UFO: Enhancing Diffusion-Based Video Generation with a Uniform Frame Organizer

Delong Liu[1], **Zhaohui Hou**[2], **Mingjie Zhan**[2], **Shihao Han**[2], **Zhicheng Zhao**[1,3,4,*], **Fei Su**[1,3,4]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]SenseTime
[3]Beijing Key Laboratory of Network System and Network Culture, China
[4]Key Laboratory of Intereactive Technology and Experience System, Ministry of Culture and Tourism, Beijing, China
{liudelong, zhaozc, sufei}@bupt.edu.cn, {houzhaohui, zhanmingjie, hanshihao}@sensetime.com

Abstract

Recently, diffusion-based video generation models have achieved significant success. However, existing models often suffer from issues like weak consistency and declining image quality over time. To overcome these challenges, inspired by aesthetic principles, we propose a non-invasive plug-in called Uniform Frame Organizer (UFO), which is compatible with any diffusion-based video generation model. The UFO comprises a series of adaptive adapters with adjustable intensities, which can significantly enhance the consistency between the foreground and background of videos and improve image quality without altering the original model parameters when integrated. The training for UFO is simple, efficient, requires minimal resources, and supports stylized training. Its modular design allows for the combination of multiple UFOs, enabling the customization of personalized video generation models. Furthermore, the UFO also supports direct transferability across different models of the same specification without the need for specific retraining. The experimental results indicate that UFO effectively enhances video generation quality and demonstrates its superiority in public video generation benchmarks.

Code — https://github.com/Delong-liu-bupt/UFO

1 Introduction

The rapid advancement of artificial intelligence has transformed the field of creative content generation. Individuals can quickly obtain personalized text (Zhao et al. 2023), images (Esser et al. 2024), sounds(Du et al. 2024), and videos (Xing et al. 2023) through simple natural language descriptions. In visual generation, diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021), which have excelled in image creation, play a crucial role. However, when applied to video generation, these models encounter challenges such as poor image quality, low aesthetic appeal, and weak consistency. For instance, as shown in Figure 1, even the most advanced open-source models cannot prevent subjects from changing shape throughout a video (e.g., the koala with the staff in Case 1, the kitten in Case 2, and the person with the bag in Case 3), or background inconsistencies (e.g., the boat in Case 2 and the advertising billboard in Case 3).

Aesthetic theory (Wu et al. 2023; Li et al. 2024) in visual media emphasizes the crucial roles of the consistency and clarity in enhancing viewer engagement and perceived quality. In video generation, where dynamic elements and transitions are essential, inconsistencies and blurring not only reduce aesthetic appeal but also undermine the effectiveness of visual communication. To address the challenges mentioned above, we propose the Uniform Frame Organizer (UFO), a non-invasive plug-in designed to enhance the consistency between the foreground and background and alleviate blurring issues, thereby improving video generation quality. Applicable to any diffusion-based video generation model, the UFO integrates a set of non-invasive adapters into the video generation model's backbone network, occupying only $0.005\times$ the size of the original model's trainable parameters. These adapters are capable of autonomously adjusting their intensity of use, featuring a tunable intensity parameter, which is tuned to optimize the balance between dynamic visual content and static precision, reflecting a direct application of aesthetic principles in video generation.

Specifically, when using a small amount of video frames or images as training data, UFO sets the intensity to the highest value, dynamically controlling each adapter's parameters and release intensity to force the model's output to approximate a static video, a scenario of extreme consistency. During this process, the UFO learns to identify and correct inconsistencies in videos. As the pre-trained model's parameters remain unchanged, the UFO's intensity can be adjusted to a lower value during application. This adjustment allows the model output to closely resemble the original while significantly enhancing the consistency between the subjects and the background in the video frames. It also markedly reduces issues such as sudden blurring of video frames.

To achieve the aesthetic consistency, during the training process, the primary optimization goal for the model integrated with the UFO is set to generate static video frames. This simplicity allows the model to learn quickly and converge after only 3000 training steps on a single GPU, using much fewer resources than fine-tuning or retraining video generation models. Moreover, once the parameters of the UFO are obtained, it supports direct transferability across

*Corresponding author

Figure 1: The left side displays three cases: Cases 1 and 3 illustrate that our proposed consistency UFO can be integrated with the model to significantly **enhance the consistency** of generation. Case 2 demonstrates that the UFO can be **directly transferred** and effectively deployed between models of the same specification without the need for training. The cases on the right show that different consistency and stylization UFOs can be **freely combined** to customize video generators.

multiple models of the same specification without the need for model-specific retuning (as shown in Figure 1, Case 2). Beyond enhancing video consistency, the UFO is capable of learning style variations from a limited amount of video-text pairs in the same style. It can also be combined flexibly with the consistency UFO to further enhance the production of videos that not only maintain consistency but also adhere more closely to specific stylistic preferences, as illustrated on the right side of Figure 1.

In practical applications, even the most advanced video generation models often require users to repeatedly adjust parameters and select results that meet their specific needs. In this process, some outcomes may become unusable due to minor consistency flaws or blurriness. The UFO resolves these issues without altering the original video content, significantly easing the challenge of achieving high-quality results. Practical tests on public video generation benchmarks Vbench (Huang et al. 2024) demonstrate that the UFO notably enhances video consistency and quality. In summary, our main contributions are:

- We propose the Uniform Frame Organizer (UFO), a non-invasive plug-in that obviously enhances video consistency and quality, and is compatible with any diffusion-based model. It features a novel adjustable intensity parameter for tuning of video effects.
- The UFO allows for direct transfer between models of the same specification and supports the modular integration of various UFOs, enabling the customization of personalized video generation models.
- Training UFO is very inexpensive, and enhances consistency without the need for video-text pairs.
- UFO significantly reduces the effort required by users to obtain high-quality videos, and the extensive experiments verify its efficiency and effectiveness.

2 Related Work

The Diffusion Model (DM) has consistently excelled in image (Nichol et al. 2022; Ramesh et al. 2022; Zhang, Rao, and Agrawala 2023) and video generation (Ma et al. 2024a; Khachatryan et al. 2023a; Lu et al. 2024), and has also expanded across various video generation tasks, including text-to-video (Luo et al. 2023; Wang et al. 2023b), image-to-video (Yin et al. 2023; Chen et al. 2023c), video-to-video (Liew et al. 2023; Ouyang et al. 2024), and applications under diverse control conditions such as pose (Karras et al. 2023; Ma et al. 2024b), depth (Chen et al. 2023b; Zhang et al. 2024), and sketch (Khachatryan et al. 2023b; Wang et al. 2024). In the past two years, the text-to-video generation, our primary focus, has made rapid progress. Early work like Image Video (Ho et al. 2022) highlighted diffusion models' ability to produce high-quality videos. However,

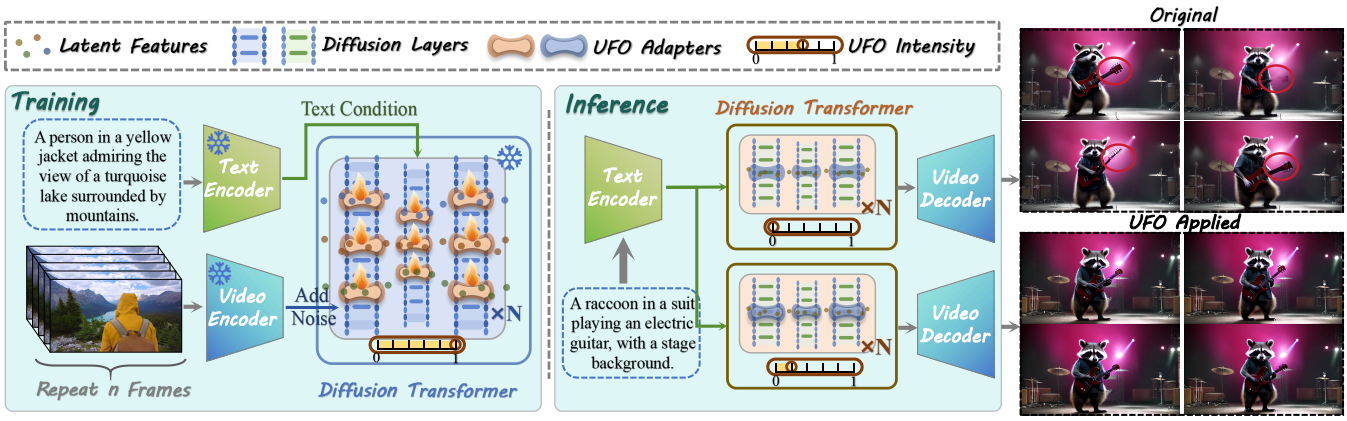

Figure 2: Training and inference of consistency UFO. During training, all parameters of the original model are frozen, and the UFO operates at maximum intensity using image-text pair data, with images duplicated across multiple frames to meet training requirements. In inference, zero intensity mirrors the original generator, while low intensity improves video consistency. The right images compare these two scenarios.

aligning videos precisely with text prompts while maintaining visual appeal remained challenging. Subsequent models, such as VideoCrafter (Chen et al. 2023a, 2024) and ModelScopeT2V (Wang et al. 2023a), were trained with large video datasets including WebVid-10M (Bain et al. 2021) and InternVid (Wang et al. 2023c). These models enhanced the aesthetics of videos but often produced shorter videos with inconsistent quality and blurred visual details. The recent Gen-2 (Esser et al. 2023) and Pika (Pika-Labs 2023), with their massive data and complex iterative sampling processes, enhanced video image quality. However, their stability and realism still need improvement, and the computational cost is also quite high.

The introduction of the Transformer-based Diffusion (DiT) architecture and it scalable parameter capability has opened new opportunities. Sora (OpenAI 2024), using DiT in its backbone network and exploiting vast datasets, has pioneered unparalleled zero-sample video generation, producing longer videos of higher quality. Although this technology is not yet publicly available, it has significantly spurred the open-source community's exploration of DiT-based video generation models (Zheng et al. 2024; PKU-Yuan-Lab and Tuzhan-AI 2024; Xu et al. 2024a), leading to a series of high-quality video generation models that substantially surpass previous models. Nonetheless, as video length increases, challenges like video consistency and blurring persist. In response, we propose a low-cost, widely applicable consistency enhancement plug-in, UFO, which is validated on the best-performing open-source models (Zheng et al. 2024; Xu et al. 2024a).

3 Methodology

As shown in Figure 2, the UFO includes a series of lightweight adapters (Section 3.1) that can be non-destructively attached to any mapping layer of the model without altering the original model's parameters. During training (Section 3.2), only the UFO's parameters are updated, while the intensity is set to the highest to achieve extremely consistent videos under text conditions, rendering all video frames static. This phase drives the UFO to develop the ability to identify and correct inconsistencies. During inference (Section 3.3), setting the UFO's intensity to 0 will result in an output that is identical to that of the original pre-trained model. When it is at low level, the output will closely resemble the original, maintaining motion in video frames. The UFO's targeted repair capabilities enhance the consistency between subjects and backgrounds and mitigate video quality degradation. For example, when applying the same prompt and fixed random seed, the UFO-generated appearance and attire of the raccoon are more consistent, notably preventing significant shape transformations in the electric guitar being played, as shown in Figure 2.

3.1 Lightweight Adapters for UFO

To achieve cost-effective improvements in video generation models and eliminate reliance on a single model framework, inspired by efficient parameter fine-tuning methods (Pfeiffer et al. 2020; Hu et al. 2022), we design a series of adapters, each of which consists of a layer with minimal input or output dimensions, and is injected into the diffusion model with minimal overhead. These adapters act as the smallest subunits for controlling the consistency of hidden features in video frames, enabling precise, targeted consistency corrections.

Specifically, in a module parameterized by $\mathbf{W} \in \mathbb{R}^{m\times n}$ in the DIT, we learn a detection layer $\mathbf{v}_{det} \in \mathbb{R}^{n\times d}$ to precisely locate features affecting video consistency. Concurrently, a correction layer $\mathbf{v}_{cor} \in \mathbb{R}^{m\times d}$ modifies the identified features. Here, d is chosen to be small to ensure parameter efficiency. Consequently, the original representation $\mathbf{y} = \mathbf{W}\mathbf{x}$ is modified as follows:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \alpha\beta(\mathbf{v}_{det}^T\mathbf{x}) \cdot \mathbf{v}_{cor} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ represent the input and output of

the intermediate layer, respectively, the superscript T indicates transposition, α is a adjustable intensity factor, and β is a learnable dynamic intensity factor, aiming to dynamically adjust the strength of each adapter of the UFO to ultimately correct the consistency of video frames.

Focusing on just two frames $y_t, y_{t+n} \in \mathbf{y}$ in the video, the difference in output between these two intermediate layers $\Delta y_{t,t+n}, = y_t - y_{t+n}$ can be expressed as:

$$\begin{aligned}\Delta y_{t,t+n} &= (\mathbf{W}x_t - \mathbf{W}x_{t+n}) + \alpha\beta((\mathbf{v}_{det}^T x_t)\cdot \mathbf{v}_{cor} \\ &\quad - (\mathbf{v}_{det}^T x_{t+n})\cdot \mathbf{v}_{cor}) \\ &= \mathbf{W}\Delta x_{t,t+n} + \alpha\beta\Delta\mathbf{v}_{det}^T\Delta x_{t,t+n}\cdot \mathbf{v}_{cor}.\end{aligned} \tag{2}$$

During the training phase, with $\alpha = 1$, the target is for all video frames to be identical, thus $\Delta y_n = 0$ regardless of the value of n. Therefore, the optimization goal for each latent feature is $-\beta\Delta\mathbf{v}_{det}^T\Delta x_{t,t+n}\cdot\mathbf{v}_{cor}$, meaning that the trained β, $\mathbf{v}_{det}$, and $\mathbf{v}_{cor}$ can adaptively identify and fill the variations in each video frame.

Utilizing this feature, during inference, α is set to a low value, ensuring that the variability $\Delta y_n \approx \mathbf{W}\Delta x_{t,t+n}$ in video frames maintains the subjects, background, and motion capabilities essentially consistent with those of the pretrained model's output videos, while the additional term $\alpha\beta\Delta\mathbf{v}_{det}^T\Delta x_{t,t+n}\cdot\mathbf{v}_{cor}$ has a comprehensive view of the changes in video frames, adaptively enhancing the consistency of the output video. Furthermore, if the intensity factor α is fixed, UFO, due to its parametric characteristics, also supports using a small batch of video-text pairs to directionally fine-tune the video generation model, customizing the video generation effects.

3.2 Training of UFO

During the training phase, video data $\mathbf{V} \in \mathbb{R}^{F\times H\times W\times C}$ is first compressed into a latent space representation $z = \mathcal{E}(\mathbf{V})$ using a pretrained variational autoencoder (VAE) (Kingma and Welling 2013). Additionally, a textual condition c is introduced, which is derived from a text encoder using prompts aligned with the video content. In the generation process, the diffusion model gradually introduces noise to simulate the diffusion of video data, forming perturbed samples $z_t = \sqrt{\overline{\alpha}_t}z + \sqrt{1-\overline{\alpha}_t}\epsilon$, where $\epsilon \sim N(0,1)$ represents noise sampled from a standard normal distribution, and $\overline{\alpha}_t$ serves as a noise scheduler, with t denoting the diffusion time step.

After integration with UFO, the parameters of the original model are denoted as θ, and only the parameters within UFO are updated during training. The reverse diffusion process, which is essentially training the model to denoise, aims to predict the less noisy z_{t-1}: $p_\theta(z_{t-1}|z_t) = N(\mu_\theta(z_t), \Sigma_\theta(z_t))$. Here, the log likelihood of the variational lower bound simplifies to $L_{vlb}(\theta) = -\log p(z_0|z_1,c) + \sum_t D_{KL}\left(q(z_{t-1}|z_t,z_0)\|p_\theta(z_{t-1}|z_t)\right)$. Since both q and p_θ are Gaussian, the D_{KL} term is determined by the mean μ_θ and covariance Σ_θ. The μ_θ is reparametrized into the denoising model ϵ_θ, which can be trained using a simple objective:

$$L_{simple}(\theta) = \mathbb{E}_{z\sim p(z),\epsilon\sim N(0,1),t,c}\left[\|\epsilon - \epsilon_\theta(z_t,t,c)\|_2^2\right], \tag{3}$$

According to (Nichol and Dhariwal 2021), it is necessary to fully optimize the D_{KL} term (i.e., train using the full L_{vlb}) to train an LDM with learnable covariance Σ_θ. Therefore, the training loss for UFOs employs both L_{simple} and L_{vlb}.

When training the consistency UFO, every frame in $\mathbf{V}$ used is identical, thus image-text pairs, which are more readily available, can be used as training data. For customizing stylization UFOs, regular video-text pairs are used as training data.

3.3 Inference

During inference, the trained UFO is integrated into the diffusion model, retaining all functionalities of the original model. For the consistency UFO, its intensity factor α tends to be set to a low value, which can be adjusted based on the performance of the original pre-trained model during video inference. If issues such as inconsistency or blurring are severe, α should be increased, which enhances video frame consistency. This adjustment allows users to control video consistency according to their needs. For stylization UFOs, α is suggested to match the level used during training, and minor adjustments can optimize personalization. Note that when combining different UFOs, the intensity of each UFO needs to be adjusted as required.

4 Experimental Results

4.1 Settings

Implementation details To ensure the rigor of our experiments, we train UFOs using two of the latest text-to-video open-source models, EasyAnimate-V2 (Easy) (Xu et al. 2024a) and OpenSora-V1.2 (Open) (Zheng et al. 2024). Training is conducted on 4 NVIDIA A100 GPUs, with inference running on a single GPU. During the training process, only the parameters of the UFOs are updated, with each UFO undergoing 3000 training steps. All adapters have a hyperparameter dimension $d = 4$, and gradient accumulation is not used. For Open, a linear warm-up strategy is employed in the first 500 steps, where the learning rate gradually increases from nearly zero to $2e-4$, and this rate is maintained after the warm-up phase. For Easy, the learning rate is set at $1e-4$ and remains constant. The rest of the training settings follow the original methods. During inference, all settings use the recommended configurations of the original methods, with videos set at 24 Frames Per Second (FPS), and all experiments and visual effects in the text use the same random seed to compare with and without UFOs. More details on training and inference can be found in the supplementary materials.

Training Datasets For training the consistency UFO, we use a subset of the LAION-Aesthetics V2 (Schuhmann et al. 2022) dataset with aesthetic scores above 6.5, from which we extract 12K image-text pairs to create static video-text pairs for training. For the training of stylization UFOs, we collect 300 videos for each of the four styles (Pixel Art, oil painting, animated style, black and white) from publicly available video resources on the internet. The text for these videos is automatically annotated using the 13B version of

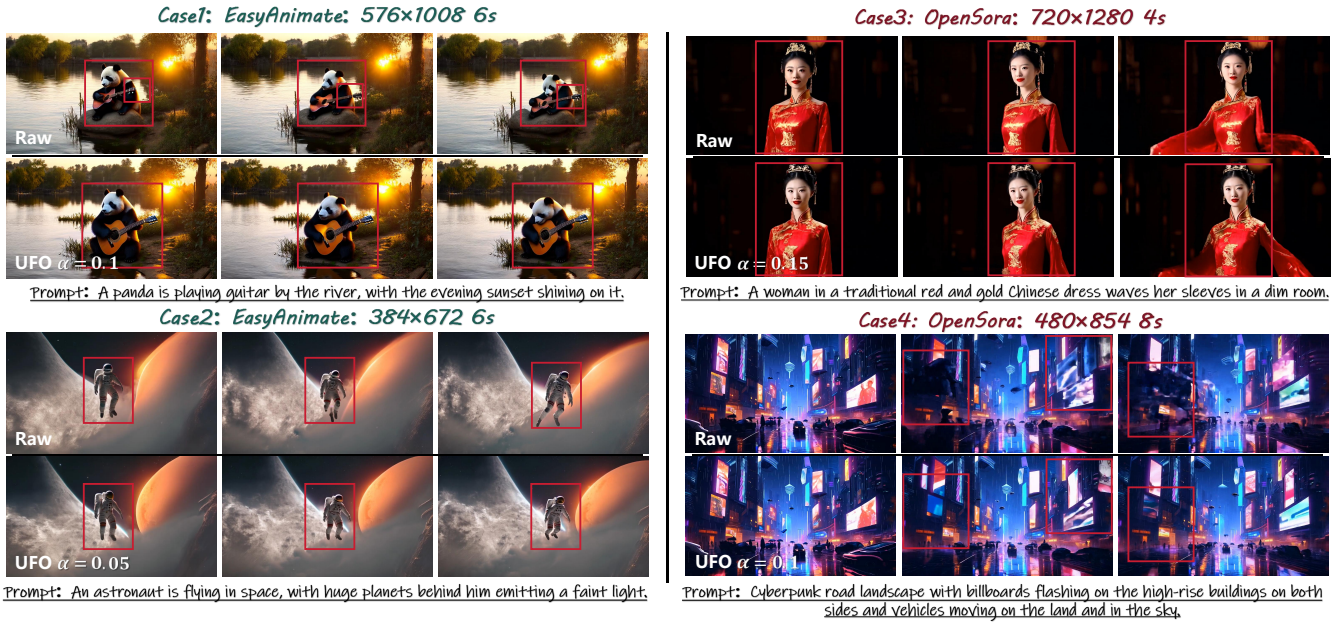

Figure 3: Visualizations of the consistency UFO. The areas highlighted in red boxes show inconsistencies or blurriness in the videos produced by the pre-trained model.

No.	Model	Resolution	UFO/α	TQ	SC	BC	TF	MS	FWQ	AQ	IQ	SQ	EC
1	Open	240 × 426	0	95.55%	93.00%	94.91%	97.84%	96.43%	55.25%	51.96%	58.54%	66.68%	-
2			0.1	96.46%	94.53%	95.52%	98.49%	97.29%	55.43%	52.25%	58.61%	67.15%	51/1165
3			0.2	**97.01%**	**95.08%**	**96.32%**	**98.71%**	**97.91%**	**56.23%**	**53.05%**	**59.40%**	**67.78%**	86/1165
4		480 × 854	0	95.24%	93.04%	93.62%	98.30%	95.99%	59.49%	57.01%	61.96%	72.13%	-
5			0.1	96.35%	94.45%	94.88%	98.97%	97.08%	60.19%	57.63%	62.75%	72.24%	35/1165
6			0.2	**97.00%**	**95.48%**	**95.68%**	**99.14%**	**97.70%**	**60.50%**	**58.03%**	**62.96%**	**72.40%**	80/1165
7		720 × 1280	0	95.33%	92.47%	95.22%	98.38%	95.23%	60.42%	57.27%	63.57%	73.00%	-
8			0.1	96.70%	94.50%	96.55%	99.06%	96.67%	60.61%	57.59%	63.62%	73.78%	39/1165
9			0.2	**97.26%**	**95.34%**	**97.08%**	**99.27%**	**97.36%**	**60.97%**	**57.91%**	**64.03%**	**73.80%**	71/1165
10	Easy	384 × 672	0	97.07%	94.67%	96.84%	99.49%	97.26%	63.53%	62.78%	64.27%	71.21%	-
11			0.07	98.20%	96.44%	97.73%	99.68%	98.94%	63.64%	62.95%	64.32%	71.48%	28/1165
12			0.15	**99.02%**	**98.06%**	**98.54%**	**99.81%**	**99.66%**	**63.77%**	**63.05%**	**64.48%**	**71.82%**	81/1165
13		576 × 1008	0	96.41%	92.94%	96.49%	99.30%	96.91%	63.57%	61.43%	65.71%	70.58%	-
14			0.07	97.53%	94.80%	97.52%	99.57%	98.22%	64.42%	62.08%	66.75%	71.64%	21/1165
15			0.15	**98.57%**	**96.69%**	**98.63%**	**99.68%**	**99.28%**	**64.98%**	**62.31%**	**67.65%**	**72.08%**	75/1165

Table 1: Impact of the consistency UFO on the performance of different base models. The values for TQ and FWQ represent the mean scores across their respective dimensions.

the PLLaVA (Xu et al. 2024b) model, with descriptions regarding the video style removed during training.

Evaluation Metrics To objectively demonstrate the improvements in video consistency and quality achieved by the UFO, we employ the latest video generation evaluation method, Vbench, using a fixed intensity setting for the consistency UFO. This evaluation encompasses two main dimensions: Video Quality (VQ) and Semantic Quality (SQ). As our approach does not specifically target enhancements in video semantic consistency, our primary focus is on the VQ metrics. These include four dimensions of "Temporal Quality" (TQ): "Subject Consistency" (SC), "Background Consistency" (BC), "Temporal Flickering" (TF), "Motion Smoothness" (MS), and two dimensions of "Frame-Wise Quality" (FWQ): "Aesthetic Quality" (AQ) and "Imaging Quality" (IQ). We also assess the dimensions related to SQ, providing only a total score. For a single complete evaluation, a total of 4720 videos inferred from Vbench's official prompts are used, of which 1165 videos relate to the four dimensions of TQ. Since some pre-trained model inferences produce videos with minimal visual changes, using a fixed intensity UFO can cause the frames to become nearly static, potentially skewing the TQ metrics. Consequently, we exclude such videos from the evaluation, recorded as "Excluded Count" (EC), to reflect the impact of the consistency UFO. More details on the criteria for judging near-static conditions and the specifics of the metrics are available in the supplementary materials.

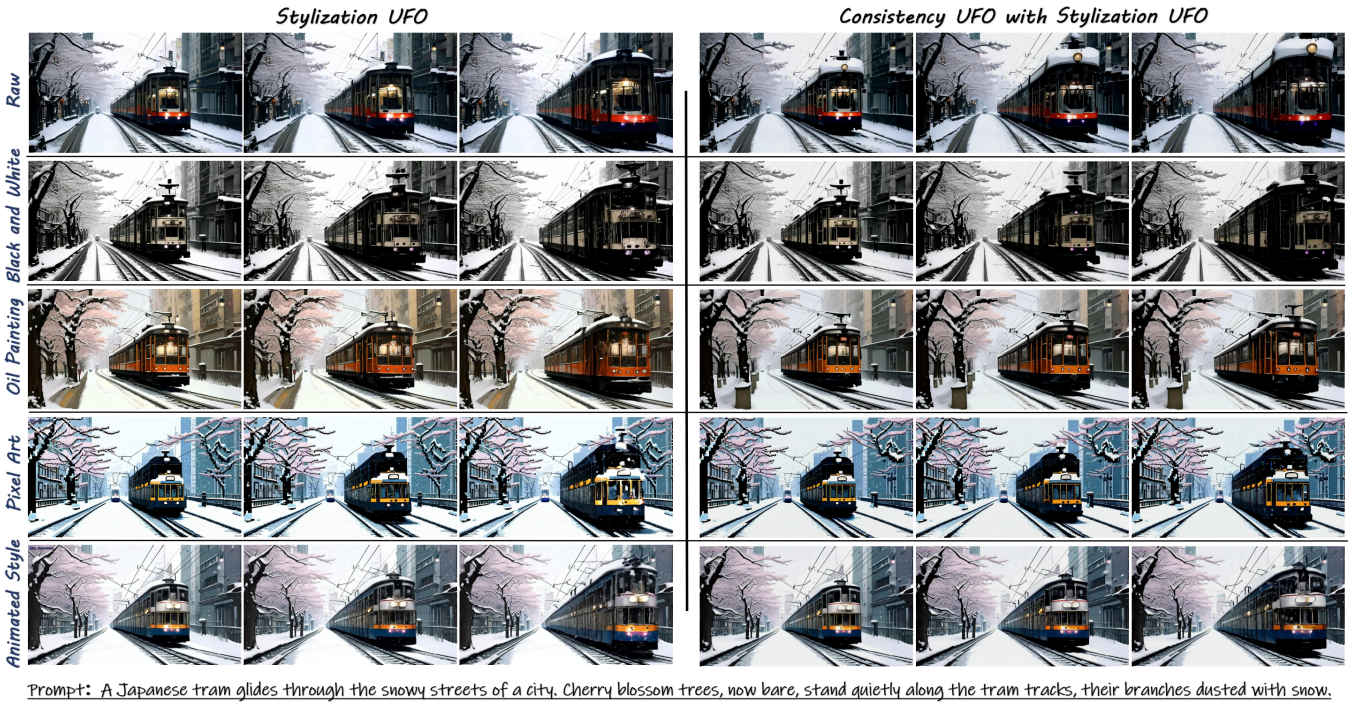

Figure 4: Examples of the effects of consistency UFO with stylization UFO. The first row illustrates the results without using stylization UFO, while rows two to five demonstrate the effects of different stylization UFOs. Videos on the right have added consistency UFO compared to those on the left. In these cases, all the stylization UFOs have $\alpha = 1$ and the consistency UFO have $\alpha = 0.1$, all generated by Open with a resolution of 720×1280 and a duration of 4 seconds.

4.2 Quantitative Results

We evaluate the consistency UFO on two baseline models, Easy and Open. Open supports high-quality video generation across multiple resolutions with a single model, while Easy performs poorly when handling different resolutions, requiring the use of two separate models for videos of various resolutions. Videos used for Vbench evaluation are rendered at typical resolutions supported by the original models, with each video running for 4 seconds. The results are shown in Table 1. When $\alpha = 0$, it reflects the performance without UFO, while $\alpha > 0$ indicates the use of UFO at varying intensities. In our primary dimension of concern, Temporal Quality (TQ), it is clear that UFO significantly enhances both the consistency between the subject and background, and the smoothness of video motion. A higher intensity of UFO leads to more pronounced improvements, but it may also cause more videos with minimal dynamics to become static. However, in practical use, users can freely adjust the intensity of UFO based on video outcomes, thus avoiding such issues. Similarly, the Frame-Wise Quality (FWQ) dimension related to image quality shows the same trend because UFO effectively eliminates blurring and flickering issues in the video, thereby enhancing image quality. Surprisingly, UFO also results in gains in the Semantic Quality (SQ) dimension, likely due to enhancements in TQ and FWQ dimensions that improve the visual expression stability of the generated videos.

Notably, although Easy uses two different models of the same specification to process videos of two resolutions, the same consistency UFO plugin is utilized. It was only trained on the model handling higher resolutions, suggesting that UFO can effectively transfer between models and achieve the desired effects.

4.3 Qualitative Results

Consistency UFO In Figure 3, we showcase four qualitative results on two baseline models to illustrate the intuitive effects and characteristics of our consistency UFO. In Case 1, without the use of the consistency UFO, the appearance and size of the panda, as well as the guitar, are inconsistent. After applying the UFO with $\alpha = 0.1$, the panda and the guitar maintain consistency, preserving the composition and main elements of the original video. Case 2 demonstrates that transferring the consistency UFO directly to another model is also effective, significantly improving the consistency of the astronaut depicted in the image. Case 3 displays the universality of the UFO, which is effective under any framework, significantly enhancing the appearance and posture consistency of the woman in the image. Case 4 primarily shows that the consistency UFO can address issues of blurring and flickering in long video generation, effectively alleviating the blurring of billboards and the flickering of black objects in long videos. In practical applications, the intensity level can be adjusted based on the degree of inconsistency or blurriness in the original video to optimize the output.

Model	Params.	Time	UFO/α	d	TQ	FWQ	SQ
Open	1.14 B	-	0	-	95.24%	59.49%	72.13%
	1.42 M	0.24%	0.1	1	95.37%	59.43%	72.09%
	2.83 M	0.33%	0.1	2	95.98%	59.89%	72.19%
	5.66 M	0.40%	0.1	**4**	**96.35%**	**60.19%**	72.24%
	11.32 M	0.51%	0.1	8	96.29%	60.11%	**72.38%**
	90.56 M	1.08%	0.1	64	96.03%	59.67%	71.98%
Easy	818.17 M	-	0	-	96.41%	63.57%	70.58%
	1.89 M	0.31%	0.07	1	96.53%	63.76%	70.76%
	3.79 M	0.40%	0.07	2	97.16%	64.28%	71.12%
	7.58 M	0.49%	0.07	**4**	**97.53%**	64.42%	**71.64%**
	15.15 M	0.58%	0.07	8	97.49%	**64.50%**	71.62%
	121.21 M	1.44%	0.07	64	97.14%	63.97%	70.89%

Table 2: Performance changes associated with different dimensions d in the consistency UFO adapters. The 'Params.' column represents the amount of the diffusion model's parameters when UFO is not used, and the trainable parameters when UFO is in use. 'Time' indicates the percentage increase in time required to infer a single video compared to the original model. All performance metrics are based on the inference of 4-second videos, with a resolution of 480×854 for Open and 576×1008 for Easy.

Consistency UFO with Stylization UFO Due to the parametric characteristics of UFO, various styles of stylization UFOs can be customized. Figure 4 demonstrates the effects of combining a pre-trained model with stylization UFOs, which can transform original videos into various styles while preserving the fundamental elements and layout of the original image. For example, despite the videos on the left side of the figure having different styles, elements such as the direction of the train, the cherry trees on the left side of the road, the buildings on the right, and the shooting angle remain consistent. However, this transformation still preserves the subject inconsistencies present in the original video. Therefore, by leveraging the flexibility of UFO, combining stylization UFO with consistency UFO can produce more consistent personalized videos. Comparing the video frames on the left and right sides of the figure, it is noticeable that maintaining consistency in the tram's front and doors is challenging across all styles, with even the colors (oil painting style) and proportions (animated style) varying. These issues are effectively addressed in the video frames on the right.

4.4 Ablition Studies

UFO Adapters Dimension Table 2 illustrates the performance changes when different dimensions d are used in the consistency UFO adapters. It is evident that increasing d up to 4 effectively enhances both consistency and image quality. However, further increases beyond 4 do not yield additional gains. When d becomes too large, the UFO begins to fit the characteristics of the limited data used for training. While this still can enhance the consistency of the images, it causes the content of the visuals to diverge from those generated by the original model. Although increasing d does not significantly reduce inference speed, it introduces more parameters, tailoring the model more closely to the specific characteristics of the training data. Therefore, after comprehensive consideration, $d = 4$ is established as the optimal setting for the UFO.

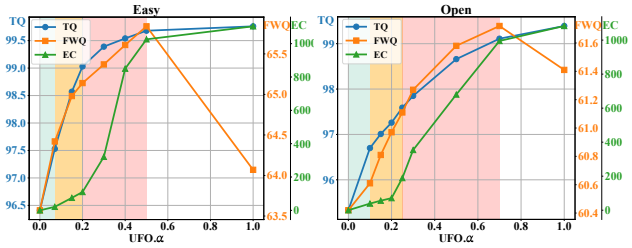

Figure 5: Metric variations with different α levels for the consistency UFO. Left image metrics are from a 4-second video at 576×1008 on Easy. Right image metrics are from a 4-second video at 720×1280 on Open. Light blue indicates conservative strategy, orange for moderate, and red for aggressive.

UFO Intensity α To achieve optimal visual effects for specific prompts, one can initially obtain preliminary results from the video generation model and then adjust the consistency UFO's α based on the degree of inconsistency observed. For scenarios requiring enhanced general generative capabilities, α needs to be preset. Figure 5 illustrates the impact of the consistency UFO with fixed α values, showing similar metric trends across two models. With a conservative strategy, videos maintain their dynamism while improving in consistency and quality. A moderate strategy enhances these aspects significantly, though it may slow down motion in some videos. An aggressive strategy markedly increases consistency and quality but can lead many videos to become nearly static. Further increasing α risks making nearly all generated videos unusable. Since videos generated by the Easy model exhibit inherently less motion than those from Open, α settings are adjusted more conservatively across these strategies.

5 Conclusion

In this paper, we propose and validate the UFO, a non-invasive plugin for diffusion-based video generation models. By integrating the UFO into existing models, its effectiveness in mitigating common problems like video quality degradation and frame inconsistency is demonstrated, and without significantly increasing computational demands. In addition, The proposed intensity α also provides users with the flexibility to control video consistency, facilitating the creation of videos that meet their specific needs. Moreover, the UFO's modular design and low resource requirements make it easily transferable between different models, thus enhancing their flexibility and scalability. In the future, we aim to improve the UFO so that it can reliably enhance video quality by automatically intensity adjusting.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants 62076033.

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; Weng, C.; and Shan, Y. 2023a. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. arXiv:2310.19512.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047.

Chen, W.; Ji, Y.; Wu, J.; Wu, H.; Xie, P.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023b. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*.

Chen, X.; Wang, Y.; Zhang, L.; Zhuang, S.; Ma, X.; Yu, J.; Wang, Y.; Lin, D.; Qiao, Y.; and Liu, Z. 2023c. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; et al. 2024. CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. *arXiv preprint arXiv:2407.05407*.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7346–7356.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Karras, J.; Holynski, A.; Wang, T.-C.; and Kemelmacher-Shlizerman, I. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 22623–22633. IEEE.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023a. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15954–15964.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023b. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15954–15964.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, X.; Yuan, K.; Pei, Y.; Lu, Y.; Sun, M.; Zhou, C.; Chen, Z.; Timofte, R.; Sun, W.; Wu, H.; et al. 2024. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6415–6431.

Liew, J. H.; Yan, H.; Zhang, J.; Xu, Z.; and Feng, J. 2023. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*.

Lu, H.; Yang, G.; Fei, N.; Huo, Y.; Lu, Z.; Luo, P.; and Ding, M. 2024. VDT: General-purpose Video Diffusion Transformers via Mask Modeling. In *The Twelfth International Conference on Learning Representations*.

Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Zhao, D.; Zhou, J.; and Tan, T. 2023. Videofusion: Decomposed diffusion models for high-quality video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10209–10218. IEEE.

Ma, X.; Wang, Y.; Jia, G.; Chen, X.; Liu, Z.; Li, Y.-F.; Chen, C.; and Qiao, Y. 2024a. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024b. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

OpenAI. 2024. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: 2024-08-01.

Ouyang, H.; Wang, Q.; Xiao, Y.; Bai, Q.; Zhang, J.; Zheng, K.; Zhou, X.; Chen, Q.; and Shen, Y. 2024. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8089–8099.

Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; and Gurevych, I. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Pika-Labs. 2023. Pika Labs. https://www.pika.art/. Accessed June 10, 2024.

PKU-Yuan-Lab; and Tuzhan-AI. 2024. Open-Sora-Plan.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2024. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023b. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; et al. 2023c. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20144–20154.

Xing, Z.; Feng, Q.; Chen, H.; Dai, Q.; Hu, H.; Xu, H.; Wu, Z.; and Jiang, Y.-G. 2023. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*.

Xu, J.; Zou, X.; Huang, K.; Chen, Y.; Liu, B.; Cheng, M.; Shi, X.; and Huang, J. 2024a. EasyAnimate: A High-Performance Long Video Generation Method based on Transformer Architecture. *arXiv preprint arXiv:2405.18991*.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024b. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. arXiv:2404.16994.

Yin, S.; Wu, C.; Liang, J.; Shi, J.; Li, H.; Ming, G.; and Duan, N. 2023. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Wei, Y.; Jiang, D.; ZHANG, X.; Zuo, W.; and Tian, Q. 2024. ControlVideo: Training-free Controllable Text-to-video Generation. In *The Twelfth International Conference on Learning Representations*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024. Open-Sora: Democratizing Efficient Video Production for All.