

EIDT-V: Exploiting Intersections in Diffusion Trajectories for Model-Agnostic, Zero-Shot, Training-Free Text-to-Video Generation

Diljeet Jagpal¹, Xi Chen^{1,2,*}, and Vinay P. Namboodiri¹

¹University of Bath; ²Fudan University

{dkjj20, xc841, vpn22}@bath.ac.uk, x.chen@fudan.edu.cn



Eight equally spaced frames from 24-frame GIFs generated by our EIDT-V model. Top row shows SD3 Medium [8] results for prompt: “A peacock displaying its feathers”. Bottom row shows SDXL [32] results for prompt: “A child blowing bubbles that float and pop gently”. These examples highlight the model’s ability to generate high-quality videos with semantic and temporal coherence.

Abstract

Zero-shot, training-free, image-based text-to-video generation is an emerging area that aims to generate videos using existing image-based diffusion models. Current methods in this space require specific architectural changes to image-generation models, which limit their adaptability and scalability. In contrast to such methods, we provide a model-agnostic approach. We use intersections in diffusion trajectories, working only with the latent values. We could not obtain localized frame-wise coherence and diversity using only the intersection of trajectories. Thus, we instead use a grid-based approach. An in-context trained LLM is used to generate coherent frame-wise prompts; another is used to identify differences between frames. Based on these, we obtain a CLIP-based attention mask that controls the timing of switching the prompts for each grid cell. Earlier switching results in higher variance, while later switching results in more coherence. Therefore, Our approach can ensure appropriate control between coherence and variance for the frames. Our approach results in state-of-the-art perfor-

mance while being more flexible when working with diverse image-generation models. The empirical analysis using quantitative metrics and user studies confirms our model’s superior temporal consistency, visual fidelity and user satisfaction, thus providing a novel way to obtain training-free, image-based text-to-video generation. Further examples and code at <https://djagpal02.github.io/EIDT-V/>

1. Introduction

Recent advances in image generation have established diffusion models as state-of-the-art (SOTA) tools for producing visually compelling and coherent images. Open source techniques such as Stable Diffusion [8, 32, 35] and Flux [10] show efficient high-quality outputs made possible by the diffusion models [23, 41]. Despite this, extending these capabilities to video remains challenging due to the unique temporal coherence and dynamic motion requirements.

SOTA, such as OpenAI’s Sora [1] and Meta’s MovieGen [33], have made significant progress in high-quality text-to-video generation. However, they rely on extensive training and complex architectures, which limits accessibility. These models typically require high-end GPUs or sit behind paywalls, making them out of reach for most users. Although lower cost models [2–4, 12, 14, 15, 18, 19, 26,

* Corresponding author.

Supported by EPSRC (EP/T518013/1, EP/Y021614/1); AAM available under a CC BY licence.

[27, 46, 49] exist, the difference in quality is significant, and the inference and training costs are still beyond the reach of most. Hence, research is also being pursued to explore training-free approaches [20, 21, 24]. These approaches require architectural adjustments tied to specific diffusion models, limiting flexibility and scalability while producing even lower-quality outputs.

We propose a model-agnostic, zero-shot text-to-video generation approach using image-based diffusion models. Our method operates entirely in the latent space, achieving compatibility with various image-based diffusion models without modifications or additional training. This model-agnostic design enables high-quality video generation across different architectures, establishing a flexible and robust foundation for video generation.

Contributions: The main contributions of our work are:

- **Grid-Based Prompt Switching for Conditional Image Generation:** We develop a novel grid-based prompt switching technique specifically for conditional image generation. This method divides each image into localized grid regions, providing fine-grained control over variance.
- **Enhanced Text control via LLM Modules :** Our approach incorporates two in-context trained modules within a LLM: one module generates frame-wise prompts from text inputs, while the other detects differences between consecutive frames to guide temporal consistency.
- **CLIP-based attention masking** - The text-guided difference between consecutive frames is used through a CLIP-based attention masking module that generates the prompt switch times, thereby controlling the variance and coherence spatially and between frames.

All these contributions result in a complete pipeline thoroughly validated compared with corresponding baselines and ablations. We provide more comprehensive metrics to validate the generated results’ temporal consistency and visual fidelity, along with suitable user studies.

2. Related Works

2.1. Image Diffusion

Diffusion models [6, 17, 38, 39, 41] have become essential for generating high-quality images, offering improved stability and scalability over previous SOTA GANs [11]. With continuous advancements [28, 42], diffusion models are evolving rapidly. The open-source Stable Diffusion (SD) models illustrate this progression: moving from SD1 [35] to SDXL [32], the architecture became significantly more complex. In SD3 [8], the original UNet architecture [36] was replaced by a multi-modal version of the diffusion image transformer [31] and enhanced with rectified flows [28]. These rapid architectural shifts often render tools for earlier versions obsolete, underscoring the need to develop adapt-

able tools.

2.2. State-of-the-Art Video Generation Models

SOTA video generation models, such as Runway’s Gen-3 Alpha [37], Meta’s MovieGen [33], and others [25], [1], have demonstrated high-quality video synthesis with solid temporal coherence. However, they are costly to train and for inference, highlighting the need for computationally efficient alternatives.

2.3. Low Cost Video Generation

Initial efforts, such as VDM [19] and MagicVideo [49], adapted image diffusion processes to video, with MagicVideo leveraging latent spaces to lower computational requirements. Multi-stage models, like Imagen Video [18], utilize cascaded sub-models to enhance video resolution and frame rate, but the complexity of these stages keeps computational demands high. Similarly, transformer-based approaches, such as WALT [14], employ memory-efficient windowed attention, though maintaining high-resolution coherence across frames still requires substantial resources.

Models like LVDM [15] and Video LDM [4] use hierarchical and latent diffusion techniques to extend video length, aiming for efficiency but still facing scalability challenges. Even efficiency-focused methods like Latent-Shift [2] and Ed-t2v [26] that enhance motion fidelity with techniques such as temporal shift modules and identity attention remain relatively expensive, making truly low-cost, accessible video generation a continuing challenge.

2.4. Zero-Shot Video Generation

Zero-shot video generation uses pre-trained image models for video synthesis tasks without additional training, enhancing accessibility. Text2Video-Zero [24] was an early example, utilizing approximated optical flow and replacing self-attention with cross-attention to maintain frame continuity, although its limited motion approximation reduces scalability. Subsequent models such as Free-Bloom [21] and DirecT2V [20] improved semantic coherence by using LLMs to generate frame-wise prompts. However, the LLMs in these models were manually configured and lacked standardized frameworks, which affects reproducibility. Moreover, these approaches are heavily tied to specific diffusion architectures, making them susceptible to becoming obsolete as diffusion models evolve.

3. Background

3.1. Foundational Concepts: Diffusion Models

Diffusion models transform an initial noise sample X_T into a final data sample X_0 through iterative denoising over a time horizon T . Two main approaches are denoising diffusion probabilistic models (DDPMs) [17, 38], which pro-

gressively remove noise in discrete steps, and score-based models [40], which leverage gradients of data densities to guide the denoising process. Karras et al. [23] unified these methods into a modular framework, showing that both aim to map noise to data through a continuous denoising trajectory.

By treating the diffusion steps as continuous over $t \in [0, T]$, Song et al. [41] reformulated diffusion as Stochastic Differential Equations (SDEs), where the reverse SDE describes the process of recovering data from noise. They further introduced Probability Flow ODEs (PF-ODEs) [41], which offer a deterministic path from noise to data while preserving the identical marginal distributions as the SDE. This continuous, deterministic trajectory simplifies the sampling process, making diffusion models efficient for high-quality image generation.

3.2. Classifier-Free Guidance

Classifier-free guidance [16] enables conditioning diffusion models on text without relying on an external classifier. During training, the model alternates between a specific condition y (e.g., text) and a null condition $\emptyset$, learning both the conditional $s_\theta(x_t, t | y)$ and unconditional $s_\theta(x_t, t | \emptyset)$ score functions. The final conditional score function at sampling is given by:

$$\begin{aligned} \nabla_{x_t} \log p_{X_t|Y}(x_t|y) &\approx s_\theta(x_t, t | y) \\ &+ \gamma (s_\theta(x_t, t | y) - s_\theta(x_t, t | \emptyset)), \end{aligned} \quad (1)$$

where γ controls the influence of the text condition. Adjusting the guidance scale, γ , allows the diffusion process to produce images that align closely with the text prompt.

3.3. Uniqueness of Diffusion Trajectories

In the deterministic framework of ODE-based diffusion, each trajectory—initiated from a state x_T under a given condition y —uniquely determines the final output x_0 , provided the ODE satisfies Lipschitz continuity [23, 41]. This property ensures that diffusion models using ODE formulations yield consistent trajectories for a given condition.

4. Methodology

We present the EIDT-V pipeline, illustrated in Fig. 1. This pipeline leverages diffusion intersections to enable frame-based video generation. The following subsections detail each component, beginning with the foundational intuition behind our approach.

4.1. Prompt Switching via Diffusion Intersection

Consider two vehicles, A and B , each travelling backwards in time from an initial time T toward a destination at time 0. The position of vehicle i at time t is denoted by $\mathbf{x}_t^{(i)}$,

where $i \in \{A, B\}$, and each vehicle follows a deterministic trajectory:

$$\mathbf{x}_t^{(i)} = \mathbf{f}(\mathbf{x}_T^{(i)}, t, y_i), \quad (2)$$

where $\mathbf{x}_T^{(i)}$ is the starting position, and y_i specifies the guidance or “route” for vehicle i .

If the trajectories of A and B intersect at time $t = t_s$, a dependency forms, constraining their maximum separation at the destination, $t = 0$. Assuming each vehicle moves with speed v , this maximum separation at $t = 0$ is given by the distance function $D(t_s) = 2 \cdot v \cdot t_s$, such that:

$$\|\mathbf{x}_0^{(A)} - \mathbf{x}_0^{(B)}\| \leq D(t_s). \quad (3)$$

In diffusion models, this analogy applies to image synthesis, where each trajectory represents the ODE evolution of an image from an initial noisy state x_T to a coherent structure x_0 , guided by a prompt y . By switching the Prompt from y to y' at time $t = t_s$, we create a similar intersection point that limits divergence between the resulting images, allowing us to control how much each prompt influences the final image.

4.2. Grid Prompt Switching

Prompt switching provides global variance control; however, in some cases, we may want more targeted variance. To address this, we introduce **grid prompt switching**. In this method, we divide the image into an $n \times n$ grid, where each cell (i, j) is assigned its own prompt switch time $t_s^{(i,j)}$. This split allows specific image regions to adopt prompt changes independently, enabling fine-grained control over which parts of the image transition to a new prompt.

Prompt switching for each grid cell individually can lead to inconsistencies, as the diffusion process is inherently global and interconnected throughout the image. We propose a hybrid approach combining the original diffusion trajectory with the updated one to maintain spatial coherence while allowing for localized prompt transitions.

As illustrated in Fig. 2, we implement this by first defining a *Switch Time Matrix* (STM), which determines the specific time $t_s^{(i,j)}$ at which each cell switches prompts. This matrix is compared with the current timestep t to create a binary mask $M_t^{(i,j)}$ for each cell, signalling whether at time t it should follow the original trajectory or denoise a new one.

$$M_t^{(i,j)} = \begin{cases} 0, & \text{if } t > t_s^{(i,j)} \\ 1, & \text{if } t \leq t_s^{(i,j)} \end{cases} \quad (4)$$

At each diffusion step t , the mask M_t dynamically determines which parts of the image should use latents from the original trajectory $X_t^{(A)}$ and which should be updated with

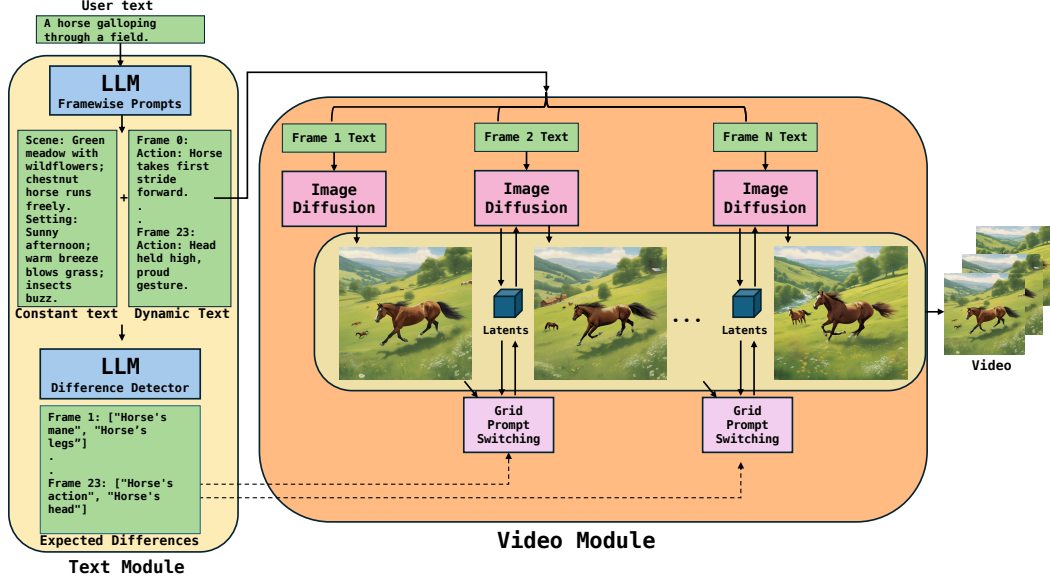


Figure 1. **EIDT-V Pipeline for Frame-Based Video Generation.** The pipeline consists of two primary modules: text and video. The text module converts the user’s input into framewise prompts and expected variations, which guide the video module in generating frames iteratively. The video module achieves controlled variance and coherence across frames by leveraging trajectory intersections. Integrating two LLM modules and the grid prompt switching enables a generic image diffusion model to synthesize coherent video sequences effectively.

latents from the new prompt trajectory $X_t^{(B)}$. The latent representation for each cell is, therefore, given by:

$$X_t^{(i,j)} = \begin{cases} X_t^{(A,i,j)}, & \text{if } M_t^{(i,j)} = 0 \\ X_t^{(B,i,j)}, & \text{if } M_t^{(i,j)} = 1 \end{cases} \quad (5)$$

This results in a composite latent X_t for the image, calculated as follows:

$$X_t = M_t \odot X_t^{(B)} + (1 - M_t) \odot X_t^{(A)} \quad (6)$$

Where $\odot$ denotes element-wise multiplication applied across each cell in the latent representation.

This technique preserves overall image coherence by seamlessly transitioning cells between prompts, effectively combining the inpainting and generation processes.

4.3. Text-Guided Grid Switching with Attention

Selecting an effective STM is essential to ensure that prompt transitions align with areas of significant variation. As shown in Fig. 2 (right panel), we automate this by comparing the initial prompt y with the target prompt y' to obtain a textual difference Δ . This difference highlights regions of high attention, where prompt y' introduces changes or motion.

Given the list of differences, expected differences in Fig. 2, we apply a CLIP-Segmentation model [29] to create attention maps over the image x_0^A . The attention maps are normalized, exponentiated, and resized to match the latent

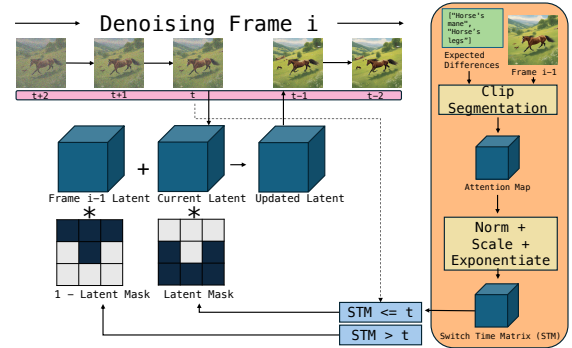


Figure 2. **Grid Prompt Switching with Text-Guided Attention.** On the right, an auxiliary block processes the previous frame and difference text using a CLIP segmentation model to generate the STM. This is converted into a binary mask (Eq. (4)) that selects, at timestep t , whether each grid cell follows the original or new prompt trajectory. In the main denoising process, the mask blends the latent representations $X_t^{(A)}$ and $X_t^{(B)}$ as per Eq. (6) to form X_t . While a coarse 3×3 grid is shown for clarity, in practice, higher-resolution masks (e.g., 128×128) are used.

size, producing the targeted STM. Cells with high attention values receive earlier switch times, allowing more variance, while low-attention areas retain stability.

As illustrated in Fig. 1, this approach connects frames in the video module, enabling controlled transitions in motion-heavy areas for smooth frame continuity.

4.4. Large Language Models

We now have a mechanism to connect frames in our video, but we require frame-wise conditional text prompts and lists of differences between them. To generate these, we use in-context learning [5, 30] with LLaMA-3 8B [7], creating two modules as shown in Fig. 1.

LLM Framewise Prompts generates prompts y_t for each frame t based on the user’s initial description. Each Prompt consists of a fixed scene descriptor and a dynamic component:

$$y_t = \text{Fixed Scene Descriptor} + \text{Dynamic Component}_t \quad (7)$$

To fit within CLIP’s [34] 77-token limit, we allocate 60 tokens for the fixed part and 15 for the dynamic component (2-token buffer), preserving prompt coherence across frames while allowing gradual transitions.

LLM Difference Detector compares two text prompts to identify differences, returning a list of distinct elements between them. We ask the model to return anything that may be in motion or varied between the two scenes described by the prompts.

Examples of outputs from both modules are in the text module in Fig. 1 and in Fig. 5.

5. Experiments

5.1. Implementation Details

To evaluate our method’s robustness, we created a set of 50 diverse video generation prompts (see Sec. 9) selected to span various visual and semantic contexts.

We implement our model with the Stable Diffusion framework, primarily SD1.5 [35] for comparability with prior work. We also tested compatibility and performance on more advanced architectures, specifically SDXL [32] and SD3 Medium [8]. All experiments ran on a single NVIDIA RTX A5000 GPU (24 GB memory), comparable to an RTX 3090, demonstrating the feasibility of our approach on consumer-grade hardware. Hyperparameter tuning details can be found in Sec. 10.

We also explored a modular alternative to cross-attention manipulation of previous works using the IP-Adapter [47], more details in Sec. 11.

5.2. Evaluation Metrics

Previous works have often relied on the CLIP score as a primary evaluation tool; however, this method only checks image-text alignment. To address this gap, we incorporated additional metrics specifically designed to assess frame-to-frame consistency. First, we employed Multi-Scale Structural Similarity (MS-SSIM) [44], quantifying the structural similarity between consecutive frames across multiple

scales. Higher MS-SSIM values indicate better preservation of structural information, which is essential for maintaining consistent content and layout between frames. Additionally, we used Learned Perceptual Image Patch Similarity (LPIPS) [48], which assesses perceptual similarity by analyzing deep features from neural networks. Lower LPIPS values correlate with better perceptual continuity across frames, contributing to smoother transitions and reducing flickering artefacts. Finally, we introduced an Optical Flow-based Temporal Consistency Loss using the Farneback method [9] to evaluate temporal motion consistency. Lower temporal consistency loss values suggest smoother, more coherent motion, an essential factor in generating realistic video sequences. These metrics provide a more comprehensive assessment of our method’s capacity to generate videos with high temporal coherence, structural integrity, and perceptual consistency across frames. For completeness, we report CLIP scores in Sec. 12, although, as expected, these show minimal variance across different models.

5.3. Results

This section presents quantitative and qualitative evaluations of our method, comparing its performance to baseline approaches under two conditions: (1) using Stable Diffusion 1.5 for fair comparisons with existing works and (2) evaluating its adaptability and scalability across different architectures. We emphasize that our comparisons are limited to training-free methods with publicly available code, as many more recent approaches require additional training or lack accessible implementations.

5.3.1 Quantitative Comparison

Our method achieved competitive performance across all three metrics compared to previous approaches using SD1.5. Specifically, our MS-SSIM score for EIDT-V SD1.5 with IP-Adapter was 0.655 ± 0.13 , closely matching the highest score of 0.672 ± 0.095 achieved by Free-Bloom [21]. Furthermore, our LPIPS score of 0.316 ± 0.089 was lower than those obtained by other SD1.5-based methods, indicating an improvement in perceptual similarity across frames. Regarding temporal consistency, our base SD1.5 model achieved a score of 0.152 ± 0.062 , suggesting that our model generated more stable motion as measured by optical flow analysis.

Our method demonstrated strong adaptability to newer frameworks when evaluating across architectures, with substantial performance improvements. For example, the SDXL [32] implementation produced an MS-SSIM score of 0.701 ± 0.089 , with LPIPS and Temporal Consistency scores of 0.28 ± 0.086 and 0.138 ± 0.054 , respectively. The SD3 Medium [8] model further enhanced performance,

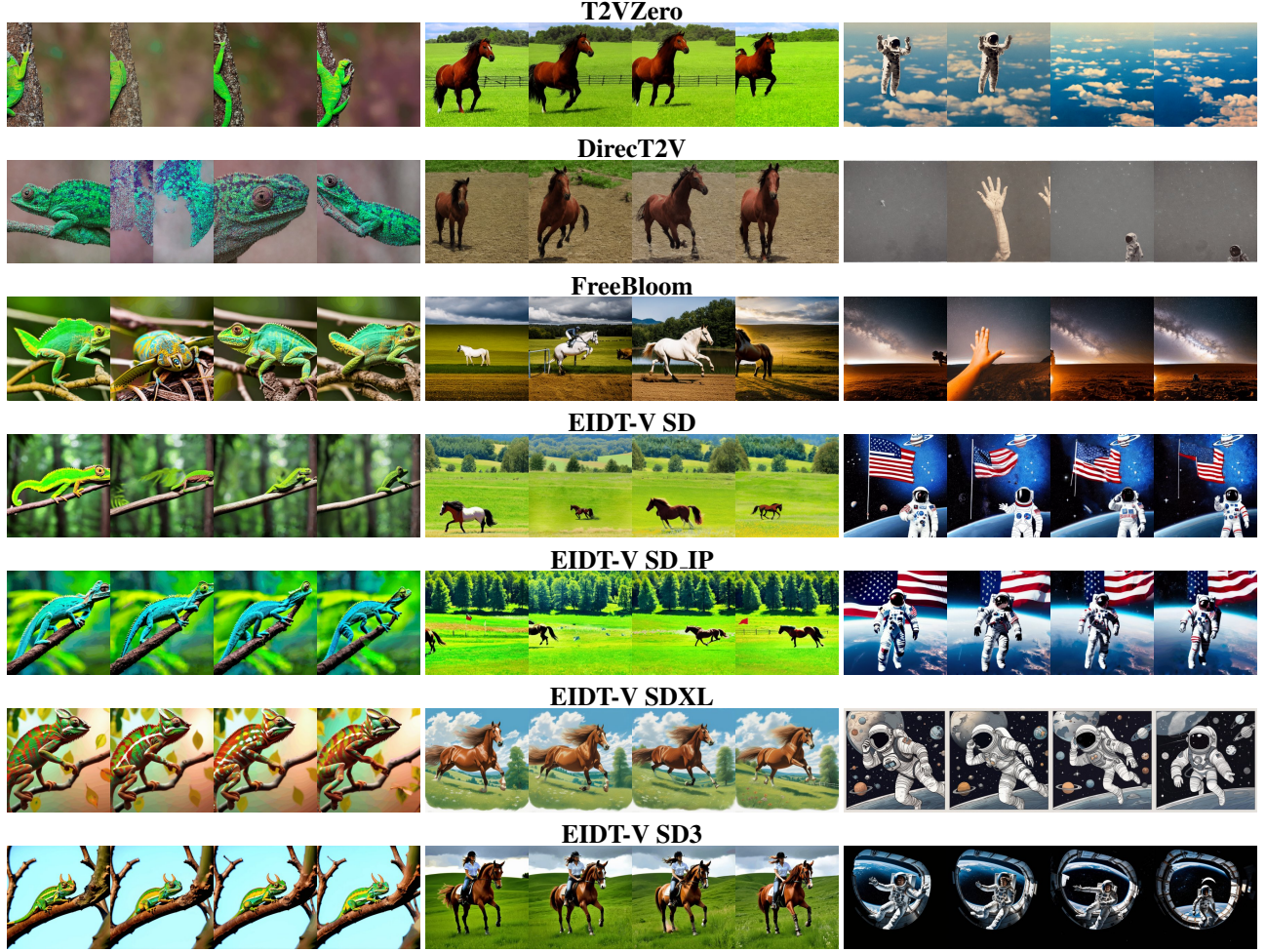


Figure 3. **Qualitative comparison of different video generation models across three prompts:** (a) “A chameleon changing colors on a branch”, (b) “A horse galloping across a field”, and (c) “An astronaut floating in space waving”. T2VZero produces coherent frames but does not fully capture the specifics of each prompt; for instance, the chameleon does not change colors, and the astronaut does not appear to be waving. DirecT2V struggles to generate coherent frames. Interestingly, both DirecT2V and FreeBloom, which are LLM-based models, capture the essence of “waving” and “space” but fail to fully integrate these concepts in each frame. They have strong semantic coherence but not temporal. Our model, however, demonstrates clear color changes in the chameleon, captures the horse’s movement (notice the legs), and shows the astronaut’s arm moving in a waving pattern while keeping the rest of the frame highly consistent.

achieving MS-SSIM, LPIPS, and Temporal Consistency scores of 0.81 ± 0.109 , 0.184 ± 0.08 , and 0.087 ± 0.042 , respectively. These results underscore the scalability of our approach, with newer architectures contributing to improved detail, reduced flickering, and greater coherence.

5.3.2 Qualitative Comparison

Qualitative comparisons with baseline methods Fig. 3 (additional in Sec. 16, Sec. 17), further illustrate our method’s improvements in temporal coherence, subtle motion accuracy and flexibility across various architectures.

When evaluated on the same SD1.5 architecture as prior methods such as DirecT2V [20], Free-Bloom [21], and

T2V-Zero [24], our approach demonstrated notable qualitative enhancements. Our method achieved smoother frame transitions and maintained consistency across sequential frames, resulting in visually coherent videos. Additionally, our approach effectively captured subtle, nuanced movements between frames, attributable to targeted attention adjustments applied across sequences.

We validated further on SDXL [32] and SD3 Medium [8] architectures, where it produced frames with improved clarity and detail, enhancing the realism of video sequences. Notably, with high fidelity, our approach captured text-prompt-specific changes, such as colour variations or subtle movements, such as hand waves. These highlight our model’s flexibility and capacity to adhere promptly across

Table 1. **Quantitative performance comparison of video generation models**, including T2V-Zero, DirecT2V, Free-Bloom, and our proposed EIDT-V method across multiple configurations. Metrics include MS-SSIM (higher indicates better structural similarity), LPIPS (lower indicates better perceptual quality), and Temporal Consistency Loss (lower indicates better temporal coherence). The table highlights the flexibility of EIDT-V across various pre-trained architectures, with the best results achieved using SD3 Medium.

Method	Pre-Trained Model	Unmodified Architecture	MS-SSIM (↑)	LPIPS (↓)	Temporal Consistency (↓)
T2V-Zero [24]	SD1.5	×	0.428 ± 0.174	0.404 ± 0.083	0.206 ± 0.066
DirecT2V [20]	SD1.5	×	0.492 ± 0.135	0.445 ± 0.089	0.185 ± 0.061
Free-Bloom [21]	SD1.5	×	0.672 ± 0.095	0.353 ± 0.082	0.159 ± 0.039
EIDT-V	SD1.5	✓	0.63 ± 0.137	0.33 ± 0.1	0.152 ± 0.062
EIDT-V	SD1.5 w/ IP-Adapter [28]	×	0.655 ± 0.13	0.316 ± 0.089	0.158 ± 0.074
EIDT-V	SDXL	✓	0.701 ± 0.089	0.28 ± 0.086	0.138 ± 0.054
EIDT-V	SD3 Medium	✓	0.81 ± 0.109	0.184 ± 0.08	0.087 ± 0.042

Table 2. **User Study Results** (mean ranking out of 4). Lower scores indicate better performance. Metrics include Temporal Coherence, Fidelity, Semantic Coherence, and Overall Score. All methods were evaluated using SD1.5. EIDT-V achieves the best overall ranking (1.9), followed closely by FreeBloom.

Method	Temporal Coherence (↓)	Fidelity (↓)	Semantic Coherence (↓)	Overall (↓)
FreeBloom	2.5	2.0	2.0	2.3
T2VZero	2.4	2.5	2.7	2.5
DirecT2V	3.2	3.5	3.1	3.3
EIDT-V	1.9	2.0	2.1	1.9

diverse video generation scenarios.

5.4. User Study

To evaluate our approach from a human perspective, we conducted a user study comparing EIDT-V SD1.5 with three baseline models: FreeBloom, T2VZero, and DirecT2V. Eight participants assessed 50 video samples each across multiple criteria, including temporal coherence, visual fidelity, and semantic alignment. Detailed study design information is available in Sec. 13.

The results of the user study, shown in Tab. 2, indicate that our model, EIDT-V SD1.5, achieved the best scores across all but semantic coherence, where it was a close second. Our model received the best mean score of 1.9 for temporal coherence, indicating smoother transitions and improved frame-to-frame consistency. Regarding fidelity, our model achieved a mean score of 2.0, which is on par with FreeBloom, with users noting the superior visual quality and fewer artifacts than other models. For semantic coherence, EIDT-V scored 2.1, just shy of FreeBloom’s 2.0, demonstrating strong alignment with the intended prompts. Our model received a mean score of 1.9 for user satisfaction, outperforming all baseline methods. Feedback shows that our model produces smoother video, focusing more on coherence, whereas FreeBloom focuses more on text alignment.

Table 3. **Ablation Study Results** for MS-SSIM, LPIPS, and Temporal Consistency Loss across different configurations: ChatGPT (CG), Our Framewise Prompts (OFP), and each with Grid Prompt Switching (GrPS). The inclusion of GrPS, particularly with OFP, demonstrates substantial improvements in frame coherence, perceptual similarity, and temporal stability.

Config.	MS-SSIM (↑)	LPIPS (↓)	Temp. Cons. (↓)
CG	0.132 ± 0.052	0.723 ± 0.043	0.389 ± 0.046
OFP	0.124 ± 0.082	0.686 ± 0.062	0.403 ± 0.078
CG + GrPS	0.588 ± 0.129	0.384 ± 0.094	0.157 ± 0.062
OFP + GrPS	0.63 ± 0.137	0.33 ± 0.1	0.152 ± 0.062

5.5. Ablation Study

To evaluate the impact of critical components in our approach, we conducted an ablation study analyzing how different configurations affect video generation quality across our metrics. Specifically, we tested configurations with and without Grid Prompt Switching (GrPS) and Our Framewise Prompts (OFP) to isolate their effects. The results of the ablation study are in Tab. 3.

The results reveal that incorporating GrPS significantly improves structural and perceptual coherence in generated videos. For instance, the MS-SSIM score increases substantially when GrPS is added, rising from 0.132 ± 0.052 for the baseline ChatGPT (CG) prompts configuration to 0.588 ± 0.129 for CG + GrPS. Similarly, LPIPS, which measures perceptual similarity, decreases from 0.723 ± 0.043 in the CG configuration to 0.384 ± 0.094 in CG + GrPS, indicating reduced perceptual artifacts across frames. Temporal Consistency Loss also improves markedly, dropping from 0.389 ± 0.046 in CG to 0.157 ± 0.062 , suggesting smoother and more coherent motion.

When OFP combines with GrPS (OFP + GrPS), the results are further enhanced, with MS-SSIM reaching 0.63 ± 0.137 , LPIPS improving to 0.33 ± 0.1 , and Temporal Consistency Loss reduced to 0.152 ± 0.062 . These values represent the best performance across all configurations, con-



Figure 4. **Ablation Qualitative Results** (see Tab. 3). Each row displays four equally spaced frames from the generated GIF. The prompt is “**A cup of coffee being poured with steam rising**”. The naive approach produces images linked only by theme. Applying OFP without GrPS offers minor improvements while incorporating GrPS (CG with GrPS) notably increases coherence. Finally, combining OFP with GrPS yields the best performance.

firming that the combination of framewise prompts and grid prompt switching provides the most consistent and visually coherent results.

The qualitative analysis shown in Fig. 4 supports these findings, illustrating that without GrPS, frame-to-frame consistency is minimal. With GrPS, frame transitions become significantly smoother, and when paired with OFP, consistency is enhanced further, resulting in highly stable and coherent video sequences. Overall, the ablation study demonstrates the critical role of GrPS and OFP in achieving high-quality, temporally consistent video generation.

6. Discussion and Future Directions

Our approach achieves notable temporal coherence and visual fidelity across various architectures. By constraining the model in novel ways, we open a new path for training-free video generation that excels in producing subtle, targeted variations to improve frame-to-frame consistency.

A key aspect of our method is using variance as a proxy for motion. Since the model inherently understands only variance, we rely on text conditioning to direct this variance toward generating sequences that appear as coherent movements. This approach assumes that the text prompts will

guide the model in applying variance in a way that visually represents a moving object. While this method proves effective in many scenarios, it has limitations. Occasionally, the model may generate frames with some visual inconsistencies, replacing expected movement with minimal changes that do not fully convey natural motion.

Artifacts are also an issue. Distortions like limb elongation appear across training-free methods, partly due to strong conditioning effects, with some base models—most notably SD3—being especially prone. In our method, significant prompt changes late in the diffusion process may prevent full refinement of fine details. We evaluated 24 variants of a horse-running sequence, adjusting hyperparameters. Although no configuration eliminates artifacts, specific settings significantly reduce these distortions.

Future research could address these limitations by incorporating methods to improve frame coherence, such as minimal training to reinforce the distinction between variance and motion. A hybrid approach that combines targeted generation with light training, similar to techniques used in AnimateDiff [13], could enable a low-cost, trained video generator with improved motion consistency. Expanding this approach within a trained, scalable environment could enhance adaptability, potentially leading to robust and resource-efficient tools for high-quality video generation.

7. Conclusion

This paper presents a novel, training-free approach to video generation. We address critical challenges in achieving temporal consistency and architectural flexibility, leveraging core diffusion mechanisms and a grid-based prompt-switching strategy to produce coherent and realistic video sequences without requiring architectural modifications.

This work’s primary contribution demonstrates that targeted variance, guided by text-based conditioning, can effectively substitute for more complex mechanisms in achieving visually coherent sequences. This approach has significant implications for enhancing the accessibility and scalability of video generation tools, narrowing the gap between high-quality output and low computational demands.

While the method has some limitations in differentiating targeted variance from actual motion, it lays a foundation for further exploration in resource-efficient video synthesis. Potential extensions include integrating lightweight training mechanisms or additional coherence-enhancing strategies to capture natural motion better and improve robustness. Overall, this study introduces a flexible, efficient, and high-fidelity video generation framework, offering valuable insights and tools for advancing the field of generative modeling.

8. Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant No. EP/T518013/1 and Grant No. EP/Y021614/1. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. The authors also acknowledge the University of Bath for providing access to the HEX high-performance computing (HPC) system, which was used for testing.

References

- [1] OpenAI . SORA Video generation models as world simulators | OpenAI, 2024. <https://openai.com/index/video-generation-models-as-world-simulators/>. 1, 2
- [2] Jie An, Songyang Zhang, Harry Yang, et al. Latent-Shift: Latent Diffusion with Temporal Shift for Efficient Text-to-Video Generation, 2023. arXiv:2304.08477 [cs]. 1, 2
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets, 2023. arXiv:2311.15127 [cs].
- [4] Andreas Blattmann, Robin Rombach, Huan Ling, et al. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023. 1, 2
- [5] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 5
- [6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 2
- [7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 Herd of Models, 2024. arXiv:2407.21783 [cs]. 5
- [8] Patrick Esser, Sumith Kulal, Andreas Blattmann, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1, 2, 5, 6
- [9] Gunnar Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer. 5
- [10] Flux. Flux, 2024. <https://blackforestlabs.ai/announcing-flux-1-l-pro-and-the-bfl-api/>. 1
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. *NeurIPS*, 27, 2014. 2
- [12] Xianfan Gu, Chuan Wen, Weirui Ye, et al. Seer: Language instructed video prediction with latent diffusion models. In *ICLR*, 2024. 1
- [13] Yuwei Guo, Ceyuan Yang, Anyi Rao, et al. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning, 2024. arXiv:2307.04725. 8
- [14] Agrim Gupta, Lijun Yu, Kihyuk Sohn, et al. Photorealistic video generation with diffusion models. In *ECCV*, pages 393–411. Springer, 2024. 1, 2
- [15] Yingqing He, Tianyu Yang, Yong Zhang, et al. Latent Video Diffusion Models for High-Fidelity Long Video Generation, 2023. arXiv:2211.13221 [cs]. 1, 2
- [16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2022. 3
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2
- [18] Jonathan Ho, William Chan, Chitwan Saharia, et al. Image Video: High Definition Video Generation with Diffusion Models, 2022. arXiv:2210.02303 [cs]. 1, 2
- [19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, et al. Video diffusion models. In *NeurIPS*, pages 8633–8646. Curran Associates, Inc., 2022. 1, 2
- [20] Susung Hong, Junyoung Seo, Heeseong Shin, et al. Direct2V: Large Language Models are Frame-Level Directors for Zero-Shot Text-to-Video Generation, 2024. arXiv:2305.14330 [cs]. 2, 6, 7
- [21] Hanzhuo Huang, Yufan Feng, Cheng Shi, et al. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *NeurIPS*, 36:26135–26158, 2023. 2, 5, 6, 7
- [22] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b, 2023. 5
- [23] Tero Karras, Miika Aittala, Timo Aila, et al. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 35:26565–26577, 2022. 1, 3
- [24] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, et al. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV*, pages 15954–15964, 2023. 2, 6, 7
- [25] Kling. KLING AI: Next-Generation AI Creative Studio, 2024. <https://klingai.com/>. 2
- [26] Jiawei Liu, Weining Wang, Wei Liu, et al. ED-T2V: An Efficient Training Framework for Diffusion-based Text-to-Video Generation. In *IJCNN*, pages 1–8, 2023. ISSN: 2161-4407. 1, 2
- [27] Shaoteng Liu, Yuechen Zhang, Wenbo Li, et al. Video-p2p: Video editing with cross-attention control. In *CVPR*, pages 8599–8608, 2024. 2
- [28] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 2, 7
- [29] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, pages 7086–7096, 2022. 4
- [30] Sewon Min, Xinxin Lyu, Ari Holtzman, et al. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022. 5
- [31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2
- [32] Dustin Podell, Zion English, Kyle Lacey, et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 1, 2, 5, 6
- [33] Adam Polyak, Amit Zohar, Andrew Brown, et al. Movie Gen: A Cast of Media Foundation Models, 2024. arXiv:2410.13720. 1, 2
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 1, 2, 5
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2
- [37] Runway. Runway | Tools for human imagination., 2024. <https://runwayml.com/>. 2
- [38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. pmlr, 2015. 2
- [39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2
- [40] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *NeurIPS*. Curran Associates, Inc., 2019. 3
- [41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, et al. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1, 2, 3
- [42] Yang Song, Prafulla Dhariwal, Mark Chen, et al. Consistency models. In *ICML*, pages 32211–32252. PMLR, 2023. 2
- [43] Patrick von Platen, Suraj Patil, Anton Lozhkov, et al. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. 5
- [44] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402, Pacific Grove, CA, USA, 2003. IEEE. 5
- [45] An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 5
- [46] Siyuan Yang, Lu Zhang, Yu Liu, et al. Video diffusion models with local-global context guidance. In *IJCAI*, pages 1640–1648, 2023. 2
- [47] Hu Ye, Jun Zhang, Sibio Liu, et al. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, 2023. arXiv:2308.06721 [cs]. 5, 3, 4
- [48] Richard Zhang, Phillip Isola, Alexei A Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5
- [49] Daquan Zhou, Weimin Wang, Hanshu Yan, et al. MagicVideo: Efficient Video Generation With Latent Diffusion Models, 2023. arXiv:2211.11018 [cs]. 2