Encapsulated Composition of Text-to-Image and Text-to-Video Models for High-Quality Video Synthesis

Tongtong Su[1,2], Chengyu Wang[2]*, Bingyan Liu[3,2], Jun Huang[2], Dongming Lu[1]*
[1] Zhejiang University, [2] Alibaba Cloud Computing,
[3] South China University of Technology
{sutongtong,ldm}@zju.edu.cn,
{chengyu.wcy,huangjun.hj}@alibaba-inc.com, eeliubingyan@mail.scut.edu.cn

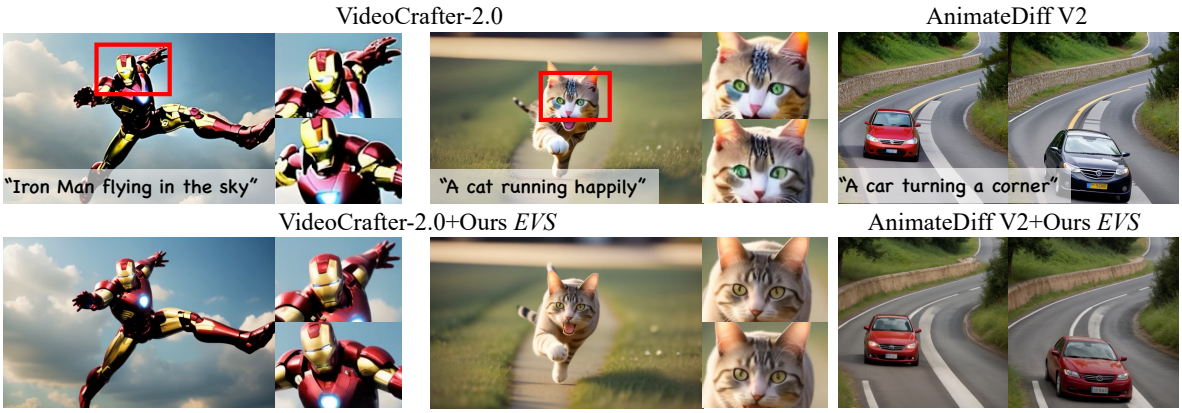

Figure 1. Comparison of videos generated by T2V models w/ and w/o *EVS*. Our method significantly improves imaging quality compared to videos solely generated by VideoCrafter-2.0 (see the Iron Man head and cat eyes), and frame consistency compared to AnimateDiff-V2 (see the stability of car color across frames).

Abstract

In recent years, large text-to-video (T2V) synthesis models have garnered considerable attention for their abilities to generate videos from textual descriptions. However, achieving both high imaging quality and effective motion representation remains a significant challenge for these T2V models. Existing approaches often adapt pre-trained text-to-image (T2I) models to refine video frames, leading to issues such as flickering and artifacts due to inconsistencies across frames. In this paper, we introduce EVS*, a training-free Encapsulated Video Synthesizer that composes T2I and T2V models to enhance both visual fidelity and motion smoothness of generated videos. Our approach utilizes a well-trained diffusion-based T2I model to refine low-quality video frames by treating them as out-of-distribution samples, effectively optimizing them with noising and denoising steps. Meanwhile, we employ T2V backbones to ensure consistent motion dynamics. By encapsulating the T2V temporal-only prior into the T2I generation process,* EVS *successfully leverages the strengths of both types of models, resulting in videos of improved imaging and motion quality. Experimental results validate the effectiveness of our approach compared to previous approaches. Our composition process also leads to a significant improvement of 1.6x-4.5x speedup in inference time.* [1]

1. Introduction

Recently, large-scale text-to-video (T2V) models [27, 28, 32, 33] have gained significant attention due to their ability to generate realistic videos from textual descriptions. These models leverage vast datasets of text-video pairs, allowing them to learn complex relationships between textual inputs

*Co-corresponding authors.

[1]Source codes: https://github.com/alibaba/EasyNLP/tree/master/diffusion/EVS

and visual outputs. Currently, the prominent T2V generation research in the open-source community can largely be categorized into two main approaches. The first approach focuses on training a general T2V diffusion model. This is achieved either by initializing certain modules with pre-trained text-to-image (T2I) models and introduces additional blocks to concentrate on learning the temporal dynamics of videos [3, 8, 40], or by training from scratch jointly on images and videos [14, 45]. In contrast, alternative methods employ T2I models for video synthesis without extensive re-training [13, 30, 47], which inflate T2I along the temporal axis (i.e., replacing self-attention layers in U-Net with cross-frame attention layers) and successfully maintain the imaging quality of generated videos at the T2I levels. Despite these advancements, as shown in Figure 2, current popular T2V models often struggle to simultaneously ensure high imaging quality and motion quality [12], which are essential challenges that need to be addressed to improve overall performance of video synthesis.

A straightforward approach to addressing the challenges is to improve the imaging quality of videos generated in the vanilla T2V pipeline by combining T2I models. Yet, T2I models can only be applied in a frame-independent manner, which may lead to flickers between frames. Thus, explicit consistency constraints are incorporated into the video synthesis pipeline. For instance, Rerender-A-Video [43] utilizes optical flow to iteratively warp latent features from the previous frame, aligning them with the current frame. TokenFlow [7] explicitly propagates diffusion features and computes inter-frame feature correspondences using Nearest-Neighbor Field (NNF) search. However, these methods are designed for real-world videos and rely on precise optical flow or NNF estimations on input videos with high motion consistency. When these techniques are directly applied to model-generated videos that may exhibit apparent inconsistencies, inaccuracies in estimation can exacerbate these inconsistencies and introduce artifacts.

We propose *EVS*, a training-free Encapsulated Video Synthesizer composing of T2I and T2V models to produce videos with significantly balanced imaging and motion qualities, together with large inference speedup compared with vanilla alternating two models. Specifically, we treat low-quality image frames as out-of-distribution samples [24, 26, 29, 46] for the T2I model and devise proper noising and denoising steps to pull them back to the high-quality imaging distribution. As for the underlying T2V model, we employ publicly available backbones that are capable of producing highly consistent and stable videos to enhance the motion smoothness. In the *EVS* framework, we encapsulate this T2V temporal-only prior into the T2I generation process, mitigating the adverse effects of poor T2V imaging quality. This can be achieved through Selective Feature Injection (SFI), which incorporates inversion

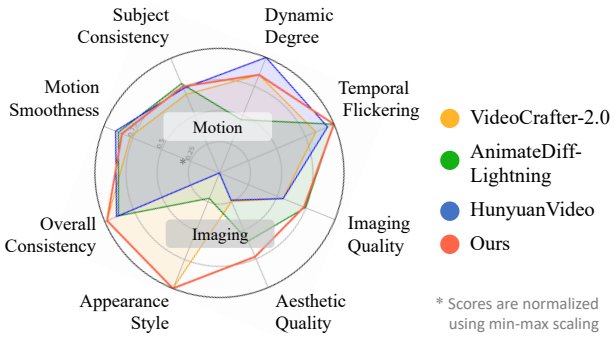

Figure 2. Evaluation results of T2I/T2V models on VBench [12]. Results show that motion and imaging qualities are hard to balance in one T2V model. With encapsulated composition of T2I and T2V models, our method effectively combines their advantages.

features representing spatial details, while allowing the remaining features to be refined by the temporal prior. Experiments on the authoritative benchmark VBench [12] demonstrate that *EVS* integrates the advantages of both two types of models, which outperforms baselines and achieves 1.6x-4.5x speedup in inference time. The results are also summarized in Figure 2. In summary, the key contributions of our paper are as follows:

- We introduce *EVS*, a training-free framework which enhances the imaging and motion qualities of synthesized videos with versatile T2I and T2V diffusion models.
- We propose a novel encapsulated injection of T2V module into T2I diffusion processes to achieve complementary advantages of T2V and T2I models.
- Experiments show that *EVS* effectively improves the imaging and motion qualities of synthesized videos, and achieves 1.6x-4.5x speedup in inference time.

2. Related Works

2.1. Video Diffusion Models

Current diffusion-based T2V methods [3, 6, 8, 10, 13, 14, 18, 30, 39, 40, 45] can be categorized into two groups. The first category comprises zero-shot methods that require only a pre-trained T2I model [13, 30, 36, 47]. During inference, these methods utilize cross-frame attention to ensure temporal consistency, which are limited to generating videos with simple dynamics and are unable to handle more complex motions. To address this limitation, some studies fine-tune the T2I model on a single video, enabling the generation of videos with similar motion patterns. However, this approach tends to overfit the specifics of the single video and lacks generalization to other motions [5, 35, 42]. The second category involves training a general T2V diffusion model on large-scale video data [3, 8, 10, 14, 45]. Due to the scarcity of high-quality video-text data, these methods either initialize the T2V spatial module with a pre-trained T2I model or jointly train using both image and video data.

Current T2V benchmarks, such as VBench [12], include evaluation metrics such as frame-wise imaging quality, temporal consistency, and dynamic degree. According to the findings [12, 23], there is currently no model that excels across all dimensions. VideoCrafter-2.0 [3], an open-source T2V model that achieves a balanced and excellent performance across all dimensions, consistently ranking above average in each. It explores various strategies to enhance the learning of temporal modules and mitigate image degradation. Nonetheless, despite these efforts, the resulting imaging quality still falls short when compared to that of dedicated T2I models. Inspired by recent advances in frame-wise video editing [7, 30], one might contemplate improving the imaging quality of individual frames using sophisticated T2I models. However, processing each frame independently may exacerbate inconsistencies among frames, leading to noticeable flickering.

2.2. Improving Temporal Consistency of Videos

Previous works [2, 7, 13, 30, 43] have considered constraining temporal consistency at both global and local levels. At the global level, they replace self-attention in U-Net with cross-frame attention to regularize the roughly unified appearance; however, this approach is insufficient for ensuring local detail consistency. Rerender-A-Video [43] and FRESCO [44] employs optical flow to warp and fuse the latent features, while TokenFlow [7] utilizes NNF to compute inter-frame feature correspondences, propagating reference frame features to others. Their explicit enforcement effectively constrains consistency at local level. Nonetheless, achieving explicit correspondence requires that input videos exhibit highly consistent and simple motion [11, 17], with minimal amplitude changes between frames.

With the rapid development of T2V models, there has been increasing interest in exploring the integration of T2I and T2V models [22]. BIVDiff [34] firstly employs a T2I model to process individual video frames, followed by a T2V model for temporal smoothing. Consequently, the overall imaging quality aligns with that of the T2V model. VideoElevator [48] adopts T2I and T2V denoising steps to enhance temporal consistency and imaging quality simultaneously. They break down each sampling step into T2I and T2V from start to finish. Introducing T2I too early can hinder motion understanding, leading to motionless frames. AnyV2V [15] processes the first frame using T2I model and leverage image-to-video (I2V) [46] model to handle the entire video. The final results is limited by current I2V ability, which struggles to process videos with complex motion.

3. *EVS*: The Proposed Method

In this section, we leverage both T2I and T2V models to enhance imaging quality and motion smoothness of generated videos, without training another refinement model. We first introduces the generation process of T2I and T2V models. After that, we describe two basic composition approaches of T2I and T2V models, discussing and presenting their limitations. We then delve into the specifics of our *encapsulated composition*, addressing how it overcomes these drawbacks.

3.1. Preliminaries and Basic Notations

T2I. T2I models, exemplified by the Latent Diffusion Model (LDM) [31], generate images based on textual descriptions. LDM comprises a pre-trained autoencoder and a U-Net architecture. The encoder compresses an image x into latent space, yielding $z_0 = \mathcal{E}(x)$, while the decoder $\mathcal{D}$ reconstructs z_0 back to the pixel space. LDM is trained in latent space by estimating various levels of noise added to z_0, with the strength parameterized by $\{\bar{\alpha}_t\}_{t=1}^{T}$:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $t = 0, \cdots, T-1$ is the timestep and ϵ is the Gaussian random noise. At the inference stage, we sample z_{t-1} based on z_t with DDIM sampling [37]:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{z_{t\to 0}}_{\text{predicted } z_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(z_t, t, c)}_{\text{direction pointing to } z_{t-1}}, \tag{2}$$

where $z_{t\to 0}$ is the predicted clean latent at timestep t:

$$z_{t\to 0} = (z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, c))/\sqrt{\bar{\alpha}_t}, \tag{3}$$

ϵ_θ denotes the noise prediction diffusion model, and c represents the text embedding. For each T2I denoising step, we obtain both z_{t-1} via Eq. (2) and $z_{t\to 0}$ via Eq. (3).

When we aim to improve the imaging quality using T2I models, SDEdit noising-denoising procedure is often leveraged [24, 26]. Specifically, a noising timestep t_I and an ending denoising timestep t'_I are determined. Noise is applied using Eq. (1) to obain z_{t_I}, then denoising steps are performed to obtain $z_{t'_\text{I}-1}, z_{t'_\text{I}\to 0}$ with enhanced imaging quality. In the following sections, we focus on the predicted $z_{t'_\text{I}\to 0}$ (denote as z_0^I), as it serves as a crucial link between T2I and T2V. In short, we re-write the process as:

$$z_0^\text{I} = \text{T2I}^\uparrow(z_0, t_\text{I}, t'_\text{I}). \tag{4}$$

T2V. Recent T2V[3, 19] training strategy and inference process are consistent with those of T2I models. Additionally, the dimensionality of latent space is expanded along the temporal axis. VideoCrafter-2.0 [3] and AnimateDiff [8] employ a frame-wise autoencoder to process video frames, ensuring that the clean distribution of T2V is analogous to that of T2I. Similar to Eq. (4), we assign a noising timestep t_V and an ending timestep t'_V to enhance the temporal consistency across the frames from z_0 to z_V:

$$z_0^\text{V} = \text{T2V}^\uparrow(z_0, t_\text{V}, t'_\text{V}). \tag{5}$$

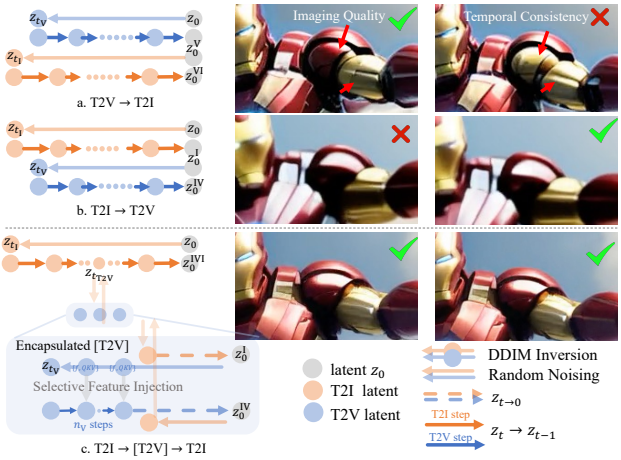

Figure 3. **a/b:** Two basic compositions of T2I and T2V. **c:** T2I denoising process encapsulated with the [T2V] block.

Note that noise schedules for T2I and T2V are different. Consequently, except for z_T (both Gaussian noises) and z_0 (T2I and T2V models share a common autoencoder), the intermediate timestep latents z_t are drawn from different distributions for T2I and T2V models. Therefore, a latent z_t from one model cannot be directly fed into the other model.

DDIM Inversion. SDEdit [24] is typically employed for conditional image generation, where the condition is implicitly remained under noisy latents. Therefore, finding a balance between maintaining condition fidelity and utilizing the diffusion prior presents a significant challenge. DDIM inversion [37] deterministically encodes the latent z_0 into noisy latent, which can be used to reconstruct z_0 through DDIM sampling. To tackle the accumulated error [25] when classifier-free guidance is applied, [4, 9, 22] collect convolutional features f_{inv} and attention features $Q_{\text{inv}}, K_{\text{inv}}, V_{\text{inv}}$ during the DDIM inversion process. These features are selectively injected to replace original features during denoising process in predefined U-Net layers.

3.2. Compositions of T2I and T2V Denoising

In this section, we introduce how to derive the encapsulated composition of T2I and T2V models to generate videos with higher imaging qualities and temporal consistency.

Two Basic Compositions. Previous works to compose T2I and T2V models can be summarized into two basic ones. Given a video latent z_0 with unsatisfied imaging quality and temporal smoothness, the first approach employs the T2V model to produce a smooth video, followed by a few T2I denoising steps to improve the frame imaging quality (see Figure 3(a)), similar to previous research on T2I-based video processing [7, 43]. The noising-denoising process is defined as:

$$z_0^{\text{V}} = \text{T2V}^{\uparrow}(z_0, t_{\text{V}}, 0), \quad z_0^{\text{VI}} = \text{T2I}^{\uparrow}(z_0^{\text{V}}, t_{\text{I}}, 0). \tag{6}$$

In this scenario, it is trivial to see that frame-wise T2I can re-introduce inconsistencies across frames. The second approach involves starting with the T2I noising and denoising steps, followed by T2V (see Figure 3(b)) [34]:

$$z_0^{\text{I}} = \text{T2I}^{\uparrow}(z_0, t_{\text{I}}, 0), \quad z_0^{\text{IV}} = \text{T2V}^{\uparrow}(z_0^{\text{I}}, t_{\text{V}}, 0). \tag{7}$$

However, this type of method may lead to a degradation in imaging quality of video frames after refinement via T2V, as the final imaging quality still depends on the T2V model. The challenge arises from the implicit condition of noisy latents in SDEdit, where it is difficult to separate spatial and temporal components.

We approach this problem from two perspectives. Considering the two basic compositions, the quality of the final video is limited by the model applied at later. Therefore, a natural idea emerges: once we obtain z_0^{VI} or z_0^{IV}, we can feed them into another round of noising and denoising. This iterative process can introduce a significant amount of redundant steps. To address this, our first strategy is *Encapsulated Composition*, which efficiently alternates between T2I and T2V stages during the denoising process, thereby eliminating redundancy. To further minimize the impact of the T2V model's disadvantages on imaging quality, we exclusively leverage its temporal prior. This introduces our second strategy of *leveraging T2V temporal-only prior with Selective Feature Injection* (SFI), where we selectively inject features from DDIM inversion into the denoising process to preserve imaging information from previous T2I steps.

Our Encapsulated Composition. The key challenge lies in bridging the gap between two latent distributions of T2I and T2V at arbitrary timesteps. Directly utilizing z_t from one model (either T2I or T2V) to another is not feasible, particularly when two models employ different sampling methods. Instead of completely denoising the latents using one model and subsequently transitioning to another model (refer to Eq. (6) and Eq. (7)), we propose to leverage the intermediate predicted clean latents to efficiently align the two latent distributions. Specifically, we start with the T2I noising-denoising process:

$$z_0^{\text{I}} = \text{T2I}^{\uparrow}(z_0, t_{\text{I}}, t_{\text{T2V}}). \tag{8}$$

Different from Eq. (7), we do not completely denoise until reaching timestep 0. At timestep t_{T2V}, the predicted clean latent z_0^{I} can serve as a shortcut to be connected with T2V, which aligns well with the distribution of clean latent representations in both T2I and T2V models. This allows us to effectively input it into the T2V noising-denoising process:

$$z_0^{\text{IV}} = \text{T2V}^{\uparrow}(z_0^{\text{I}}, t_{\text{V}}, t_{\text{V}} - n_{\text{V}}), \tag{9}$$

where n_{V} represents how many times the T2V step is executed. This process effectively stabilizes inconsistent video latents z_0^{I} from frame-wise T2I denoising into consistent

z_0^{IV}. The t_{V} value is independent with T2I shortcut timestep t_{T2V}, and can therefore be regarded as an *encapsulated block* outside the T2I process. For varying levels of inconsistency in video frames (such as global inconsistency of changing color, or local inconsistency of changing details), different noisy timestep t_{V} will be required. By decoupling this T2V stabilization block from T2I process, we can introduce a significantly larger noise at a later stage for T2I denoising. This approach allows us to explore optimal methods for balancing stabilization with minimal degradation in imaging quality. Similarly, z_0^{IV} can serve as a shortcut to be connected back to the T2I denoising process for the remaining T2I steps:

$$z_0^{\text{IVI}} = \text{T2I}^{\uparrow}(z_0^{\text{IV}}, t_{\text{T2V}}, 0), \tag{10}$$

where z_0^{IVI} is the finally enhanced video latent. Instead of implementing T2V stabilization during every T2I denoising step, we investigate the potential efficiency gained by selecting only one time of applying the block. This aims to conserve computational resources while maintaining effectiveness. Once the T2V block is introduced, the previously improved imaging quality from T2I may deteriorate again, necessitating another round of T2I and T2V processing. This indicates that the final stability of the video frames is significantly influenced by the timing of the last use of the T2V model stabilization, leading to a pipeline represented as T2I+[T2V]+T2I, where the *encapsulated block* [T2V] needs to be applied only once.

Leveraging Temporal-Only Prior of T2V. Directly applying T2V-based SDEdit [24] simultaneously introduces imaging and temporal prior. Applying DDIM inversion based reconstruction [37], on the other hand, will not introduce any prior. In our work, we wish to maintain imaging quality from T2I steps, and only leverage the temporal prior of T2V. In practice, it is feasible to attain a balanced noising strength with SDEdit. We aim to further minimize the imaging degradation caused by T2V.

We start by considering two extreme cases: (1) reconstructing the original video to the maximum extent, using DDIM inversion and full features injection; and (2) leveraging the T2V prior to the maximum extent using DDIM inversion. For a well-trained T2I model, DDIM inversion can reconstruct arbitrary image with enough steps [37], even for out-of-distribution (OOD) images (as shown in Figure 4 first row, $T = 50$). With limited steps ($T = 5$), even the reconstruction will fail, it allows for the T2I prior to self-rectify low-quality images. Based on this observation, we regard this partial reconstruction as an opportunity for the T2V model to utilize its prior knowledge. As shown in Figure 4 second row, using the T2V model with fewer steps ($T = 8$) can temporally smooth the OOD videos generated from frame-wise T2I-based approaches.

With the above two extreme cases, we can start our probing analysis of U-Net layers to find a satisfied partial reconstruction point, which only leverages the temporal prior of the T2V model. Recent analysis of T2I U-Net decoder layers has shown that, (1) self-attention Softmax$(QK^{\top})$ at deeper layers represent structural information of original images [20, 21]. (2) V is closely related to imaging information, i.e, styles and colors, while shallower layers of the U-Net architecture are more related to detailed textures [4, 38]. For our task, we aim at maintaining the imaging information; therefore, we inject $K_{\text{inv}}, V_{\text{inv}}$ at predefined layers, while *slightly* perturbing self-attention maps through blending Q_{inv} with Q during denoising:

$$\phi_{\text{out}} = \text{Attn}(\gamma \cdot Q_{\text{inv}} + (1-\gamma) \cdot Q, K_{\text{inv}}, V_{\text{inv}}), \tag{11}$$

where $Q_{\text{inv}}, K_{\text{inv}}, V_{\text{inv}}$ are collected attention features during DDIM inversion process, Q is original query smoothed by T2V during the denoising steps, and γ is the blending rate. When $\gamma = 1$ and injection is applied in all U-Net layers, the process should result in the upper limit of reconstruction. Then we gradually decrease γ and the number of injected layers, progressing from the shallow to the deep layers of the U-Net. This process involves incrementally introducing temporal priors while ensuring the preservation of image quality. Our *Selective Feature Injection* (SFI) strategy simplifies the identification of the balancing point compared to SDEdit, since spatial and temporal information are decoupled in an explainable manner, rather than being implicitly mixed within noisy latent space as in SDEdit.

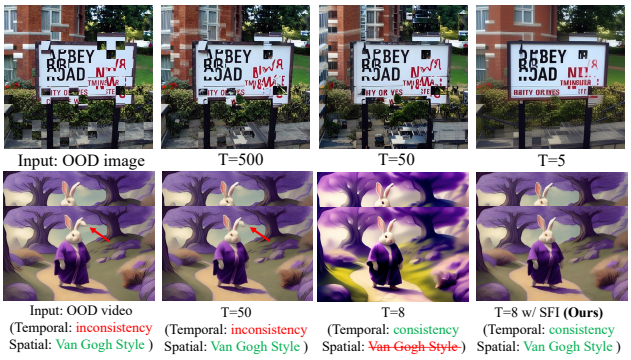

Figure 4. **Row 1:** T2I-based DDIM inversion with different steps. With limited ($T = 5$), T2I prior helps self-rectify images. **Row 2:** T2V-based DDIM inversion with limited steps ($T = 8$) simultaneously introduces temporal and spatial prior, but the latter fails to fully capture the original watercolor style. Our Selective Feature Injection (SFI) strategy can exclusively introduce the temporal prior without imaging style degradation.

3.3. Summary of Our *EVS* Method

To summarize, the optimal pipeline in *EVS* can be characterized as follows: the primary denoising step involves the T2I process for imaging quality enhancement, with inter-

Algorithm 1 The *EVS* Algorithm

Input: z_0: Original video latent; c: Text embedding; t_{T2V}: [T2V] employ timestep during T2I process; $t_{\text{I}}, t_{\text{V}}$: Noising timestep of T2I, T2V; n_{V}: [T2V] block denoising steps;
Output: z_0^{IVI}: Improved video latent;

$z_{t_{\text{I}}} = \sqrt{\bar{\alpha}_{t_{\text{I}}}} z_0 + \sqrt{1-\bar{\alpha}_{t_{\text{I}}}}\epsilon$;
for $t = t_{\text{I}}, t_{\text{I}}-1, \ldots, 1$ **do**
  $z_{t\to 0}, z_{t-1} \leftarrow z_t, \epsilon_\theta^I(z_t, t, c)$;
  if $t = t_{\text{T2V}}$ **then**
    $z_0^{\text{I}} := z_{t_{\text{T2V}}\to 0}$; # bridge to [T2V]
    $z_{t_{\text{V}}}, \{f, QKV\}_{\text{inv}} = \text{DDIM-inv}(z_0^{\text{I}}, t_{\text{V}})$;
    for $t' = t_{\text{V}}, t_{\text{V}}-1, \ldots, t_{\text{V}}-n_{\text{V}}+1$ **do**
      $z_{t'\to 0}, z_{t'-1} \leftarrow z_{t'}, \epsilon_\theta^V(z_{t'}, t, c, \{f, QKV\}_{\text{inv}})$;
    end for
    $z_0^{\text{IV}} := z_{(t_{\text{V}}-n_{\text{V}})\to 0}$; # bridge back to T2I
    $z_{t_{\text{T2V}}} = \sqrt{\bar{\alpha}_{t_{\text{T2V}}}} z_0^{\text{IV}} + \sqrt{1-\bar{\alpha}_{t_{\text{T2V}}}}\epsilon$;
  end if
end for
Return $z_0^{\text{IVI}} := z_0$.

mediate timestep for the application of the [T2V] *encapsulated block* for temporal motion consistency enhancement. DDIM inversion with *Selective Feature Injection* is *optional* only for challenging cases where the imaging information required and obtained through the T2I model falls entirely outside the T2V domain, making it challenging for SDEdit to obtain a balanced point for both factors. Finally, We present the *EVS* algorithm pseudo-code in Algorithm 1.

4. Experiments

4.1. Experimental Settings

Dataset. We utilize videos from VBench [12], generated by VideoCrafter-2.0 [3] **(VC2)** and AnimateDiff-V2 [8] **(AD2)** as our original video dataset. We use the *Overall Consistency* subset[2], due to its inclusion of videos characterized by complex movements, and intricate details in the objects. This subset is particularly useful for evaluating temporal motion consistency and imaging quality. The global consistency of videos produced by VideoCrafter-2.0 is well-preserved, particularly regarding the shape and color of subjects. However, the primary issue arises from local inconsistencies, which manifest as flickering and artifacts in details (illustrated in Figure 5 Left, where details of Iron Man's legs exhibit noticeable changes across frames). In contrast, AnimateDiff-V2 videos present a more challenging scenario, characterized by global inconsistency (as illustrated in Figure 5 Right, where the color of the moving car transitions from white to red). To clearly illustrate the improved imaging quality achieved by T2I models, we resize the videos to twice their original dimensions (**VC2**: 320×512, **AD2**: 512×512) for input to all baselines.

Baselines. We compare our approach with two streams of baselines. Rerender-A-Video [43], FRESCO [44] and TokenFlow [7] utilize T2I for frame-wise processing. TokenFlow also necessitates a precise DDIM inversion and the storage of intermediate features to compute NNF. Since our task involves substantial yet specific adjustments, such as enhancing image quality and adding details, we apply the same inversion strength $s_{\text{I}} = t_{\text{I}}/T_{\text{I}} = 0.4$ for above methods, and our method to ensure a fair comparison. BIVDiff employs a mixed inversion of T2I and T2V inversion to align frame-wise generated latents of T2I with the T2V denoising process, leveraging the T2V model to achieve temporal smoothness of videos. It requires an inversion strength of $s_{\text{I}} = 1.0$ for mixed inversion in the initial random noise at timestep $t_{\text{I}} = T_{\text{I}}$. AnyV2V [15] applys T2I model for first frame processing, followed by the application of the I2V inversion (I2Vgen [46]) using the same strength as ours. All methods utilize *epiCRealism*[3] as T2I, which is specialized in high-quality image synthesis, with total timesteps $T_{\text{I}} = 50$. For BIVDiff and our method, we adopt AnimateDiff-Lightning [19] with default $T_{\text{V}} = 8$ as the foundational model for the T2V model due to its superior temporal motion consistency ranking on VBench.

Evaluation Benchmarks. For imaging quality assessment, we adopt **DOVER** [41] and the Aesthetic Predictor **(AP)** V2.5[4]. In terms of motion consistency, we utilize two metrics from VBench [12]: Motion Smoothness **(MS)** and Subject Consistency **(SC)**. The former metric focuses on local consistency by interpolating frames [16] $t-1$ and $t+1$ and computing the error with frame t. The latter metric emphasizes global consistency by leveraging DINO [1] feature similarity across all frames. We further calculate an **Overall** score by averaging the normalized values of the four scores mentioned above. The normalization range is derived from the VBench LeaderBoard[5]. Additionally, we emphasize time efficiency as we introduce a new composition technique, without time-consuming explicit consistency computing and redundant denoising steps.

4.2. Comparisons Against Baselines

Table 1 shows that our *EVS* enhances temporal motion consistency and imaging quality of videos generated by VC2 and AD2, achieving an overall better video quality. As illustrated in Figure 5 Left, the inaccurate estimation of optical flow in VC2 videos results in significant repainting in Rerender-A-Video. Similarly, the inaccurate estimation of NNF in TokenFlow leads to mismatches between adjacent patches, ultimately resulting in blurring (see mechanical details of Iron Man's legs). In the case of AD2, the inconsistency also exacerbates the issues stemming from inaccuracies in optical flow or NNF estimation. As demonstrated in Figure 5 Right, while Rerender-A-Video and TokenFlow may achieve some level of global color consistency, the in-

[2]For each of the 93 prompts provided in the benchmark, we select videos (with id=0) for both VC2 and AD2, totaling 186 videos.

[3]https://huggingface.co/emilianJR/epiCRealism
[4]https://github.com/discus0434/aesthetic-predictor-v2-5
[5]https://huggingface.co/spaces/Vchitect/VBench_Leaderboard

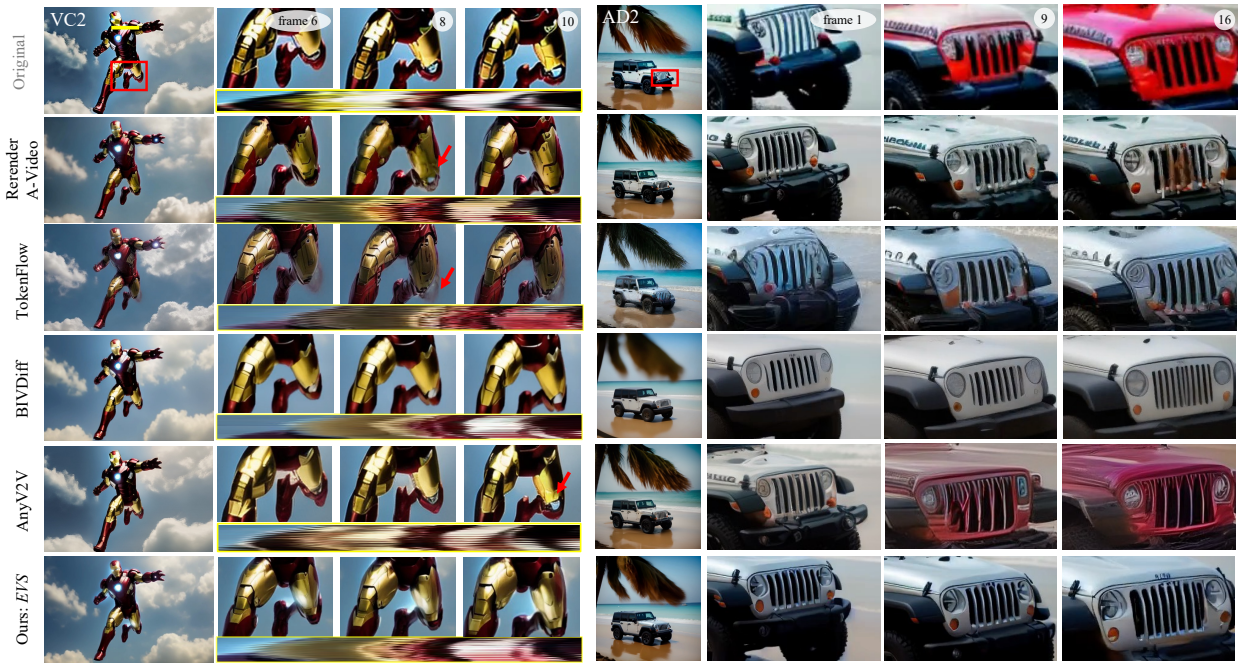

Figure 5. Comparison with baselines. Our method enhances both imaging quality and temporal consistency. Rerender-A-Video and TokenFlow introduce inconsistencies, evidenced by blur and artifacts (see red frames and flickers in yellow line pixels across frames). T2V temporal smoothing of BIVDiff reverts imaging quality to T2V level, lacking realistic details. AnyV2V is unable to effectively propagate the first frame's enhancement to subsequent frames, resulting in persistent artifacts. We refer the reader to our supplementary material for comprehensive video comparisons with baselines.

	Consistency (↑)		Imaging (↑)		Overall (↑)	Time (↓)		Consistency (↑)		Imaging (↑)		Overall (↑)	Time (↓)
	MS	SC	DOVER	AP				MS	SC	DOVER	AP		
VC2-ori	0.9829	0.9738	55.17	4.50	0.6435	-	AD2-ori	0.9769	0.9484	73.32	4.69	0.5806	-
Rerender	0.9820	0.9745	**76.39**	5.29	0.8385	532.15	Rerender	0.9750	0.9496	**85.05**	**5.73**	0.8197	1174.08
FRESCO	0.9745	0.9706	73.59	5.44	0.7917	206.89	FRESCO	0.9667	0.9396	82.28	5.08	0.6968	494.37
TokenFlow	0.9696	0.9786	64.47	4.39	0.6759	546.94	TokenFlow	0.9630	**0.9541**	72.77	4.65	0.5369	1091.74
BIVDiff	**0.9885**	0.9800	64.62	5.05	0.7707	352.70	BIVDiff	0.9758	0.9513	74.39	4.91	0.6107	1073.66
AnyV2V	0.9775	0.9540	73.10	5.11	0.7483	377.21	AnyV2V	0.9751	0.9270	76.93	4.89	0.5998	672.71
Ours	0.9881	**0.9808**	73.20	**5.46**	**0.8545**	**120.86**	Ours	**0.9825**	0.9530	84.30	5.42	**0.8243**	**302.82**

Table 1. Quantitative comparison with baselines. First line refers to VC2 and AD2 generated original (-ori) videos. Our method enhances both imaging quality and consistency and achieves the highest overall score, with 1.6x-4.5x speedup.

accurate estimation of pixel correspondence continues to induce local inconsistencies (notably seen in the decorative details on the front of the car). This observation is further evidenced in Table 1, which indicates that FRESCO encounters a similar issue. The SC metric, which indicates global consistency, shows improvement across all methods. Conversely, the MS metric, reflecting local inconsistency, experiences a decline for most baselines, whereas our approach demonstrates a notable enhancement. Apart from the overall quality, the inference speed of *EVS* is significantly improved. Rerender-A-Video and FRESCO requires iterative T2I usage across frames to incorporate optical flow, resulting in a time complexity that scales linearly with frames N. Tokenflow requires precise T2I DDIM inversion with sufficient number of steps to ensure that the propagation of intermediate features can match the original video content. For two T2V based baselines, BIVDiff requires an inversion strength of $s_{\text{I}} = 1.0$ for mixed inversion in the initial random noise. AnyV2V relies on precise T2V DDIM inversion to maintain the structural integrity of the source video. Our EVS batchify process all frames without the need for time-consuming accurate inversion. Overall, *EVS* achieves 1.6x-4.5x speedup on these datasets.

4.3. Ablation Studies

Compositions Strategy. As shown in Table 2, pure T2V or T2I can achieve optimal consistency or imaging quality. T2I achieves the highest imaging quality score; however, consistency deteriorates compared original videos due to frame-wise operation. In contrast, T2V produces the highest score in consistency, along with marginally improved imaging quality. Two basic composi-

Figure 6. T2I+T2V declines to T2V imaging quality (zoom in to see the blurred face of Iron Man in red frames). T2V+T2I introduces inconsistency similar to T2I (see flickers in yellow line). Our T2I+[T2V]+T2I balances both aspects.

tions, T2I+T2V/T2V+T2I, can slightly balance two aspects, but their results still tends to favor the model used later. As shown in Table 2, T2I+T2V significantly enhances consistency compared to T2I, but results in a notable decline in imaging quality. Conversely, T2V+T2I achieves the highest imaging quality at the expense of poorer consistency. T2I+[T2V]+T2I achieves a balance with comparable consistency from T2V and imaging quality from T2I.

	Consistency (↑)		Imaging (↑)		Overall (↑)	Time (↓)
	MS	SC	DOVER	AP		
VC2-ori	0.9829	0.9738	55.17	4.50	0.6435	-
T2I	0.9666	0.9710	71.96	**5.48**	0.7567	108.62
T2V	**0.9903**	**0.9817**	62.83	5.03	0.7655	60.03
T2I+T2V	0.9885	0.9800	64.62	5.05	0.7707	155.07
T2V+T2I	0.9868	0.9796	**75.27**	5.44	0.8474	160.31
T2I+[T2V]+T2I	0.9881	0.9808	73.20	5.46	**0.8545**	120.86

Table 2. Ablation study of T2I and T2V compositions.

Hyperparameter Analysis. In Algorithm 1, we utilize four hyperparameters: $t_{\mathrm{I}}, t_{\mathrm{V}}, t_{\mathrm{T2V}}, n_{\mathrm{V}}$. As illustrated in Figure 7, a larger value of t_{T2V} (indicating earlier insertion of the T2V block during T2I) allows for more timesteps to be available for T2I, resulting in improved imaging quality but increased inconsistency. A larger t_{V} (addition of more noise) can more effectively eliminate inconsistencies. However, excessively large t_{V} may cause the frames to converge too closely to the T2V imaging distribution, adversely affecting imaging quality. In the supplementary material, we explore the full combination of these hyperparameters and also apply additional T2V models for enhanced consistency.

Analysis of SFI. To validate the effectiveness of the *Selective Feature Injection* technique in preserving imaging information, we conduct tests on T2I-refined stylized videos from the VBench *Appearance Style* subset. These styles are completely beyond the comprehension of T2V model (e.g. Van Gogh style in Figure 4). For these videos, we first apply

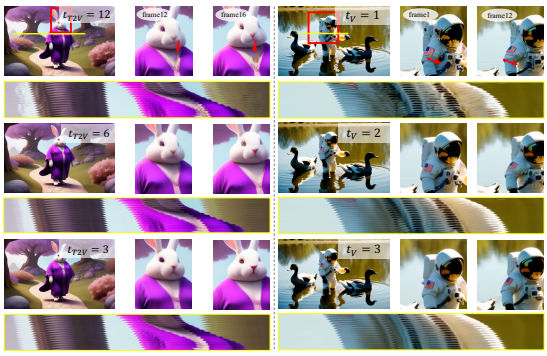

Figure 7. Hyperparameter analysis of the T2V block.

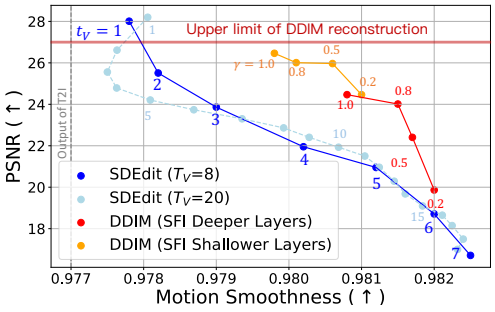

Figure 8. T2V SDEdit v.s. DDIM. It is challenging for SDEdit to strike a balance of maintaining imaging information (PSNR) and introducing T2V motion prior (MS) across all timesteps. DDIM with Selective Feature Injection (SFI) can find empirical points.

SDEdit with different noising strength. Results in Figure 8 indicate that regardless of the total timesteps (8 or 20), as the noising strength increases, there is a noticeable enhancement in temporal smoothness, but this comes at the cost of significant loss in imaging details (lower PSNR). For DDIM inversion, the overall reconstruction PSNR is notably higher than that of SDEdit under same level of motion smoothness. One can selectively choose injected layers and Q_{inv} injected rate γ: injecting in shallower U-Net layers with higher γ can maintain original video in a larger extent. Compared to SDEdit, it is easier to achieve a balanced point. As shown in Figure 8, injecting deeper layers with $\gamma = 0.8$ or shallower layers with $\gamma = 0.5$ are two optical points. Transitioning from higher γ to these two points can enhance motion smoothness with negligible PSNR degradation.

5. Conclusion

In conclusion, our novel training-free encapsulated video synthesizer, *EVS*, successfully bridges the gap between existing pre-trained T2I and T2V models, resulting in higher-quality video synthesis with enhanced visual fidelity and motion smoothness. It also achieves a significant 1.6x-4.5x speedup in inference time. For our future work, we will continue to improve the video quality for T2V pipelines by refining the T2I/T2V denoising process.

Acknowledgment

This work is supported by Key Scientific Research Base for Digital Conservation of Cave Temples (Zhejiang University), State Administration for Cultural Heritage, and Alibaba Research Intern Program.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[2] Duygu Ceylan, Chun-Hao P. Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23206–23217, 2023. 3

[3] 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. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2, 3, 6

[4] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024. 4, 5

[5] Zhongjie Duan, Lizhou You, Chengyu Wang, Cen Chen, Ziheng Wu, Weining Qian, and Jun Huang. Diffsynth: Latent in-iteration deflickering for realistic video synthesis. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2024*, pages 332–347. Springer, 2024. 2

[6] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025. 2

[7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2, 3, 4, 6

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3, 6

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 4

[10] 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

[11] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 3

[12] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2, 3, 6

[13] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 2, 3

[14] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[15] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 3, 6

[16] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 6

[17] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2024. 3

[18] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 2

[19] Shanchuan Lin and Xiao Yang. Animatediff-lightning: Cross-model diffusion distillation. *arXiv preprint arXiv:2403.12706*, 2024. 3, 6

[20] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7817–7826, 2024. 5

[21] Bingyan Liu, Chengyu Wang, Jun Huang, and Kui Jia. Attentive linguistic tracking in diffusion models for training-free text-guided image editing. In *Proceedings of 32nd ACM International Conference on Multimedia*, pages 4158–4166. ACM, 2024. 5

[22] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 3, 4

[23] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Ray-

mond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 3

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3, 4, 5

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 4

[26] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 2, 3

[27] OpenAI. Sora. [Online] https://openai.com/index/video-generation-models-as-world-simulators/, 2024. 1

[28] PikaLabs. Pika 1.0. [Online] https://www.pika.art/, 2023. 1

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[30] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 2, 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[32] Runway. Gen-2. [Online] https://research.runwayml.com/gen2, 2023. 1

[33] Runway. Gen-3. [Online] https://runwayml.com/research/introducing-gen-3-alpha, 2024. 1

[34] Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7393–7402, 2024. 3, 4

[35] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *Asian Conference on Machine Learning*, pages 1215–1230, 2024. 2

[36] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4, 5

[38] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 5

[39] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[40] 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

[41] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 6

[42] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2

[43] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023*, pages 1–11, 2023. 2, 3, 4, 6

[44] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8703–8712, 2024. 3, 6

[45] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[46] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2, 3, 6

[47] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 2

[48] Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, Xiangyang Ji, and Wangmeng Zuo. Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. *arXiv preprint arXiv:2403.05438*, 2024. 3