Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065529

ARTICLE

Optimizing Semantic and Texture Consistency in Video Generation

Xian Yu, Jianxun Zhang*, Siran Tian and Xiaobao He

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

*Corresponding Author: Jianxun Zhang. Email: zhangjxcqut@sina.com

Received: 15 March 2025; Accepted: 17 July 2025; Published: 29 August 2025

ABSTRACT: In recent years, diffusion models have achieved remarkable progress in image generation. However, extending them to text-to-video (T2V) generation remains challenging, particularly in maintaining semantic consistency and visual quality across frames. Existing approaches often overlook the synergy between high-level semantics and low-level texture information, resulting in blurry or temporally inconsistent outputs. To address these issues, we propose Dual Consistency Training (DCT), a novel framework designed to jointly optimize semantic and texture consistency in video generation. Specifically, we introduce a multi-scale spatial adapter to enhance spatial feature extraction, and leverage the complementary strengths of CLIP and VGG—where CLIP focuses on high-level semantics and VGG captures fine-grained texture and detail. During training, a stepwise strategy is adopted to impose semantic and texture losses, constraining discrepancies between generated and ground-truth frames. Furthermore, we propose CLWS, which dynamically adjusts the balance between semantic and texture losses to facilitate more stable and effective optimization. Remarkably, DCT achieves high-quality video generation using only a single training video on a single NVIDIA A6000 GPU. Extensive experiments demonstrate that our method significantly improves temporal coherence and visual fidelity across various video generation tasks, verifying its effectiveness and generalizability.

KEYWORDS: Diffusion model; dynamic weighting; text-to-video; one-shot

1 Introduction

With the rapid development of AIGC, the diffusion model [1] has become an important breakthrough in the field of generative models. Although traditional generative models such as generative adversarial networks (GANs) [2] have achieved remarkable results in image generation tasks, they often face problems of instability and mode collapse during training.

Diffusion models have advanced image synthesis [3], yet video generation remains challenging [4] due to requirements for high-quality frames and temporal coherence. Prior works [1,5] incorporate semantic consistency in diffusion for images, with CLIP guidance improving resolution [6]. VDM [7] applies 3D convolutions and spatiotemporal attention but faces instability. Zhang et al. [8] show that disentangling semantics and texture enhances frame quality; however, their joint optimization for temporal coherence is still unresolved.

To address the challenges of video generation, recent studies [9] explore combining diffusion models with other mechanisms to enhance diversity, detail, and consistency. A key approach is jointly optimizing semantic and texture feature consistency, especially in complex dynamic scenes [10]. Leveraging semantic and texture features from deep models like CLIP [11] and VGG improves not only individual frame quality but also temporal smoothness and semantic coherence, thereby elevating overall video quality.

This paper proposes a video generation method, DCT, based on a diffusion model, which improves video generation by optimizing semantic and texture consistency. We use CLIP to capture high-level semantic information and VGG to extract low-level textures and details, ensuring that each frame is visually coherent and temporally consistent. By dynamically adjusting the weights of semantic and texture consistency, we balance their contributions and gradually improve the generation quality. Experimental results show that the proposed method effectively enhances video consistency and overcomes the limitations of traditional diffusion models in video generation. Our main contributions include:

- A multi-scale spatial adapter module is designed to enhance the expressiveness of input features through multi-scale feature fusion and optional skip connections.
- A new dual consistency training method DCT is proposed, which balances the generation quality and training cost by optimizing the consistency of semantic and texture features, reduces the computational complexity, and ensures the high quality and consistency of video generation.
- A dynamic weight adjustment mechanism CLWS is introduced, which adjusts the semantic and texture consistency weights according to the time step during the training process and flexibly controls the optimization degree of the two, thereby improving the quality of the generated video and the training efficiency.

2 Related Work

2.1 Text-to-Video Generation

Video generation has become an important research area in GANs and diffusion models. Traditional methods use GANs for realistic video generation through adversarial training between the generator and discriminator. However, due to the temporal dimension in video data, GANs often struggle with maintaining both spatial and temporal consistency. To address this, approaches like MoCoGAN [12] and TGAN [13] introduced temporal models, but they still face issues with mode collapse and inadequate temporal consistency. Existing work [14] extends the diffusion process to the temporal dimension, using noise progressive restoration to generate high-quality, temporally consistent videos. While existing text-to-video (T2V) generation models have achieved impressive results, their performance heavily relies on training with large-scale video datasets. In contrast, we propose a novel framework that enables efficient T2V generation by effectively adapting a pretrained text-to-image (T2I) diffusion model using only a single text-video pair.

2.2 Semantic Consistency

The two-stream model of CLIP [5] embeds images and text into a shared high-dimensional feature space, allowing it to compute the semantic distance between the generated frame and the target frame through feature alignment, ensuring semantic accuracy of the generated content. VQ-VAE-2 [15] combines semantic feature extraction and modulation modules, increasing generation diversity while ensuring content consistency. Additionally, StyleGAN3 [16] and VideoGPT [17] also integrate semantic feature extraction modules to optimize visual and semantic consistency during generation.

2.3 Texture Consistency

Texture consistency requires not only the restoration of detailed texture in a single frame but also maintaining texture coherence across frames, especially in dynamic scenes such as flowing water, fluttering leaves, or moving vehicles. The classic perceptual loss method [18] uses a pre-trained VGG model to extract deep features, ensuring visual consistency. Subsequent research introduced multi-scale texture modeling,

using pyramid structures or multi-resolution feature extraction to constrain both low and high-resolution content. Texture Networks [19] captures texture features at multiple levels, TecoGAN [20] improves cross-frame texture consistency by adding temporal consistency constraints, and CVT [21] optimizes texture consistency by constraining the similarity between video frames. Building upon these foundations, our proposed model introduces Dual Consistency Training (DCT), which jointly optimizes the alignment of semantic and texture features. This design effectively enhances temporal coherence and visual fidelity in text-to-video generation tasks.

2.4 Dynamic Weighting Strategy

In video generation, optimizing both semantic and texture feature losses is a complex challenge. Semantic features capture content and contextual consistency, while texture features focus on visual details. The dynamic weighting strategy [22] adjusts the weights of semantic and texture losses during training for balanced optimization. Mao et al. [23] proposed a time-step-based linear weighting strategy, prioritizing semantic consistency initially and gradually shifting to texture consistency. Existing dynamic weighting methods often ignore detailed frequency information important for balancing semantic and texture losses. To overcome this, we propose a multi-scale Fourier weighting strategy that uses spectral cues to adaptively adjust loss weights, improving optimization stability and video quality.

3 Method

3.1 Preliminaries of Stable Diffusion

The diffusion model is an important branch of generative model research in recent years. Its core idea is to transform data $x_0 \in X$ into random noise and gradually restore the original data through the reverse process. In this section, we will introduce the forward diffusion process, reverse restoration process and training objectives of the diffusion model.

The forward diffusion process gradually converts data into Gaussian noise via a Markov chain [24]. At each of the T steps, noise is incrementally added, with the transition probability defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{1}$$

where $t = 1, \ldots, T$ is the time step; β_t is a hyperparameter that controls the noise intensity at each step, which is usually increased in training, such as linear or cosine increase. After T steps, the data will gradually evolve into standard Gaussian noise $x_t \sim N(0, I)$. At any time step t, the data x_t can be expressed as a linear weighted combination of the original data x_0 and the noise:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_t)$ is the cumulative scaling factor at time step t and ϵ represents the noise sampled from a standard Gaussian distribution.

The goal of the reverse diffusion process is to gradually restore each step of adding noise in the forward diffusion process and generate the original data x_0 from the noise x_t. The conditional probability of the reverse process is:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \tag{3}$$

where $\mu_\theta(x_t, t)$ is the mean estimated by the neural network and $\Sigma_\theta(x_t, t)$ is the covariance matrix (usually fixed to be a diagonal matrix).

The core training objective of the diffusion model is to learn the noise ϵ added during the $\epsilon_\theta\ (x_t, t)$ forward diffusion process through a neural network. By simplifying the formula, the training objective can be defined as the mean square error (MSE):

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,\epsilon,t}[\| \epsilon - \epsilon_\theta\ (x_t, t)\|^2] \tag{4}$$

3.2 Framework Overview

The proposed Dual Consistency Training (DCT) framework aims to address critical challenges in text-to-video generation, namely maintaining semantic alignment with the input text and ensuring temporal continuity of visual textures across frames. To comprehensively tackle these issues, DCT constructs an end-to-end pipeline comprising three core components. First, a multi-scale spatial adapter processes input features at multiple resolutions to capture both coarse and fine-grained spatial information. Next, the framework employs dual supervision signals derived from distinct pre-trained networks: a semantic supervision branch based on CLIP enforces semantic consistency with the text, while a texture supervision branch based on VGG strengthens the preservation of fine visual details. These complementary losses guide the diffusion-based generation network to produce video frames that are temporally coherent and semantically faithful. Finally, the framework incorporates a dynamic loss weighting strategy (CLWS) that adaptively balances the contributions of semantic and texture constraints throughout training. This holistic design enables DCT to effectively learn from minimal data, achieving high-quality video generation with strong temporal consistency.

3.3 Network Architecture

Our model is built upon the pre-trained Stable Diffusion framework [1], and its overall architecture is illustrated in Fig. 1. During training, each action is represented by a video clip and its corresponding text prompt. Video frames are first encoded into latent feature representations X_i by the pre-trained encoder, then passed through the forward diffusion process where noise is gradually added to simulate feature degradation. The backward diffusion process employs an extended U-Net network [25] to remove the noise and recover clean features. To capture temporal dependencies between video frames, we extend the ResNet modules [26] to 3D convolutional blocks and introduce a multi-scale spatial adapter to enhance feature expressiveness. Our current model primarily targets single-object scenarios, with plans to address challenges from multi-object interactions—such as increased scene complexity and temporal dependencies—by incorporating temporal attention mechanisms and instance segmentation in future work. At each reverse diffusion step, the denoised frames are fed into pre-trained CLIP and VGG models to extract semantic and texture features, respectively. During training, the denoising loss is jointly optimized with semantic consistency loss and texture consistency loss. Some results are shown in Fig. 2.

3.3.1 Multi-Scale Spatial Adapter

Large-scale text-image pre-trained T2I models demonstrate excellent performance in personalization generation [27] and image editing [28] tasks, with significant transferability. Therefore, we believe that using a lightweight fine-tuning method can effectively utilize the spatial information in video generation. Inspired by efficient fine-tuning techniques in NLP [29] and vision tasks, we chose the adapter method because it is simple to implement and efficient.

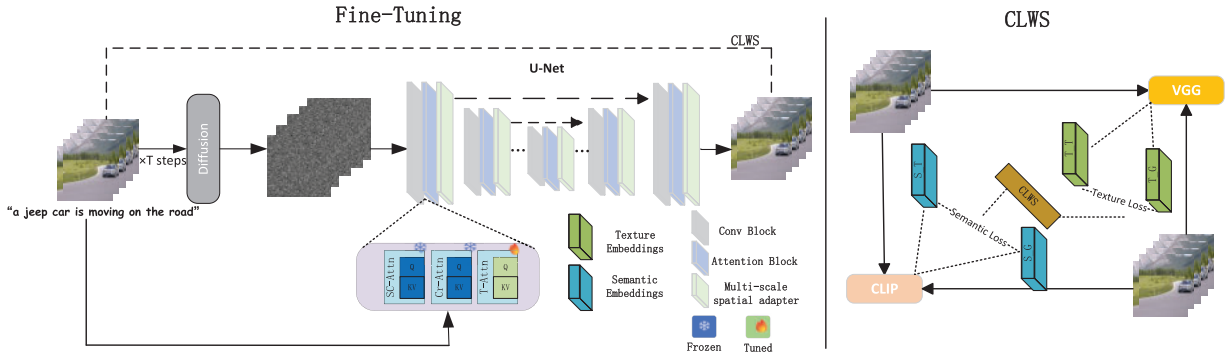

Figure 1: Pipeline of DCT: Given a text-video pair ("a jeep car is moving on the road"), the method leverages pretrained T2I diffusion models for T2V generation. It integrates three attention modules: **SC-Attn** models intra-frame dependencies; **Cr-Attn** aligns external semantics (e.g., CLIP features); and **T-Attn** captures temporal consistency across frames

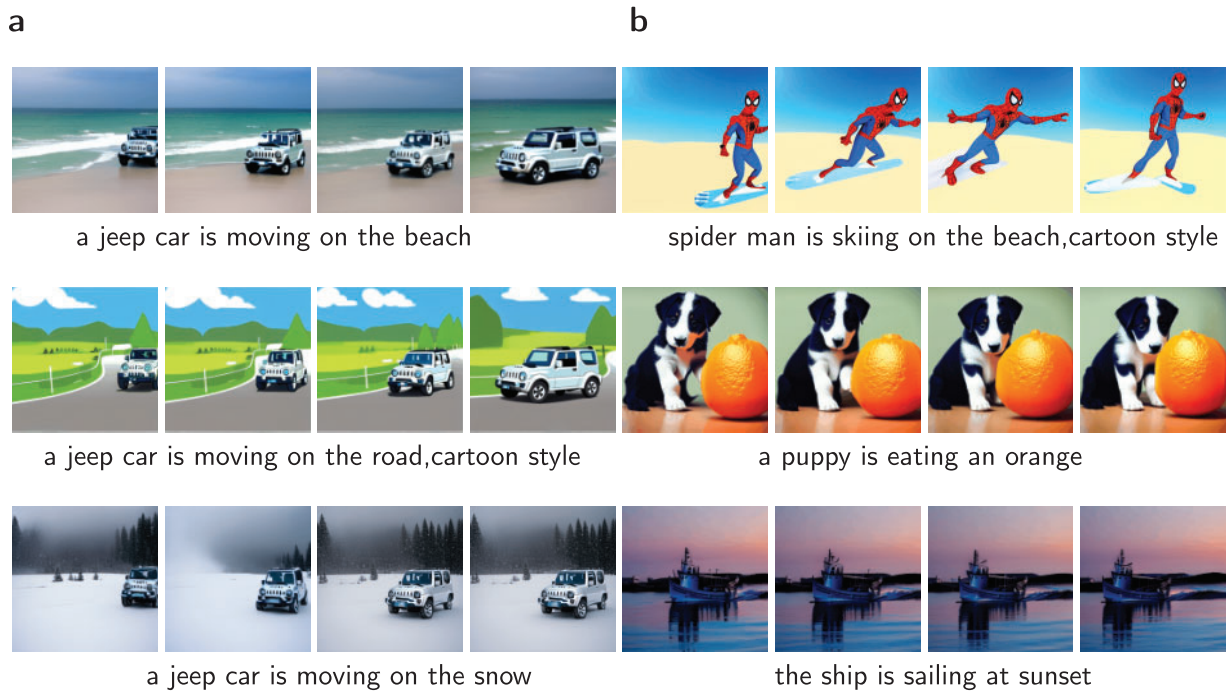

Figure 2: Sample results of our method. (**a**) Text-guided video editing results via one-shot text-video pair tuning. (**b**) Open-domain text-to-video generation results in the wild

In our DCT framework, the core objective of the multi-scale spatial adapter is to enhance the representation capability of the input features by fusing features at different scales. Specifically, as shown in Fig. 3, the adapter first uses two sets of fully connected layers for downsampling and upsampling, respectively, to extract the input features at different scales, denoted as $\hat{X}$, and then reconstructs them to the original input dimension. This approach not only enhances the multi-scale information of the features but also effectively captures both global and local spatial information. The multi-scale spatial adapter can be written as:

$$\text{MS-Adapter}(X) = X + W_{\text{up}}^{1}\left(\text{GELU}\left(W_{\text{down}}^{1}(X)\right)\right) + W_{\text{up}}^{2}\left(\text{GELU}\left(W_{\text{down}}^{2}(X)\right)\right) \tag{5}$$

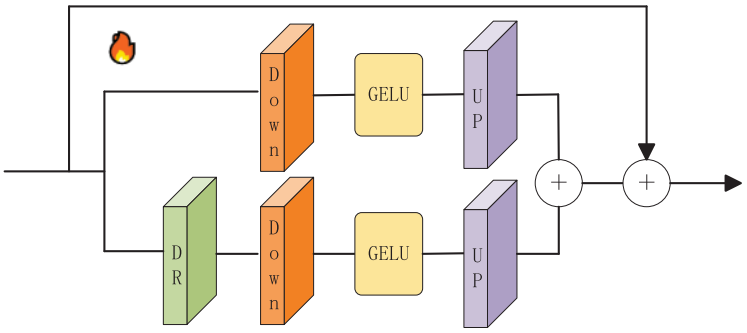

Figure 3: MS-Adapter. DR stands for dimensionality reduction, Down stands for downsampling, Up stands for upsampling, and the activation function is GELU

The downsampling and upsampling processes: The input feature matrix $X \in \mathbb{R}^{N \times d}$ goes through downsampling layers W^1_{down} and W^2_{down}, as well as upsampling layers W^1_{up} and W^2_{up}, to perform feature extraction and reconstruction at different scales. The downsampling operation helps the model encode input information at various scales, while the upsampling operation aids in restoring and enhancing higher-level feature information.

Interaction of multi-scale features: In diffusion models, the input data undergoes several rounds of denoising. The role of the multi-scale spatial adapter is to enrich the diffusion process by adding features from different scales to the original input features during each generation round. Specifically, this fusion of multi-scale features helps the diffusion model retain global structure while capturing local details when processing fine-grained spatial information.

This module adopts a lightweight design, inserting only a small number of parameters into the pre-trained model for fine-tuning, thereby avoiding large-scale adjustments to the main model and significantly reducing parameter updates and computational overhead. Meanwhile, the multi-scale spatial adapter efficiently integrates features at different scales, enhancing spatial information representation and promoting faster training convergence. Its simple structure and efficient computation further reduce resource consumption and time costs. Overall, this design effectively improves training efficiency and shortens training time while maintaining performance.

3.3.2 Semantic-Texture Coherence

Semantic-texture consistency is a key goal in video generation, ensuring that the generated content is both semantically and visually aligned with the target. We use the CLIP model [5] to extract semantic features from images and texts, optimizing the semantic distance between the generated and target frames using mean square error (MSE). Texture consistency is crucial for visual quality and detail preservation, especially in dynamic scenes. To address this, we use the VGG network [30] to extract multi-level texture features, ensuring visual consistency and improving detail representation by comparing texture features between the generated and target frames.

3.3.3 Consistency Loss Weight Scheduler (CLWS)

In computer vision, images comprise both spatial information and rich frequency-domain features [31]. Conventional weighting schemes, such as Gaussian or exponential functions, employ fixed weights that lack adaptability to frequency characteristics, limiting their ability to capture structural and detailed information in complex scenarios. The Fourier transform decomposes images into low- and high-frequency components,

where low frequencies represent global structure and high frequencies encode textures and fine details [32]. Leveraging the Fourier spectrum allows dynamic adjustment of semantic and texture loss weights, enabling adaptive balance between structural and detailed features during training. This content-aware mechanism outperforms fixed-weight methods, enhancing model stability and performance in complex visual tasks.

1) Multi-Scale Fourier Transform.

To capture frequency features at multiple levels, we apply Fourier transforms to image representations downsampled at different scales s. For a scaled image $I_s(x, y)$, the frequency representation is computed as:

$$F_s(u, v) = \int_{-\infty}^{\infty} I_s(x, y) e^{-j2\pi(ux+vy)} dxdy \tag{6}$$

where $I_s(x, y)$ is the downsampled version of image $I(x, y)$ at scale s. By calculating the frequency domain transform at multiple scales, information at different frequency levels can be captured.

2) Combining Amplitude and Phase.

Amplitude and phase information are essential for image reconstruction. The loss function integrates both amplitude and phase spectra, enabling optimization based on energy distribution (amplitude) and spatial structure (phase). The Fourier transform decomposes a signal into its amplitude spectrum $|F(u, v)|$ and phase spectrum $\angle F(u, v)$. This separation is expressed as:

$$F(u, v) = |F(u, v)| e^{j \angle F(u,v)} \tag{7}$$

3) Adaptive Weight Adjustment.

According to the frequency domain features, especially the energy distribution of the spectrum, we dynamically adjust the weights of the semantic loss and texture loss in the loss function through the frequency domain energy distributions $E_{semantic}$ and $E_{texture}$.

$$E_x = \sum_{u,v} |F_x(u, v)|^2, \quad x \in \{\text{semantic}, \text{texture}\} \tag{8}$$

where $F_{semantic}(u, v)$ and $F_{texture}(u, v)$ represent the Fourier transform results of the semantic image and texture image.

Based on these energy distributions, an adaptive weight α is defined:

$$\alpha = \frac{E_{semantic}}{E_{semantic} + E_{texture} + \epsilon} \tag{9}$$

4) Consistency Loss.

To ensure the generated video frames remain consistent with the target in both semantic content and fine-grained details, we define two frequency-domain loss components: semantic loss $L_{semantic}$ and texture loss $L_{texture}$. These losses are computed by measuring the difference between the Fourier transforms of the predicted outputs and the corresponding ground-truth frames. Semantic and texture losses jointly measure the frequency-domain consistency between the generated output and the target frame. Specifically, the loss is defined as:

$$L_x = \sum_{u,v} \left| F_x(u, v) - F_{\text{target}}(u, v) \right|^2, \quad x \in \{\text{semantic}, \text{texture}\} \tag{10}$$

Here, $F_x(u, v)$ represents the frequency-domain representation of either semantic or texture features extracted from the generated frame, and $F_{target}(u, v)$ is the corresponding target spectrum. The semantic

loss emphasizes global content alignment, while the texture loss focuses on preserving local structural details such as edges and fine textures. These objectives are complementary and are later combined into a unified consistency loss. To effectively integrate both objectives, we formulate a unified consistency loss as a weighted sum:

$$L_{\text{consistency}} = \alpha L_{\text{semantic}} + (1-\alpha) L_{\text{texture}} \tag{11}$$

The weight $\alpha \in [0,1]$ is adaptively computed based on the relative spectral energy of the semantic and texture components. This dynamic weighting allows the model to automatically emphasize different types of consistency during training, depending on the current characteristics of the generated output. As training progresses, the model balances structural alignment and detail synthesis, enhancing semantic fidelity and visual realism in generated frames.

4 Experiments

4.1 Implementations

Our experiments are based on the latent diffusion model, using a single motion video and a text description of the video for training. We uniformly sample 24 frames from the video at a resolution of 512×512 and perform 500 steps of fine-tuning with a learning rate of 3×10^{-5}, a batch size of 1, and a random seed of 33. Mixed-precision training (FP16) is enabled to reduce memory usage and accelerate training. During inference, we use the DDIM sampler and do not use classifier guidance. Fine-tuning a single video takes approximately 34 minutes, with all experiments performed on a single A6000 GPU, consuming 15GB of VRAM during training and 8 GB during inference. A brief scalability test shows that training time increases approximately linearly with frame count and quadratically with resolution; e.g., for 48 frames at 768×768, training took about 75 minutes. Mixed-precision training helps maintain practical efficiency.

Datasets: To evaluate our method, we selected 25 videos from the DAVIS dataset, a widely used benchmark in video analysis known for its high quality and diverse, well-annotated content. Although originally designed for object segmentation, DAVIS offers rich motion patterns and complex scenes, making it suitable for text-driven video generation. We chose videos from three representative categories: animals, vehicles, and humans. For each video, we created four prompts involving object editing, background modification, and style transformation, resulting in 100 evaluation samples.

4.2 Comparisons

During DCT training, we randomly select a video from the corresponding set for each motion type to enhance diversity and improve generalization. We then compare our method with five publicly available baselines: 1) CogVideo [33]: A zero-shot T2V model that generates high-quality videos from text prompts. 2) Plug-and-Play [34]: A frame-by-frame image editing method that enhances content flexibility. 3) Text2LIVE [35]: A text-driven video editing approach based on hierarchical neural graphs. 4) AnimateDiff [36]: Generates continuous animations from text, improving image-to-video translation. 5) Text2Video-Zero [37]: A zero-shot video generation model that operates without large-scale training data.

Quantitative results. We evaluate our method against five baseline models based on text alignment, inter-frame continuity, and generation diversity, as shown in Table 1.

Table 1: Quantitative comparison with text-to-video methods

Method	Alignment	Consistency	Diversity
CogVideo [33]	26.2227	94.2312	80.6236
Plug-and-Play [34]	26.9424	94.7236	83.0136
Text2LIVE [35]	28.5779	93.1931	79.4723
AnimateDiff [36]	27.0168	90.9635	81.6238
T2V-Zero [37]	28.0254	91.2359	78.1596
DCT (Ours)	28.5874	94.8742	79.1256

We evaluate video-text alignment using the average CLIP score per frame [11], where higher scores indicate better semantic consistency. Frame consistency is measured by the cosine similarity between CLIP embeddings of adjacent frames, reflecting temporal smoothness. Video diversity is assessed via the cosine distance between video embeddings, with smaller distances indicating lower diversity. Experimental results show that CogVideo achieves strong temporal consistency but lacks semantic richness due to limited text understanding. Plug-and-Play ensures smooth transitions but offers limited content diversity due to its frame-by-frame editing. Text2LIVE captures semantics well but struggles with fluency and detail. AnimateDiff lacks temporal coherence, affecting visual continuity, while T2V-Zero aligns well with text but still falls short in overall quality. In contrast, our method outperforms all baselines in both semantic consistency and temporal smoothness.

We use PSNR and optical flow consistency to evaluate the CLWS strategy, as shown in Table 2. Mul-Fourier outperforms other algorithms with fast convergence and high-quality output. Gaussian performs poorly in motion consistency and quality, while Fourier and Exponential are stable but slightly less effective than Mul-Fourier.

Table 2: CLWS method is quantitatively compared, PSNR represents peak signal-to-noise ratio, and OF-Consistency represents optical flow consistency

Method	PSNR	OF-Consistency
Exponential	31.54	2.85
Gaussian	31.42	2.74
Fourier	31.61	2.95
Mul-Fourier	31.70	2.67

Qualitative results. Fig. 4 presents a comparison for the prompt "Spider-Man skiing on the beach, cartoon style." CogVideo fails to capture the text concept and yields low-quality results; Plug-and-Play lacks temporal coherence; Text2LIVE struggles with accurate content expression; AnimateDiff shows limited performance; heterogeneous T2I models offer consistency and diversity but lack texture alignment; and T2V-Zero, while visually appealing, lacks realistic motion. In contrast, our method achieves superior temporal consistency, semantic alignment, and overall video quality.

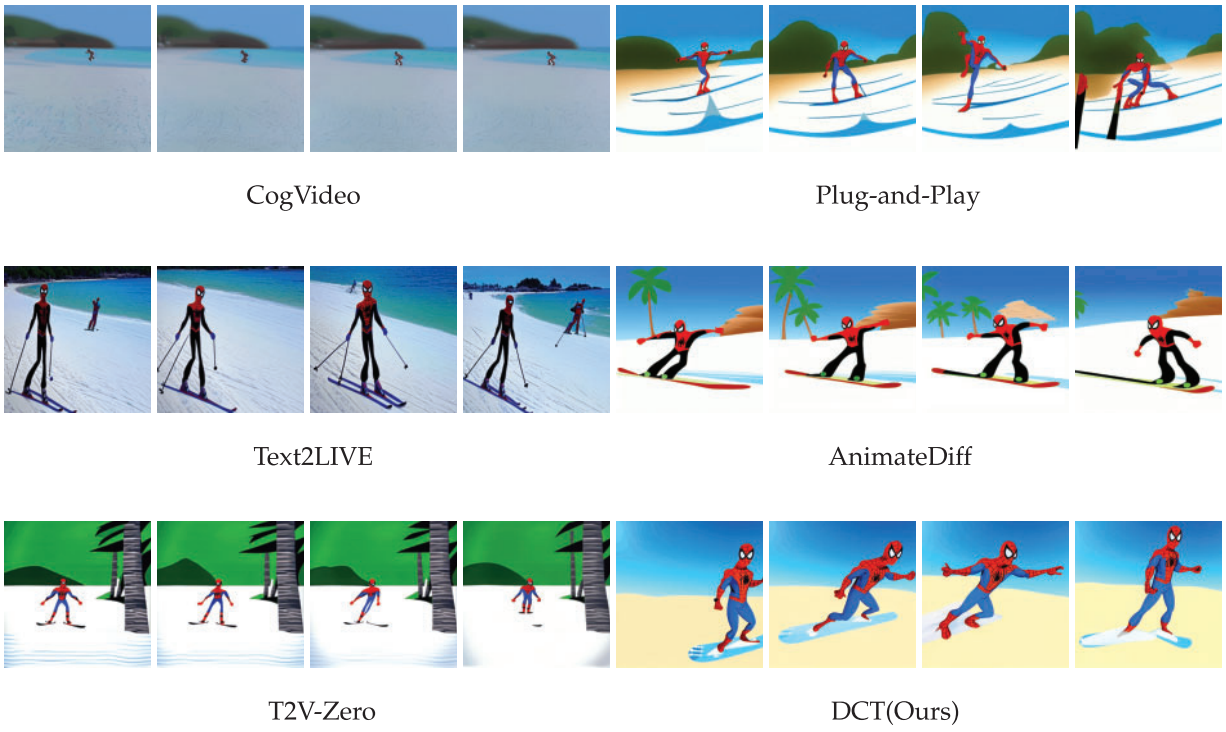

Figure 4: Qualitative comparison between the proposed DCT and five baselines. **Zoom in for the best view**

As shown in Fig. 5, the four weighting strategies yield distinct results in the vehicle region. Gaussian weighting overly smooths the image, blurring edges and losing high-frequency details. Exponential weighting improves smoothness but still distorts the vehicle's structure. The single Fourier method enhances some frequencies but lacks edge clarity. In contrast, the Mul-Fourier strategy effectively fuses multi-scale frequency information, producing sharper edges, richer textures, and better structural consistency.

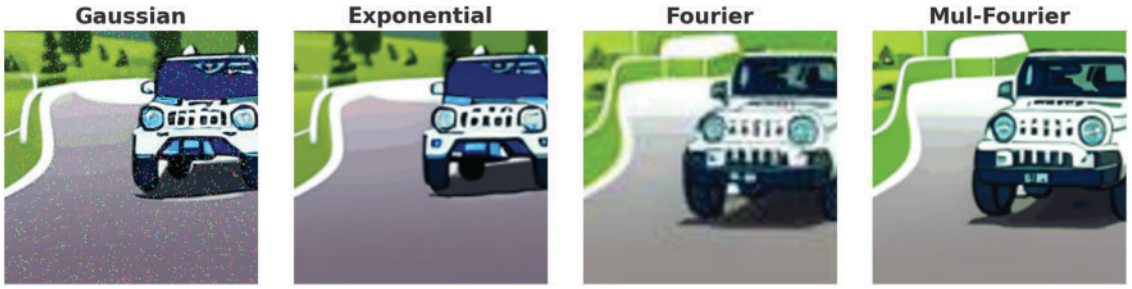

Figure 5: Qualitative comparison of CLWS

4.3 Ablation Study

We conducted an ablation study using the text prompt "a puppy is eating an orange," as shown in Fig. 6. In Fig. 6b, although detail and complexity have improved, the theme is unclear and the image appears cluttered. Fig. 6c suffers from unnatural detail handling and lacks clarity, theme focus, and compositional balance. Similarly, Fig. 6d shows weak theme prominence and poor naturalness. These results highlight the

essential role of each module in enhancing image quality. As shown in Table 3, all three metrics improve significantly with the full model.

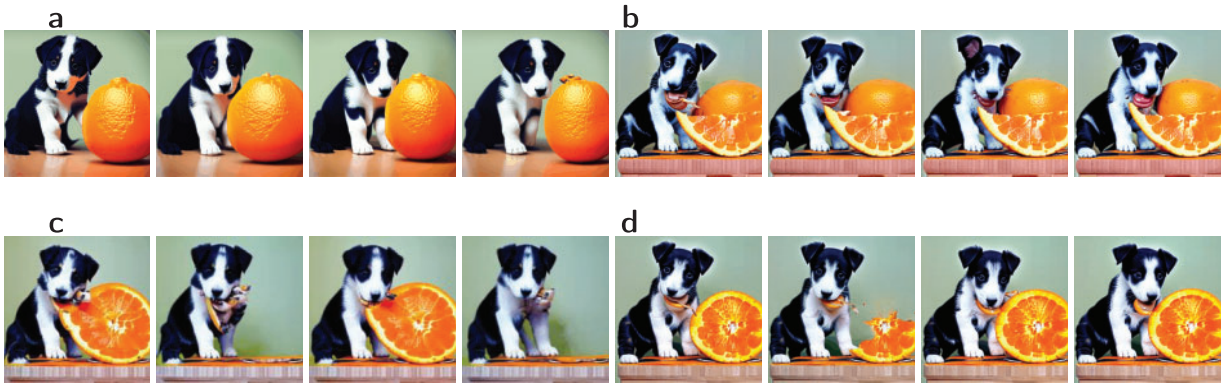

Figure 6: Ablation experiment. The text prompt is "a puppy is eating an orange". (**a**) corresponds to the full DCT model, (**b**) represents the model variant without semantic consistency, (**c**) indicates the model variant without texture consistency, (**d**) indicates the model variant without the multi-scale spatial adapter module

Table 3: Ablation study. MS: Multi-Scale Spatial Adapter, S: Semantic Consistency, T: Texture Consistency

Method	MS	S	T	Alignment	Consistency	FVD	PSNR	OF-Consistency
Baseline	×	×	×	27.4986	93.4018	230.53	29.78	2.87
MS	✓	×	×	27.6512	93.5862	227.44	29.83	2.81
S	×	✓	×	28.5069	94.2653	224.92	30.93	2.75
T	×	×	✓	28.1234	94.1234	225.67	30.51	2.77
MS + S	✓	✓	×	28.5167	94.5001	223.21	31.25	2.72
MS + T	✓	×	✓	28.3543	94.2345	222.19	30.98	2.75
S + T	×	✓	✓	28.5582	94.6992	220.13	31.57	2.70
DCT (Ours)	✓	✓	✓	28.5874	94.8742	219.47	31.70	2.67

As shown in Fig. 7, without the texture consistency module (left), the snow and sky appear blurred and lack detail. Without the semantic consistency module (middle), the pink outfit deviates from the text prompt. In contrast, the full model (right) preserves both texture and semantics, producing more natural and coherent results.

Figure 7: Qualitative comparison under different consistency settings, the text prompt is "a man, wearing yellow clothes, is skiing"

Table 4 shows that while ViT and BLIP improve consistency scores, they incur substantial memory and training time overheads. In contrast, DCT achieves comparable performance with significantly lower memory usage and greater training efficiency. Thus, considering both effectiveness and efficiency, DCT represents a more practical choice.

Table 4: Comparison of different vision model methods

Method	Alignment	Consistency	Memory (Gib)	Training (min)
Baseline	27.4986	93.4018	16	49
Baseline+ViT	27.3451	93.4251	22	37
Baseline+BLIP	29.7753	95.6329	27	105
DCT (Ours)	28.5874	94.8742	15	34

5 Conclusion

This paper presents Dual Consistency Training (DCT), a text-driven video generation method using single-video fine-tuning. It combines multi-scale spatial adapters with CLIP-based semantic and VGG-based texture features. A progressive training scheme and dynamic loss balancing (CLWS) enhance frame quality and consistency. Experiments confirm improved text alignment, temporal coherence, and generalization, advancing text-to-video generation research.

6 Limitations and Future Work

While the model performs well on static backgrounds and single-object scenes, it struggles with complex multi-object interactions and dynamic motions (see Fig. 8). This limitation likely stems from text-to-image models' inherent difficulty in modeling multiple interacting objects. Future work may improve this by integrating finer temporal modeling, such as temporal attention, and incorporating instance segmentation to better distinguish and manage overlapping or occluded objects, thereby enhancing performance in multi-object scenarios.

Figure 8: In this high-motion multi-object scene, some horses exhibit deformation, overlap, or blurring, indicating limitations in structural preservation and temporal consistency

7 Ethical Considerations

Robust ethical safeguards are essential to prevent misuse of video generation technologies like deepfakes. Combining resilient digital watermarking with multimodal metadata ensures traceability and authenticity. Multimodal detection frameworks integrating visual, auditory, and textual cues, supported by

adversarial and self-supervised learning, improve detection accuracy and generalization. Privacy-preserving federated learning enables secure collaborative training across institutions. Establishing a comprehensive content authentication and governance system is vital for responsible deployment of text-to-video generation technologies.

Acknowledgement: We sincerely thank the editors and reviewers for their valuable feedback, which greatly improved this paper.

Funding Statement: This work is supported in part by the National Natural Science Foundation of China [Grant number 62471075], the Major Science and Technology Project Grant of the Chongqing Municipal Education Commission [Grant number KJZD-M202301901]. Graduate Innovation Project Funding of Chongqing University of Technology [Grant number gzlcx20253249].

Author Contributions: Xian Yu designed the methodology, developed network modules, implemented the code, and drafted the manuscript. Jianxun Zhang supervised the work and contributed to theoretical analysis. Siran Tian analyzed the dataset and optimized inference. Xiaobao He revised the manuscript and conducted literature review. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository, code is available at https://github.com/Powder-lab/DCT, accessed on 29 April 2025.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA; 2022. p. 10684–95. doi:10.1109/CVPR52688.2022.01042.
2. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44. doi:10.1145/3422622.
3. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. Glide: towards photorealistic image generation and editing with text-guided diffusion models. arXiv:2112.10741. 2021.
4. Xing Z, Feng Q, Chen H, Dai Q, Hu H, Xu H, et al. A survey on video diffusion models. ACM Comput Surv. 2024;57(2):1–42. doi:10.1145/3696415.
5. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125. 2022.
6. Tewel Y, Shalev Y, Schwartz I, Wolf L. ZeroCap: zero-shot image-to-text generation for visual-semantic arithmetic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); New Orleans, LA, USA; 2022. p. 17897–907. doi:10.1109/cvpr52688.2022.01739.
7. Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ. Video diffusion models. Adv Neural Inf Process Syst. 2022;35:8633–46.
8. Zhang S, Wang J, Zhang Y, Zhao K, Yuan H, Qin Z, et al. I2vgen-xl: high-quality image-to-video synthesis via cascaded diffusion models. arXiv:2311.04145. 2023.
9. Kim G, Shim H, Kim H, Choi Y, Kim J, Yang E. Diffusion video autoencoders: toward temporally consistent face video editing via disentangled video encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Vancouver, BC, Canada; 2023. p. 6091–100. doi:10.1109/cvpr52729.2023.00590.
10. Tian F, Du S, Duan Y. MonoNeRF: learning a generalizable dynamic radiance field from monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); Paris, France; 2023. p. 17857–67. doi:10.1109/iccv51070.2023.01641.

11. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning; 2021. p. 8748–63.
12. Tulyakov S, Liu MY, Yang X, Kautz J. MoCoGAN: decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, USA; 2018. p. 1526–35. doi:10.1109/cvpr.2018.00165.
13. Saito M, Matsumoto E, Saito S. Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); Venice, Italy; 2017. doi:10.1109/iccv.2017.308.
14. Wu JZ, Ge Y, Wang X, Lei SW, Gu Y, Shi Y, et al. Tune-a-video: one-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; Paris, France; 2023. p. 7623–33. doi:10.1109/iccv51070.2023.00701.
15. Razavi A, Van den Oord A, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2. Adv Neural Inf Process Syst. 2019;32:14866–76.
16. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, et al. Alias-free generative adversarial networks. Adv Neural Inf Process Syst. 2021;34:852–63.
17. Nie Y, Zelikman E, Scott A, Jain R, Zeng Y, Zhang Y, et al. SkyGPT: probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained VideoGPT. Adv Appl Energy. 2024;14:100172. doi:10.1016/j.adapen.2024.1001721.
18. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands: Springer; 2016. p. 694–711. doi:10.1007/978-3-319-46475-6_43.
19. Ruder M, Dosovitskiy A, Brox T. Artistic style transfer for videos. In: Pattern Recognition: 38th German Conference, GCPR 2016; 2016 Sep 12–15; Hannover, Germany: Springer; 2016. p. 26–36. doi:10.1007/978-3-319-45886-1_3.
20. Chu M, Xie Y, Mayer J, Leal-Taixé L, Thuerey N. Learning temporal coherence via self-supervision for GAN-based video generation. ACM Trans Graph. 2020;39(4):75. doi:10.1145/3386569.3392457.
21. Narasimhan M, Ginosar S, Owens A, Efros AA, Darrell T. Strumming to the beat: audio-conditioned contrastive video textures. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); Waikoloa, HI, USA; 2022. p. 507–16. doi:10.1109/WACV51458.2022.00058.
22. Li Z, Tucker R, Snavely N, Holynski A. Generative image dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Seattle, WA, USA; 2024. p. 24142–53. doi:10.1109/cvpr52733.2024.02279.
23. Mao J, Huang X, Xie Y, Chang Y, Hui M, Xu B, et al. StoryAdapter: a training-free iterative framework for long story visualization. arXiv:2410.06244. 2024.
24. Capel EH, Dumas J. Denoising diffusion probabilistic models for probabilistic energy forecasting. In: Proceedings of the 2023 IEEE Belgrade PowerTech; Belgrade, Serbia; 2023. p. 1–6. doi:10.1109/powertech55446.2023.10202713.
25. Daneshfar F, Saifee BS, Soleymanbaigi S, Aeini M. Elastic deep multi-view autoencoder with diversity embedding. Inf Sci. 2025;689:121482. doi:10.1016/j.ins.2024.121482.
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Las Vegas, NV, USA; 2016. p. 770–78. doi:10.1109/cvpr.2016.90.
27. Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K, et al. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Vancouver, BC, Canada; 2023. p. 22500–10. doi:10.1109/cvpr52729.2023.02155.
28. Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; Paris, France; 2023. p. 3836–47. doi:10.1109/iccv51070.2023.00355.
29. Ye H, Xie X, Xie F, Zuo J, Bu C. LoRA-Adv: boosting text classification in large language models through adversarial low-rank adaptations. IEEE Access. 2025;13:103234–44. doi:10.1109/access.2025.3579539.

30. Zhang H, Yang YJ, Zeng W. Self-supervised multi-scale semantic consistency regularization for unsupervised image-to-image translation. Comput Vis Image Underst. 2024;241:103950. doi:10.1016/j.cviu.2024.103950.
31. Jiang L, Dai B, Wu W, Loy CC. Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; Montreal, QC, Canada; 2021. p. 13919–29. doi:10.1109/iccv48922.2021.01366.
32. Huang Z, Zhang Z, Lan C, Zha ZJ, Lu Y, Guo B. Adaptive frequency filters as efficient global token mixers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; Paris, France; 2023. p. 6049–59. doi:10.1109/iccv51070.2023.00556.
33. Hong W, Ding M, Zheng W, Liu X, Tang J. CogVideo: large-scale pretraining for text-to-video generation via transformers. arXiv:2205.15868. 2022.
34. Tumanyan N, Geyer M, Bagon S, Dekel T. Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Vancouver, BC, Canada; 2023. p. 1921–30. doi:10.1109/cvpr52729.2023.00191.
35. Bar-Tal O, Ofri-Amar D, Fridman R, Dekel T, Rubinstein M, Shamir A. Text2Live: text-driven layered image and video editing. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland; 2022. p. 707–23. doi:10.1007/978-3-031-19784-0_41.
36. Guo Y, Yang C, Rao A, Liang Z, Wang Y, Qiao Y, et al. AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning. arXiv:2307.04725. 2023.
37. Khachatryan L, Movsisyan A, Tadevosyan V, Henschel R, Wang Z, Navasardyan S, et al. Text2Video-Zero: text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; Paris, France; 2023. p. 15908–18. doi:10.1109/iccv51070.2023.01462.