huang, Adishree Ghatare, Yuanzhe Liu, Ziniu Hu, Qinsheng Zhang, Chandramouli S Sastry, Siddharth Gururani, Sageev Oore, and Yisong Yue. Symbolic music generation with non-differentiable rule guided diffusion. *arXiv preprint arXiv:2402.14285*, 2024. 2, 3, 4, 6

[18] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 6

[19] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2

[20] Hilbert J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011, 2005. 4

[21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 1

[22] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 2024. 5

[23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 5

[24] K Li et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 5

[25] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 2

[26] Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020. 3

[27] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[28] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. 3

[29] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 4

[30] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 2

[31] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024. 2, 3

[32] 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*, pages 8748–8763. PMLR, 2021. 6

[33] 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*, pages 10684–10695, 2022. 1

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR*, 2021. 3

[36] 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*, 2021. 1, 6

[37] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 8

[38] K Sun et al. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv*, 2024. 5, 6

[39] Evangelos Theodorou, Jonas Buchli, and Stefan Schaal. A generalized path integral control approach to reinforcement learning. *The Journal of Machine Learning Research*, 11: 3137–3181, 2010. 4

[40] Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024. 4

[41] Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024. 4

[42] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7280–7290, 2023. 2, 3

[43] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2, 6, 7, 8, 3

[44] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[45] David Williams and L Chris G Rogers. *Diffusions, Markov processes, and martingales*. John Wiley & Sons, 1979. 4

[46] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336, 2024. 2

[47] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2

[48] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6

[49] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024. 2

[50] Po-Hung Yeh, Kuang-Huei Lee, and Jun-Cheng Chen. Training-free diffusion model alignment with sampling demons. *arXiv preprint arXiv:2410.05760*, 2024. 2, 3, 5

[51] Hongkai Zheng, Wenda Chu, Austin Wang, Nikola Kovachki, Ricardo Baptista, and Yisong Yue. Ensemble kalman diffusion guidance: A derivative-free method for inverse problems. *arXiv preprint arXiv:2409.20175*, 2024. 2, 3

[52] Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 567–578, 2023. 2