CoSeC-LCD: Controllable Self-Contrastive Latent Consistency Distillation for Better and Faster Human Animation Generation

Anonymous authors
Paper under double-blind review

Figure 1: Showcase of performance across various styles. Our method achieves comparable performance to the teacher model in just 4 steps, while also exhibiting advantages in fine-grained detail control. Notably, in example **f**, our method better preserves the reference image's overall style.

Abstract

Generating pose-driven and reference-consistent human animation has significant practical applications, yet it remains a prominent research challenge, facing substantial obstacles. A major issue with widely adopted diffusion-based methods is their slow generation speed, which is primarily due to multi-step iterative denoising processes. To tackle this challenge, we take the pioneering step of proposing the ReferenceLCM architecture, which utilizes latent consistency models (LCM) to facilitate accelerated generation. Additionally, to address hallucinations in fine-grained control, we introduce the Controllable Self-Contrastive Latent Consistency Distillation (CoSeC-LCD) regularization method. Our approach introduces a novel perspective by categorizing tasks into various classes and employing contrastive learning to capture underlying patterns. Building on this insight, we implement a hierarchical optimization strategy that significantly enhances animation quality across both spatial and temporal aspects. Comprehensive qualitative and quantitative experiments reveal that our method achieves results comparable to, or even surpassing, many state-of-the-art approaches, enabling high-fidelity human animation generation within just 2-4 inference steps.

1 INTRODUCTION

Generating videos that closely align with a reference image of a given person using pose-guidance inputs, such as depth map representations (Jeon et al. (2015)), has significant implications for digital human and virtual reality applications (da Silva et al. (2022), Wohlgenannt et al. (2020)). This topic has garnered increasing research interest in recent years (Karras et al. (2023); Wang et al. (2023a); Hu (2024); Xu et al. (2024b); Zhu et al. (2024)). However, existing methods encounter challenges in achieving both high efficiency and output quality, highlighting the need for further advancements to enable a real-time, high-fidelity video synthesis framework.

Generating videos from reference images presents more stringent demands than text-driven tasks (Loeschcke et al. (2022); Jiang et al. (2023)), particularly in preserving intricate details and stylistic consistency while accurately reproducing complex motion sequences (Chan et al. (2019); Siarohin et al. (2019b)). This requires precise modulation of extensive semantic information, challenging current generative models. Diffusion models (Song et al. (2020b); Ho et al. (2020)) excel in this area, with the **ReferenceNet** architecture demonstrating remarkable effectiveness in producing high-quality, temporally consistent videos for applications like facial expression (Tian et al. (2024); Xu et al. (2024a)) and dance generation (Hu (2024); Zhu et al. (2024)). It effectively encodes complex semantic details from reference images into a coherent latent space, enhancing fine detail preservation. However, current methods face challenges with slow generation speeds and high computational costs due to the iterative denoising process in diffusion models (Watson et al. (2022); Lu et al. (2022)), which hinders their suitability for real-time applications. Thus, developing a computationally efficient ReferenceNet that maintains comparable performance is essential.

The Consistency Model (CM) is an advanced generative framework capable of producing high-quality images in just a few steps, significantly reducing computational complexity compared to traditional diffusion models (Song et al. (2023)). Building on this, Luo et al. extends CM into the latent space (Kingma & Welling (2014)), establishing a foundation for efficient high-resolution image generation. In the realm of video generation, the Latent Consistency Model (LCM) has been applied by Wang et al. (2023b), demonstrating its potential for enhanced temporal coherence. LCM has also achieved significant acceleration across diverse tasks, such as motion generation (Dai et al. (2024)) and audio synthesis (Liu et al.), underscoring its versatility and efficiency improvements. However, these advancements have primarily focused on text-driven tasks, emphasizing general alignment with prompts rather than the high-precision control required for controllable and consistent generation. This leaves a gap in research on acceleration algorithms tailored for such tasks. Moreover, the performance of the Latent Consistency Model (LCM) itself has significant room for optimization, which constitutes another limitation.

In light of the current situation, to efficiently generate high-quality, controllable, and consistent videos, we have made a series of innovative advancements, addressing both speed and quality. **First**, we introduce the **ReferenceLCM** architecture, the first known consistency distillation framework that combines the robust control capabilities of ReferenceNet with the significant acceleration advantages of LCM, facilitating efficient and high-fidelity video synthesis. **Besides**, to further enhance the performance of the ReferenceLCM framework, we designed a hierarchical **Co**ntrollable **Se**lf-**C**ontrastive **L**atent **C**onsistency **D**istillation (**CoSeC-LCD**) regularization. Specifically, drawing from insights in contrastive learning, we innovatively construct different categories based on the significant semantic differences between intra-source and inter-source videos. By optimizing the distance relationships between generated samples across these categories, the model can better understand the underlying patterns in the generation tasks. We introduced **E**quivalent **T**arget **A**ggregation (**ETA**) to ensure the cohesion among generated equivalent samples and **Co**ntrastive **N**egative **S**ampling (**CoNS**) to enhance the distinction among inter-source samples. This approach collectively optimizes the sample distribution (Park et al. (2019a); Cui et al. (2021)), thereby increasing the model's confidence in generation targets and ultimately achieving higher quality and efficient video generation. We performed hierarchical optimization from two aspects: spatial quality, referring to the visual quality of video frames, and temporal consistency, addressing the overall coherence of the video. A detailed discussion of this approach can be found in section 2.2. Extensive qualitative and quantitative experiments demonstrate that our proposed method achieves comparable or superior results compared to various state-of-the-art (SOTA) methods, all while achieving acceleration factors of **10 to 50** times.

Our contributions can be summarized as follows:

- We introduced the ReferenceLCM architecture, which substantially reduces denoising steps and surpasses the speed bottlenecks of traditional ReferenceNet-based methods.
- We leveraged the semantic features of intra-source and inter-source videos from a novel perspective, further optimizing the performance of the LCD through the perspective of contrastive learning. This approach offers new insights into accelerating high-quality, controllable, and consistent video generation.
- Extensive experiments demonstrate that our method maintains results comparable to state-of-the-art models while achieving significantly high generation efficiency.

2 METHOD

In this section, we will first propose the ReferenceLCM architecture for efficient generation, which will be detailed in Section 2.1. To further enhance video generation quality, we introduce the CoSeC-LCD hierarchical regularization method, extending the ReferenceLCM framework, in Section 2.2.

2.1 REFERENCELCM

To enhance efficiency in controllable and consistent generation, we introduce the ReferenceLCM architecture, which combines ReferenceNet's strong detail control with LCM's high-efficiency acceleration capabilities. We decompose the distillation process into two phases: the first emphasizes accelerating the generation of high-quality video frames, while the second focuses on improving temporal coherence (Hu (2024); Wang et al. (2023b)). For the teacher model, we selected a leading state-of-the-art controllable consistent generation model Zhu et al. (2024).

Overall Previous works like Wang et al. (2023b); Li et al. (2024) have not explored the integration of LCD into the ReferenceNet architecture. We pioneer this integration to enhance ReferenceNet's generation speed. Since ReferenceNet is a dual-core U-Net (Ronneberger et al. (2015)) architecture that includes both Denoising UNet (D-UNet) and Reference UNet (R-UNet), applying LCD to the entire model would introduce significant computational overhead. To address this, we propose a reusable, lightweight architecture within the distillation pipeline. Specifically, we decouple ReferenceNet: the teacher, target, and student networks share the same weight initialization, forming the D-UNet group, where only the student D-UNet is trainable. The target-student updates are performed using Exponential Moving Average (EMA) Hunter (1986). R-UNet, serving as a conditional input module similar to CLIP (Radford et al. (2021)), supplies consistent inputs to the D-UNet group, thereby facilitating reusability. The overall architecture is illustrated in Figure 2.

Training We denote the R-UNet and D-UNet as $\mathcal{F}^{\mathcal{R}}$ and $\mathcal{F}^{\mathcal{D}}_{\theta}$, respectively. Following the framework of Champ (Zhu et al. (2024)), the guidance encoder group is denoted as $\mathcal{E}^{\mathcal{G}}$. The target sequence is represented as $\{x_{0,i}\}^{1:f}$, where f denotes the length of the video segment. The pose-guidance sequence extracted from the Skinned Multi-Person Linear (SMPL) (Loper et al. (2023)) model includes semantic maps, depth maps, normal maps, and skeleton maps, denoted as $\{x^{smt}_{0,i}, x^{dpt}_{0,i}, x^{nml}_{0,i}, x^{skl}_{0,i}\}^{1:f}$, respectively. The pose-guidance condition is represented as:

$$\mathbf{c}^{p}_{i} = \mathcal{E}^{\mathcal{G}}(x^{dpt}_{0,i}) \oplus \mathcal{E}^{\mathcal{G}}(x^{smt}_{0,i}) \oplus \mathcal{E}^{\mathcal{G}}(x^{skt}_{0,i}) \oplus \mathcal{E}^{\mathcal{G}}(x^{nml}_{0,i}), \tag{1}$$

where $\oplus$ denotes the feature fusion operator. The reference image is denoted as $\mathcal{I}$. We consider the attention weights transferred from R-UNet to D-UNet as input conditions, represented as $\mathcal{F}^{\mathcal{R}}(\mathcal{I})$. The CLIP embedding of the reference image is expressed as $\mathbf{c}^{\mathcal{I}}$, serving as the conditional input. $\mathbf{x}_{t,i}$ represents the noisy latent space input encoded by the Variational Autoencoder (VAE) (Kingma & Welling (2014)) at the t^{th} timestep of the i^{th} frame in the target video sequence. Consequently,

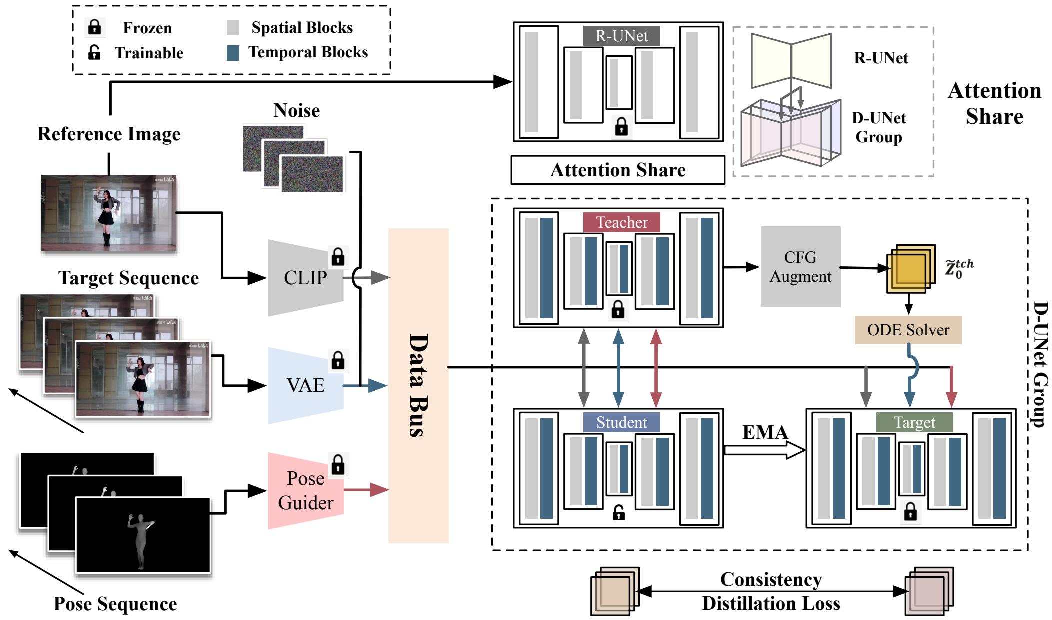

Figure 2: Overview of the ReferenceLCM architecture: We decouple the traditional ReferenceNet by utilizing R-UNet as a shared conditional input. The attention weights from R-UNet are distributed across three distinct D-UNet modules, where the consistency distillation process takes place. As detailed in 2.2, the loss function of ReferenceLCM works with **CoSeC** regularization to **further** enhance video generation quality.

the output of D-UNet can be formulated as $\mathcal{F}_\theta^{\mathcal{D}}(\mathbf{x}_{t,i}, t, \mathbf{c}_i^p, \mathcal{F}^{\mathcal{R}}(\mathcal{I}), \mathbf{c}^{\mathcal{I}})$. Following the approach in (Luo et al.), the consistency D-UNet $f_\theta^{\mathcal{D}}$ can be formulated as:

$$f_\theta^{\mathcal{D}}(\mathbf{x}_{t,i}, \mathbf{c}_i^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}) = c_{\text{skip}}(t)\mathbf{x}_{t,i} + c_{\text{out}}(t)\mathcal{F}_\theta^{\mathcal{D}}(\mathbf{x}_{t,i}, t, \mathbf{c}_i^p, \mathcal{F}^{\mathcal{R}}(\mathcal{I}), \mathbf{c}^{\mathcal{I}}) \tag{2}$$

where $c_{\text{skip}}(t)$ and $c_{\text{out}}(t)$ are differentiable functions, which satisfies $c_{\text{skip}}(0) = 1$ and $c_{\text{out}}(0) = 0$, please refer to Song et al. (2023) for details. The teacher, target, and student models are denoted by $f_{\theta^{\text{tch}}}^{\mathcal{D}}, f_{\theta^{\text{tgt}}}^{\mathcal{D}}, f_{\theta^{\text{stu}}}^{\mathcal{D}}$, respectively. According to previous work Luo et al.; Wang et al. (2023b), the consistency distillation (CD) loss is given by:

$$\mathcal{L}_{\text{CD}}(\theta^{\text{stu}}, \theta^{\text{tgt}}, \Psi) = \mathbb{E}\left[\mathcal{D}\left(f_\theta^{\text{stu}}\left(\mathbf{x}_{t_{n+k}}, \mathbf{c}^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}\right), f_{\theta^-}^{\text{tgt}}\left(\hat{\mathbf{x}}_{t_n}^{\Psi,\omega}, \mathbf{c}^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}\right)\right)\right] \tag{3}$$

where $\mathcal{D}(\cdot, \cdot)$ denotes a distance metric, such as Huber loss Huber (1992). Using the PF-ODE solver Ψ (e.g., DDIM Song et al. (2020a)), an accurate estimation of $\mathbf{x}_t$ from $\mathbf{x}_{t_{n+k}}$ is obtained, denoted as $\hat{\mathbf{x}}_{t_n}^{\Psi,\omega}$, where ω adjusts the strength of the classifier-free guidance (Ho & Salimans (2021)). Details for this can be found in Appendix B.1. In the first training phase, $\mathbf{x}$ represents a single video frame, focusing on spatial quality; in the second phase, $\mathbf{x}$ represents a video sequence, emphasizing temporal coherence. For a detailed description of the training algorithm, refer to Appendix D.

2.2 CoSeC-LCD

While the ReferenceLCM architecture can generate videos in just a few steps with quality comparable to more computationally intensive methods, it has limitations. Specifically, it may experience hallucinations in fine-grained control (see Figure 6), and improvements are needed in temporal coherence (see Figure 5). To address these issues, we innovatively propose the CoSeC-LCD regularization method from the perspective of contrastive learning. Now we will start with the modeling of the problem, outline our motivations, and provide a detailed introduction to our method.

2.2.1 Problem Formulation

Building on previous outstanding works (Kuang et al. (2021); Lin et al. (2021)), we define reference frames from the same video as intra-source (**RAS**) references, while frames from different videos

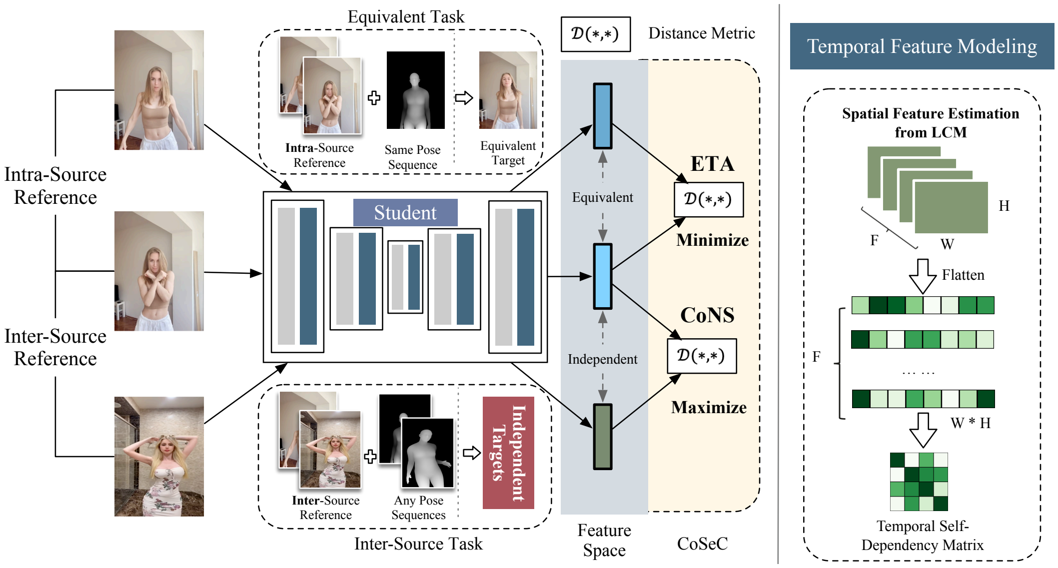

Figure 3: The architecture diagram of the self-contrastive regularization method, along with illustrations of equivalent tasks and inter-source tasks, is presented. We also demonstrate how to construct the self-dependency matrix.

are termed int**er-s**ource (**ERS**) references. Within a temporal window, RAS references exhibit a high degree of semantic similarity (Wray et al. (2021)). Unlike text-driven generation tasks, controllable consistency generation tasks yield relatively **deterministic** results. When the reference and pose guidance are given, the desired generation target is fundamentally determined. Additionally, considering the inherent semantic similarity of the RAS references, we define a target set guided by the same pose-guidance sequence and utilizing RAS references as **equivalent targets**. Otherwise, they are termed non-equivalent targets. Furthermore, when the reference images are ERS, we refer to them as **inter-source targets**, which exhibit significant semantic differences.

2.2.2 Self-Contrastive Regularization

Motivation Numerous outstanding works across various fields (Park et al. (2019b); Khosla et al. (2020); Mikolov et al. (2013)) have demonstrated the importance of the relative distance relationships of samples in feature space, which play a critical role in understanding the underlying patterns of a task (Liu et al. (2018)). Building on the previous definitions, we can apply a similar approach to enable the model to capture the underlying patterns of the task. Specifically, we can consider a set of equivalent targets as belonging to the same category, while the inter-source targets can be viewed as different categories. We aim to enhance the cohesion among generated samples related to equivalent targets while clearly preserving the distinction between generated samples under inter-source targets, thereby ensuring that the unique characteristics of each category are maintained.Since we use the generated features of the student model itself, we refer to this as **self-contrastive regularization**.

Equivalent Target Aggregation Formally, for a set of equivalent tasks $\{(\mathcal{I}^i, p)\}^{1:k}$, where $\{\mathcal{I}^i\}^{1:k}$ are k reference frames derived from the RAS video $\mathcal{V}$ and p represents the same pose guidance, the generated samples' features should be as consistent as possible. We utilize the sampled features , which can represent spatial or temporal features, as estimations. Our goal is to aggregate these features to minimize the distance between them as much as possible, we refer to this regularization as **E**quivalent **T**arget **A**ggregation (**ETA**). This aligns with intuitive reasoning: for a complex task, having greater overlap between results obtained from different perspectives (i.e., different references) generally leads to outcomes that are closer to the optimal solution (Wang et al. (2022)).

Contrastive Negative Sampling To avoid blurry images, see Figure 6, caused by focusing solely on minimizing distances, inspired by previous works (Chen et al. (2020b)), we introduce an innovative **Co**ntrastive **N**egative **S**ampling (**CoNS**) regularization. This regularization ensures that results from inter-source tasks maintaining an appropriate degree of separation. Specifically, for any two output features derived from inter-source tasks, we introduce a penalty term against the ETA. The overall architecture diagram of the CoSeC method is illustrated in Figure 3.

2.2.3 HIERARCHICAL OPTIMIZATION

Spatial Self-Contrastive Regularization In the first phase of ReferenceLCM training, we apply this regularization method, where both the input and output are single video frames. The generated output is denoted as $\hat{\mathbf{x}}_{0,\theta^{\text{stu}}}^{\mathcal{I},p} = f_\theta^{\text{stu}}\left(\mathbf{x}_{t_{n+k}}, \mathbf{c}^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}\right)$ as the spatial feature, with $\mathcal{I}$ as the reference image and p as the pose guidance. The parameter θ^{stu} represents the student model's parameters optimized in Section 2.1. The CM's notable ability to map any noisy latent $\mathbf{x}_t$ at timestep t back to the estimated original state $\hat{\mathbf{x}}_0$ allows us to use this mapped value for spatial feature estimation without iterative denoising. Since our focus is on the distribution of model outputs in latent space, there's no need to revert to pixel space. We define the regularization term $\mathcal{R}_{\text{spt}}^{\theta^{\text{stu}}}$ using $\hat{\mathbf{x}}_{0,\theta^{\text{stu}}}^{\mathcal{I},p}$ as:

$$\mathcal{R}_{\text{spt}}^{\theta^{\text{stu}}}(\phi_1, \phi_2) = \phi_1 \mathbb{E}_{\mathcal{I}_i, \mathcal{I}_j \sim \mathcal{V}_{\text{RAS}}}\left[\mathcal{D}(\hat{\mathbf{x}}_{0,\theta^{\text{stu}}}^{\mathcal{I}_i,p^\#}, \hat{\mathbf{x}}_{0,\theta^{\text{stu}}}^{\mathcal{I}_j,p^\#})\right] - \phi_2 \mathbb{E}_{\mathcal{I}_k, \mathcal{I}_l \sim \mathcal{V}_{\text{ERS}}}\left[\mathcal{D}(\hat{\mathbf{x}}_{0,\theta^{\text{stu}}}^{\mathcal{I}_k,p^*}, \hat{\mathbf{x}}_{0,\theta^{\text{stu}}}^{\mathcal{I}_l,p^*})\right], \quad (4)$$

where ϕ_1 and ϕ_2 are hyperparameters that adjust the weights of ETA and CoNS, respectively. Here, $\mathcal{I}_i$ and $\mathcal{I}_j$ denote any intra-source references, while $\mathcal{I}_k$ and $\mathcal{I}_l$ indicate inter-source references. The symbol $\#$ signifies the use of the same action guidance within equivalent tasks, and $*$ denotes a wildcard. $\mathcal{D}$ represents the distance metric. Therefore, the total training loss function for the first phase can be expressed as follows:

$$\mathcal{L}_1(\theta^{\text{stu}}, \theta^{\text{tgt}}, \Psi, \phi_1, \phi_2) = \mathcal{L}_{\text{CD}}(\theta^{\text{stu}}, \theta^{\text{tgt}}, \Psi) + \mathcal{R}_{\text{spt}}^{\theta^{\text{stu}}}(\phi_1, \phi_2). \quad (5)$$

Temporal Self-Contrastive Regularization We propose temporal self-contrastive regularization in the the second phase of ReferenceLCM training for smoother video generation, leveraging the temporal dependencies of video frames (Zhou et al. (2018)). We use a self-dependency matrix (Jeong et al. (2024)) to quantify frame changes. The Temporal Self-Dependency Matrix $\mathcal{T}_{\mathbf{v_0}^{1:f}}$ for frames in a video is defined as:

$$\mathcal{T}_{\mathbf{v_0}^{1:f}} = \text{diag}(\mathbf{d}^{-1})\mathbf{Z}^\top\mathbf{Z}\text{diag}(\mathbf{d}^{-1}); \mathbf{Z} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_f]^\top; \mathbf{d} = [\|\mathbf{v}_1\|, \|\mathbf{v}_2\|, \ldots, \|\mathbf{v}_f\|] \quad (6)$$

Here, $\mathbf{v}_i$ represents the flattened latent representation of the i^{th} frame, and diag$(\mathbf{d}^{-1})$ is a diagonal matrix with inverse latent vector norms. In $\mathcal{T}_{\mathbf{v_0}^{1:f}} \in \mathbb{R}^{f \times f}$, larger values indicate greater cosine similarity between frame pairs. The temporal feature is represented as $\mathcal{T}_{\hat{\mathbf{v}}_{0,\theta^{\text{stu}}}}^{\mathcal{I},p}$, where $\hat{\mathbf{v}}_{0,\theta^{\text{stu}}}$ is the latent representation predicted by the student model. Our optimization objective is:

$$\theta^{\text{stu},*} = \arg\min_{\theta^{\text{stu}}} \mathcal{R}_{\text{tmp}}^{\theta^{\text{stu}}} = \mathbb{E}_{\mathcal{I}_i, \mathcal{I}_j \sim \mathcal{V}_{\text{RAS}}}\left[\mathcal{D}(\mathcal{T}_{\hat{\mathbf{v}}_{0,\theta^{\text{stu}}}}^{\mathcal{I}_i,p^\#}, \mathcal{T}_{\hat{\mathbf{v}}_{0,\theta^{\text{stu}}}}^{\mathcal{I}_j,p^\#})\right], \quad (7)$$

The total training loss function, where λ is a hyperparameter, for the second phase of training is:

$$\mathcal{L}_2(\theta^{\text{stu}}, \theta^{\text{tgt}}, \Psi) = \mathcal{L}_{\text{CD}}(\theta^{\text{stu}}, \theta^{\text{tgt}}, \Psi) + \lambda \mathcal{R}_{\text{tmp}}^{\theta^{\text{stu}}}. \quad (8)$$

3 EXPERIMENT

We conducted comprehensive experiments to validate our method's superiority. In the **Main Experiment**, we evaluate video generation quality against state-of-the-art methods on standard datasets. The **Efficiency Experiment** assesses inference time and generation performance. In the **Ablation Study**, we analyze the effectiveness of each sub-module within the CoSeC-LCD framework. The **Generalization Experiment** examines performance on unseen tasks to evaluate generalization capabilities. Finally, in the **Zero-Shot Experiment**, we demonstrate our method's rapid video generation across various domains.

Method	Inference Steps ↓	SSIM ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
FOMM (NeurIPS'19)	-	0.648	0.335	85.03	90.09	405.22
MRAA (CVPR'21)	-	0.672	0.296	54.47	66.36	284.82
DreamPose (CVPR'23)	100	0.509	0.450	79.46	80.51	551.56
DisCo (CVPR'24)	50	0.674	0.285	**28.31**	55.17	267.75
Magic Animate (CVPR'24)	25	0.714	**0.239**	32.09	**21.75**	**179.07**
Animate Anyone (CVPR'24)	20	0.718	0.285	-	-	**171.90**
MagicDance (ICML'24)	50	0.752	0.292	**25.50**	46.30	216.01
Champ (ECCV'24)	20	**0.804**	**0.231**	30.17	**21.23**	**162.62**
ReferenceLCM (Ours)	**2**	**0.766**	0.259	32.11	22.86	203.37
CoSeCLCD (Ours)	**2**	**0.769**	**0.253**	**29.13**	**21.01**	181.72

Table 1: The quantitative results comparison for the Tik Tok dataset, the top 3 methods for each metric are prominently **highlighted** to emphasize their superior performance.

Method	Inference Steps ↓	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
MRAA (CVPR'21)	-	0.749	0.212	23.42	253.65
TPSMM (CVPR'22)	50	0.746	0.213	22.87	247.55
PIDM (CVPR'23)	50	0.713	0.288	30.28	1197.39
DreamPose (ICCV'23)	100	0.885	0.068	**13.04**	238.74
Animate Anyone (CVPR'24)	20	**0.931**	**0.044**	-	**81.60**
Champ (ECCV'24)	20	0.908	0.067	**16.01**	88.06
ReferenceLCM (Ours)	**2**	0.890	0.069	17.16	94.26
CoSeCLCD (Ours)	**2**	**0.908**	**0.066**	16.92	**87.45**

Table 2: The quantitative results comparison for the UBC dataset, the top 2 methods are **highlighted**.

3.1 Details

Benchmark We utilized two widely adopted open-source datasets, TikTok Jafarian & Park (2022) and UBC Fashion Zablotskaia et al. (2019), as benchmarks for our research Hu (2024); Zhu et al. (2024); Xu et al. (2024b). The TikTok dataset features diverse actions and is primarily used to evaluate video quality under complex movements, while the UBC Fashion dataset focuses on clothing displays with minimal motion, emphasizing detail consistency. To further assess the generalization capability of our method, we also collected a test dataset, **Wild-TikTok**, which is similar to the TikTok dataset but offers higher video quality. As for more details about training, please refer to C.

Metrics To quantitatively evaluate the comprehensive performance of different method, we employ several wild-adopted metrics, including, Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018),Frechet Inception Distance (FID), Video Frechet Inception Distance(Vid-FID) and Frechet Video Distance (FVD) Unterthiner et al. (2019). These metrics provide a comprehensive assessment of the quality of generated results and their discrepancies from real data.

3.2 Results

3.2.1 Main Experiment

To validate the effectiveness of our method, we conducted a comprehensive quantitative comparison against several state-of-the-art approaches. The selected methods include FOMM Siarohin et al. (2019a), MRAA Siarohin et al. (2021), DreamPose Karras et al. (2023), DisCo Wang et al. (2023a), TPSMM Zhao & Zhang (2022), PIDM Bhunia et al. (2023), Magic Animate Xu et al. (2024b), Animate Anyone Hu (2024), MagicDance Chang et al. (2023), and the teacher model Champ Zhu et al. (2024). We listed the inference steps and corresponding metrics for each method. To ensure fairness, we used the same settings as in DisCo, which is widely adopted..

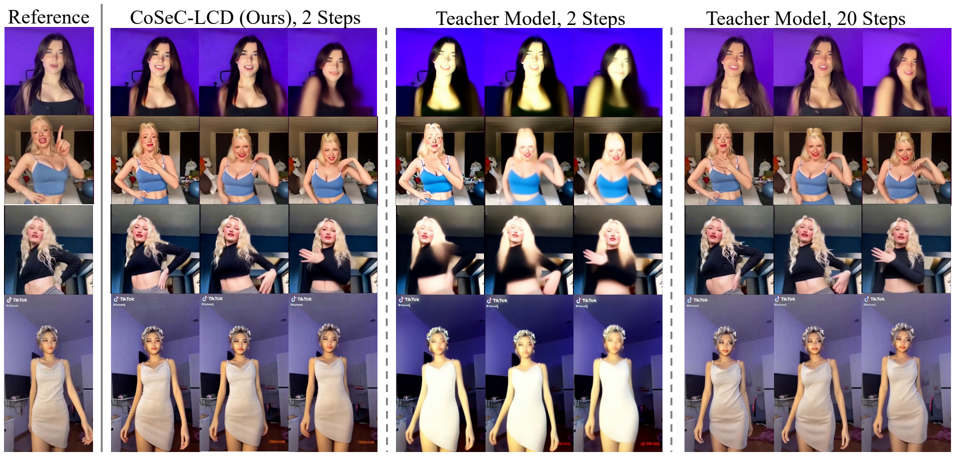

Figure 4: Main experiment's visualization. We provided qualitative results for our method and the teacher model at the same inference step, as well as a comparison to the teacher model with 20 steps.

Step	**Method**	**SSIM** ↑	**LPIPS** ↓	**FID** ↓	**FVD** ↓	**FID-VID** ↓
2	Champ (Teacher)	0.691	0.290	81.41	408.20	45.65
4	Champ (Teacher)	0.707	0.281	48.25	313.34	37.02
6	Champ (Teacher)	0.742	0.267	35.18	209.67	26.91
8	Champ (Teacher)	0.759	0.258	32.68	192.66	23.54
2	**CoSeC-LCD (Ours)**	0.769	0.253	29.13	181.72	21.01

Table 3: Comparison of performance between our model and the teacher under low-step inference.

Human Dance Generation We conducted comparative experiments on the Tik Tok dataset. The results are shown in Table 1. Our proposed two methods, ReferenceLCM and CoSeC-LCD, demonstrate superior performance compared to several state-of-the-art approaches, especially in the FID and FID-VID metrics, outperforming nearly all baselines. Notably, our method achieved this with just 2 inference steps, frequently ranking among the top performers in both tables. Moreover, among the two our proposed approaches, CoSeC-LCD significantly surpasses ReferenceLCM, demonstrating the substantial contribution of the CoSeC-LCD. Additionally, we conducted a qualitative analysis of the main experiment results, as shown in Figure 4. Our method demonstrated comparable performance to the teacher model with just 2 inference steps, while significantly outperforming the teacher model at the same step count. These results demonstrate that our method can generate high-quality, complex character dance sequences even at extremely low inference steps.

Fashion Style Video Synthesis We conducted comparative experiments on the test split of the UBC Fashion dataset, with results presented in Table 2. Our method demonstrated superior performance, even slightly surpassing the teacher model in LPIPS and FVD metrics. This further underscores the remarkable enhancement of the CoSeC-LCD approach in fine-grained control capabilities.

3.2.2 Efficiency Compare

Another critical question is how significant our advantage over the teacher model is under low inference steps. To illustrate the performance comparison between the two methods, we recorded the performance metrics in the TikTok dataset under low-step inference, with the teacher model set to 2-8 steps. The results, as shown in Table 3, indicate that our method maintains a significant edge over the teacher model, achieving an impressive 4X speedup while delivering superior overall performance. This clearly highlights the effectiveness of our innovative approach in optimizing video generation efficiency. Additionally, we provide more qualitative results in Appendix E to visually demonstrate the performance differences between the two methods.

Figure 5: We provide a visualization of the temporal self-contrastive regularization method. To clearly illustrate jitter and abrupt changes, we employed heatmaps to show pixel-level differences between consecutive frames. Areas of abnormal change reflecting the jitter occurs.

Optimization Aspect	**Method**	**SSIM** ↑	**LPIPS** ↓	**FID** ↓	**FVD** ↓
-	ReferenceLCM	0.766	0.259	32.11	203.37
Spatial	+ ETA	0.768	0.256	30.29	193.85
	+ CoNS	0.770	0.254	29.59	189.36
Temporal	+ ETA	0.769	0.253	29.13	181.72

Table 4: Each subsequent row builds on the previous one to highlight the performance improvements each method contributes.

3.2.3 Ablation Study

We conducted two types of ablation experiments to validate the effectiveness of our proposed spatial and temporal self-contrastive regularization methods, using a progressive addition approach to highlight the contribution of each method. The results in TikTok dataset are presented in Table 4.

Effectiveness of Spatial Self-Contrastive We evaluated the enhancements brought by adding ETA and CoNS at the spatial level, i.e., frame quality. Both methods showed further improvements over the previous baseline. We also provide high-resolution reference images in Figure 6 to illustrate the advancements of our approach. While the pure ReferenceLCM achieved efficient generation speeds, it often lacked satisfactory detail control, resulting in some unreasonable artifacts. The ETA method, lacking sufficient distinction, faced clarity loss. However, by incorporating CoNS, we achieved satisfactory results in both clarity and detail control.

Figure 6: Qualitative comparison results of the spatial-level ablation experiments clearly highlight the defects in the generated outputs, where the Self-Contrastive method was not fully applied, providing better clarity for understanding.

Effectiveness of Temporal Self-Contrastive As shown in Table 4, our proposed Temporal Self-Contrastive (TSC) method demonstrates superior performance in quantitative metrics. To provide a clearer illustration, we analyzed pixel-level differences between adjacent frames in a consecutive video and visualized these differences using a heatmap. This visualization effectively demonstrates the influence of incorporating the Temporal Self Contrastive method on the temporal smoothness of the video flow, as illustrated in Figure 5.

3.2.4 Generalization Experiment

To evaluate the generalization capability of our method, we tested its performance on the unseen datasets. Given that the TikTok dataset often suffers from low resolution, we quantitatively compared our method and the teacher model under the same inference steps, as well as the teacher model's full inference scenario. The results, shown in Table 5, demonstrate that our method main-

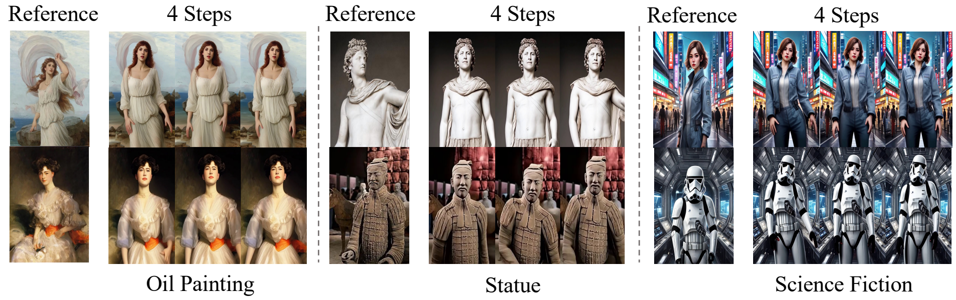

Figure 7: Our proposed method generates high-quality, controllable, consistent videos across multiple domains in just 4 inference steps.

Inference Step	Method	SSIM ↑	LPIPS ↓	FVD ↓	FID-VID ↓
20	Champ (Teacher)	0.764	0.249	173.01	18.43
1	Champ (Teacher)	0.605	0.381	562.20	84.02
	CoSeC-LCD (Ours)	0.720	0.292	325.84	38.33
2	Champ (Teacher)	0.641	0.344	495.01	66.95
	CoSeC-LCD (Ours)	0.732	0.274	198.43	22.51
4	Champ (Teacher)	0.707	0.305	305.23	35.75
	CoSeC-LCD (Ours)	0.739	0.268	192.53	19.58

Table 5: Results of Generalization Experiments in Wild-Tiktok Dataset

tains comparable performance to the teacher model on the unseen high-definition dataset. Notably, under equal inference steps, our method consistently exhibits a significant performance advantage.

3.2.5 Zero-shot Experiment

We further demonstrate the performance of our method in unseen cross-domain scenarios, where there is a substantial gap from the examples in the training dataset. To this end, we collected a diverse set of samples featuring varying styles, specifically three reference image styles—science fiction, sculptures and oil paintings—that differ significantly from those in the training set. The results, shown in Figure 7, illustrate that our method exhibits robust cross-domain generalization capability, even under low inference steps.

4 Conclusion and Future Work

In conclusion, our work presents significant advancements in controllable human animation by tackling both the speed and quality limitations that exist in current video generation methods. By introducing the ReferenceLCM architecture, we dramatically improve the efficiency of video synthesis while maintaining high fidelity, thus addressing the common challenge of slow generation times. Additionally, our hierarchical CoSeC-LCD regularization framework leverages contrastive learning to optimize both spatial and temporal dimensions, ensuring that the generated videos exhibit consistent and coherent motion. Key methods like Equivalent Target Aggregation (ETA) ensure cohesion among equivalent samples, while Contrastive Negative Sampling (CoNS) enhances the distinction between inter-source samples, collectively improving generation precision. Extensive qualitative and quantitative experiments show that our approach not only matches state-of-the-art techniques in terms of output quality but also achieves significant acceleration, making it ideal for real-time applications. Moreover, our method demonstrates strong zero-shot capabilities, effectively generating high-quality, controllable videos without the need for fine-tuning on specific datasets. This further highlights the robustness and flexibility of our approach in various scenarios. Future research will focus on extending our method to multi-person and multi-view generation, thus broadening its applicability and reinforcing its impact across a wider range of animation tasks.

REFERENCES

Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5968–5976, 2023.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pp. 561–578, 2016.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5933–5942, 2019.

Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023.

Naihan Chen, Wei Ping, Ruoming Pang, and Ron J Weiss. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

Adailton Goncalves da Silva, Marcus Vinicius Mendes Gomes, and Ingrid Winkler. Virtual reality and digital human modeling for ergonomic assessment in industrial product development: a patent and literature review. *Applied Sciences*, 12(3):1084, 2022.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision (ECCV)*, 2024.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alex Song, Paul Chu, Ya Chen, Ilya Sutskever, and Pieter Abbeel. Cascaded diffusion models for high fidelity image generation. In *Journal of Machine Learning Research*, 2022.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

J Stuart Hunter. The exponentially weighted moving average. *Journal of quality technology*, 18(4): 203–210, 1986.

Yasamin Jafarian and Hyun Soo Park. Self-supervised 3d representation learning of dressed humans from social media videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (7):8969–8983, 2022.

Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1547–1555, 2015.

Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similarity score distillation for zero-shot video editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22747–22757, 2023.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22680–22690, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Zehua Kong, Wei Ping, Jingxin Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2020.

Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3195–3204, 2021.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6050–6059, 2017.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024.

Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8239–8249, 2021.

Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jialei Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. Echoaudio: Efficient and high-quality text-to-audio generation with minimal inference steps. In *ACM Multimedia 2024*.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, 2018.

Sebastian Loeschcke, Serge Belongie, and Sagie Benaim. Text-driven stylization of video objects. In *European Conference on Computer Vision*, pp. 594–609. Springer, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3967–3976, 2019a.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3967–3976, 2019b.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. In *ACM Transactions on Graphics (TOG)*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf`.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019b.

Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040*, 2023a.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*, 2022.

Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz. Virtual reality. *Business & Information Systems Engineering*, 62:455–461, 2020.

Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3650–3660, 2021.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024a.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024b.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3657–3666, 2022.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 803–818, 2018.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024.

A Related Work

A.1 Diffusion Models

Diffusion models have gained significant attention in the field of generative modeling due to their ability to generate high-quality data samples by reversing a gradual noise addition process. The theoretical foundations of diffusion models are rooted in nonequilibrium thermodynamics and SDEs. Early work by Sohl-Dickstein et al. (2015) introduced the concept of deep unsupervised learning using nonequilibrium thermodynamics, laying the groundwork for diffusion models. The subsequent development of DDPMs by Ho et al. (2020) provided a practical framework for denoising diffusion-based generative modeling. Song & Ermon (2019) extended this framework by introducing score-based generative models, where the data distribution's gradients, or scores, are directly estimated. Further advancements, such as the work of Song et al., have refined the understanding of diffusion models using SDEs, enabling the generation of high-fidelity data samples across various domains.

The underlying idea is to transform a complex data distribution into a simple, known prior distribution, typically a Gaussian, through a sequence of small perturbations, and then reverse this process to generate new data. The diffusion process involves adding noise to the data over a continuous time horizon, transforming the original data distribution into a simple prior distribution. This transformation can be mathematically formulated as a forward SDE:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + \sqrt{g(t)}d\mathbf{w}_t \tag{9}$$

where $\mathbf{x}_t$ represents the data at time t, $\mathbf{f}(\mathbf{x}_t, t)$ is the drift term controlling the deterministic part of the evolution, $g(t)$ modulates the stochastic component, and $d\mathbf{w}_t$ is the increment of a Wiener process, representing the noise. A common choice for $\mathbf{f}(\mathbf{x}_t, t)$ is $-\frac{1}{2}\beta(t)\mathbf{x}_t$, with $\beta(t)$ as the noise strength parameter. This configuration ensures that as time progresses, the data distribution converges to a prior distribution, typically a standard Gaussian $\mathcal{N}(0, I)$.

To generate data, the reverse of this diffusion process is considered. The reverse-time SDE, according to the theory of reversing stochastic processes, is given by:

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - \mathbf{g}(\mathbf{x}_t, t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)\right] dt + \sqrt{\mathbf{g}(\mathbf{x}_t, t)}d\bar{\mathbf{w}}_t \tag{10}$$

Here, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ is the score function, representing the gradient of the log-probability density of the data at time t. This term guides the reverse process towards higher-probability regions of the data distribution, effectively reconstructing the data from noise. Accurately estimating this score function is crucial and is typically achieved through a neural network trained using score matching techniques. The network, denoted as $\mathbf{s}_\theta(\mathbf{x}, t)$, is optimized to match the true score function by minimizing the loss:

$$L(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t}\left[\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)\|^2\right] \tag{11}$$

By minimizing this loss, the model learns to approximate the score function, enabling the reverse process to generate high-quality data samples.

These advancements have established diffusion models as a versatile and robust framework for generative modeling. By leveraging the theoretical properties of SDEs and the flexibility of neural networks, these models achieve a delicate balance between high-quality data generation and computational tractability. Their application spans a wide range of areas, including image synthesis Dhariwal & Nichol (2021); Ho et al. (2022), audio generation Kong et al. (2020); Chen et al. (2020a), and beyond Rombach et al. (2022); Nichol et al. (2021), making them a cornerstone of modern generative modeling research.

A.2 Skinned Multi-Person Linear

SMPL (Skinned Multi-Person Linear Model) is a widely-used 3D human body model designed to provide a realistic and controllable representation of the human body. It represents the human body as a mesh with a fixed topology and a set of parameters that describe the shape and pose of the body, making it a powerful tool for various applications in computer vision, computer graphics, and machine learning. The model is parameterized by shape parameters , capturing body shape variations, and pose parameters representing joint rotations. These parameters are used together with Linear Blend Skinning (LBS) to deform the mesh according to a skeletal structure, providing a versatile representation of human body shapes and poses. SMPL has been successfully applied in human pose estimation Bogo et al. (2016), motion capture Lassner et al. (2017), clothing simulation Pons-Moll et al. (2017), and human reconstruction from partial observations Pavlakos et al. (2019).

B Preliminary

B.1 Latent Consistency Distillation

Latent Consistency Distillation (LCD) is a training framework designed to accelerate the convergence of diffusion models by enforcing *self-consistency* in the latent space. The core principle of LCD is based on minimizing discrepancies between latent states across different time steps within a denoising process, ensuring that they follow a consistent trajectory along a predefined Probability Flow ODE (PF-ODE).

The LCD method leverages the self-consistency property, where the model, for any noised latent variable x_t, is trained to map it to a corresponding denoised estimate along the PF-ODE path at an arbitrary time step t. Mathematically, this self-consistency can be expressed as:

$$f_\theta(x_t, t) = f_\theta(x_{t'}, t'), \quad \forall t, t' \in [\epsilon, T]$$

where t and t' represent different time steps, T is the total number of denoising steps, and ϵ is a small positive constant representing the start of the denoising process. This ensures that the model's output at different points in the trajectory remains consistent.

To encourage the self-consistency property, the model parameters θ are trained using a *consistency distillation* loss function, which minimizes the distance between latent states at subsequent time steps. The distillation loss can be formulated as:

$$\mathcal{L}(\theta, \theta^*; \Phi) = \mathbb{E}_{x,t}\left[d(f_\theta(x_{t+1}, t_{n+1}), f_{\theta^*}(\hat{x}_{t_n}, t_n))\right]$$

Here, θ^* represents the exponentially weighted moving average (EMA) of the model parameters θ, and $d(\cdot, \cdot)$ is a distance metric (e.g., ℓ_2-norm) used to measure the deviation between the predicted latent state and the true state at time t_n. The function Φ corresponds to a numerical ODE solver used to approximate the denoising process. The next latent estimate $\hat{x}_{t_n}$ is computed as:

$$\hat{x}_{t_n} = x_{t_{n+1}} + (t_n - t_{n-1})\Phi(x_{t_{n+1}}, t_{n+1}; \phi)$$

C Settings

We trained our model on the open-source training samples provided by Champ, consisting of approximately 800 videos. The training process was conducted in two distinct phases. In the **first phase**, which focused on spatial aspects, specifically the visual quality of individual video frames, the model was trained for 3000 steps. The classifier-free guidance (CFG) scale, ω, was set to 2.5. The ETA weight, ϕ_1, was 0.1, and the CoNS weight was 0.02, aimed at maintaining a consistent balance across different loss components. In the **second phase**, focused on temporal aspects, the model was trained for 2000 steps. The ETA weight was set to 0.05, CoNS weight was reduced to 0, and the CFG scale ω was set to 1.5. For both phases, the learning rate was $1e^{-6}$, and the Exponential Moving Average (EMA) decay factor α was set to 0.95. Training was conducted using four A800 GPUs, while inference requires one A800 GPU, with CFG disabled during this stage.

Algorithm 2 The training algorithm for ReferenceLCM.

Input: Target video sequence $\{x_0\}$, pose-guidance sequence $\{\mathbf{c}_i^p\}$ from SMPL, reference image $\mathcal{I}$, number of diffusion timesteps T, distance metric $\mathcal{D}$ (e.g., Huber Loss), PF-ODE solver Ψ, classifier-free guidance weight ω, EMA decay rate α.
Initialize: Model parameters θ^{stu}, θ^{tgt}, and θ^{tch} with the same initial weights.
for each epoch **do**
  Encode noisy latent input $\mathbf{x}_{t_{n+k}}$ using VAE.
  Compute pose-guidance condition: $\mathbf{c}^p = \mathcal{E}^{\mathcal{G}}(x_0^{dpt}) \oplus \mathcal{E}^{\mathcal{G}}(x_0^{smt}) \oplus \mathcal{E}^{\mathcal{G}}(x_0^{skt}) \oplus \mathcal{E}^{\mathcal{G}}(x_0^{nml})$
  Obtain attention weights from R-UNet $\mathcal{F}^{\mathcal{R}}(\mathcal{I})$ and CLIP embedding $\mathbf{c}^{\mathcal{I}}$.
  Compute the output of student D-UNet:

$$f_\theta^{\text{stu}}(\mathbf{x}_{t_{n+k}}, \mathbf{c}^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}) = c_{\text{skip}}(t_{n+k})\mathbf{x}_{t_{n+k}} + c_{\text{out}}(t_{n+k})\mathcal{F}_\theta^{\text{stu}}(\mathbf{x}_{t_{n+k}}, t_{n+k}, \mathbf{c}^p, \mathcal{F}^{\mathcal{R}}(\mathcal{I}), \mathbf{c}^{\mathcal{I}})$$

  Classifier-Free Guidance: Use ODE solver Ψ to estimate $\hat{\mathbf{x}}_{t_n}^{\Psi,\omega}$:

$$\hat{\mathbf{x}}_{t_n}^{\Psi,\omega} \leftarrow \mathbf{x}_{t_{n+k}} + (1+\omega)\Psi\left(\mathbf{x}_{t_{n+k}}, \mathbf{c}^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}, t_{n+k}, t_n\right) - \omega\Psi\left(\mathbf{x}_{t_{n+k}}, \mathbf{c}^p, \mathcal{I}, \emptyset, t_{n+k}, t_n\right)$$

  Update model parameters:

$$\mathcal{L}_{\text{CD}}(\theta^{\text{stu}}, \theta^{\text{tgt}}, \Psi) = \mathbb{E}\left[\mathcal{D}\left(f_\theta^{\text{stu}}\left(\mathbf{x}_{t_{n+k}}, \mathbf{c}^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}\right), f_\theta^{\text{tgt}}\left(\hat{\mathbf{x}}_{t_n}^{\Psi,\omega}, \mathbf{c}^p, \mathcal{I}, \mathbf{c}^{\mathcal{I}}\right)\right)\right]$$

  EMA Update: Update weights θ^{tgt} using student model θ^{stu}:

$$\theta^{\text{tgt}} \leftarrow \alpha\theta^{\text{tgt}} + (1-\alpha)\theta^{\text{stu}}$$

end for
Output: Trained model parameters θ^{stu}.

Figure 8: More showcase in Zero-Shot domains.

D THE TRAINING ALGORITHM FOR REFERENCELCM

We provide a detailed description of the training algorithm for ReferenceLCM, where the EMA decay weight α is a hyperparameter.

E MORE QUALITATIVE RESULTS

We visualized the results to provide an intuitive comparison that clearly demonstrates the significant improvements of our method over the teacher model in low-step inference. These visualizations effectively highlight the enhancements in video quality and consistency achieved through our ap-

Figure 9: More showcase in the Main Experiment.

Figure 10: More showcase in unseen reference.

proach, particularly in scenarios where the teacher model struggles with maintaining fidelity and coherence.

We further showcase the effectiveness of our model from three aspects:

Examples from the TikTok test set, where our method produces high-quality and consistent animations under challenging conditions; Animations generated from unseen real human references, demonstrating the model's robustness in generalizing to new inputs; Zero-shot generation results, where our method exhibits strong performance even in previously unexplored domains, highlighting its adaptability across different domains.