VideoMaker: Zero-shot Customized Video Generation with the Inherent Force of Video Diffusion Models

Tao Wu [1,2 *], Yong Zhang [3 *], Xiaodong Cun [3 *], Zhongang Qi [4 †], Junfu Pu [2],
Huanzhang Dou [1], Guangcong Zheng [1], Ying Shan[2,3], Xi Li [1 †]
[1]College of Computer Science and Technology, Zhejiang University
[2]ARC Lab, Tencent PCG [3]Tencent AI Lab [4]Huawei Noah's Ark Lab

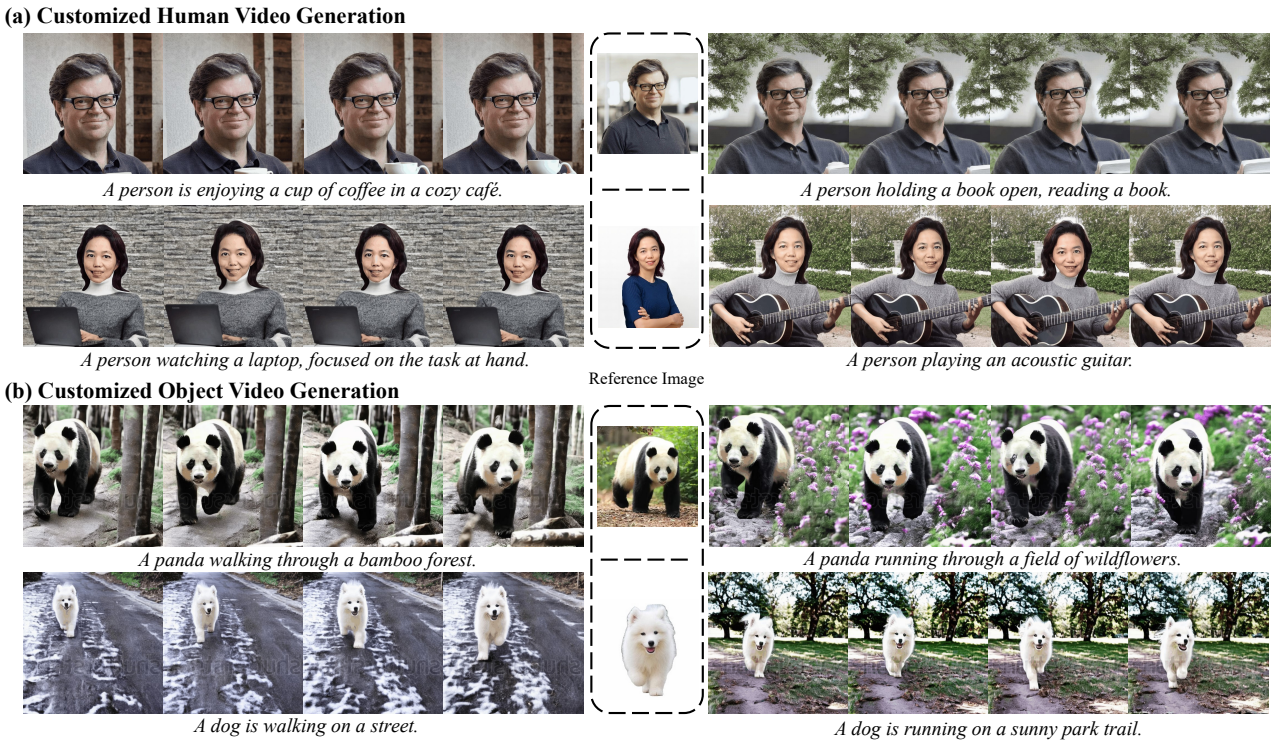

Figure 1. Visualization for our VideoMaker. Our method achieves high-fidelity zero-shot customized human and object video generation based on AnimateDiff [26].

Abstract

Zero-shot customized video generation has gained significant attention due to its substantial application potential. Existing methods rely on additional models to extract and inject reference subject features, assuming that the Video Diffusion Model (VDM) alone is insufficient for zero-shot customized video generation. However, these methods often struggle to maintain consistent subject appearance due to suboptimal feature extraction and injection techniques. In this paper, we reveal that VDM inherently possesses the force to extract and inject subject features. Departing from previous heuristic approaches, we introduce a novel framework that leverages VDM's inherent force to enable high-quality zero-shot customized video generation. Specifically, for feature extraction, we directly input reference images into VDM and use its intrinsic feature extraction process, which not only provides fine-grained features but also significantly aligns with VDM's pre-trained knowledge. For feature injection, we devise an innovative bidirectional interaction between subject features and generated content through spatial self-attention within VDM, ensuring that VDM has better subject fidelity while maintaining the diversity of the generated video. Experiments on both customized human and object video generation validate the effectiveness of our framework.

[1]* These authors contributed equally. † Corresponding author.
[2]Work done during Zhongang Qi's tenure at Tencent PCG ARC Lab.

1. Introduction

Video Diffusion Models (VDMs) [5, 9, 19, 57, 70] can generate high-quality videos from a given text prompt. However, these pretrained models unable to create specific videos from a given subject since this customized subject is hard to be described by a text prompt only. This problem is so-called customized generation and has been explored by personalized fine-tuning [6, 53, 65, 67]. Yet, the time-consuming subject-specific finetune limits its usage in the real world. Recently, Some methods [23, 32] based on [58, 71] have initially explored zero-shot customized video generation. But these methods still fail to maintain a consistent appearance with the reference subject.

Two keys for customized video generation are **subject feature extraction** and **subject feature injection**. Current methods rely on additional models to extract and inject subject features, often overlooking the inherent capabilities of VDMs. e.g., some methods [26, 58, 68] inspired by [79], employ an additional ReferenceNet for feature extraction and directly add the subject features to the VDMs for injection (Figure 2 (a)). However, these methods introduce numerous additional training parameters, and this pixel-wise injection method significantly restricts the diversity of the generated videos. Other methods [23, 32, 38, 71] employ the pre-trained cross-modal alignment model [43, 45, 50, 69] as feature extractors and inject subject feature by cross-attention layer (Figure 2 (b,c)). Nevertheless, these methods produce only coarse-grained, semantic-level features from the pre-trained extractor, which fail to capture the details of the subject. Consequently, these well-designed heuristic methods have not achieved satisfactory results in customized video generation. A question naturally arises: *Perhaps VDMs have the force to extract and inject subject features, and we only need to activate and use these forces in a simple way to achieve customized generation ?*

Rethinking the VDMs, we identified some potential inherent forces. For subject feature extraction, since inputting a noise-free reference image can be seen as a special case with a timestep of 0, the pre-trained VDM is already capable of extracting features from this without additional training. For subject feature injection, the spatial self-attention in the VDM primarily models the relationships between different pixels within a frame, making it more suitable for injecting subject reference features that are closely related to the generated content. Moreover, due to the self-adaptive nature of spatial self-attention, it can selectively interact with these features, which helps prevent overfitting and promotes diversity in the generated videos. Therefore, if we utilize VDM itself as a fine-grained feature extractor for the subject and then interact the subject features with the generated content through spatial self-attention, we can leverage the inherent force of VDM to achieve customized generation.

Inspired by the above motivation, we present our VideoMaker, a novel framework that leverages the inherent force of VDMs to enable high-quality zero-shot customized generation. We register the reference image as part of the model's input, utilizing the VDM itself for feature extraction. The extracted features are not only fine-grained but also closely aligned with the VDM's inherent knowledge, eliminating the need for additional alignment. For subject feature injection, we use VDM's spatial self-attention to explicitly interact with the subject feature close to VDM's inherent knowledge with the generated content when generating content per frame. Additionally, to ensure the model can effectively distinguish between reference information and generated content during training, we have designed a simple learning strategy to enhance performance further. Our framework employs a native approach to complete subject feature extraction and injection without adding additional modules. It only requires fine-tuning the pre-trained VDM to activate the model's inherent force. Through extensive experiments, we provide both qualitative and quantitative results that demonstrate the superiority of our method in zero-shot customized video generation. Our contributions are summarized as follows:

- We use the inherent force of the video diffusion model to extract fine-grained appearance features of the subject, and the extracted subject appearance information is more friendly to learn for the video diffusion model.
- We revolutionize the previous method of information injection, innovatively using the native spatial self-attention computation mechanism in the video diffusion model to complete subject feature injection.

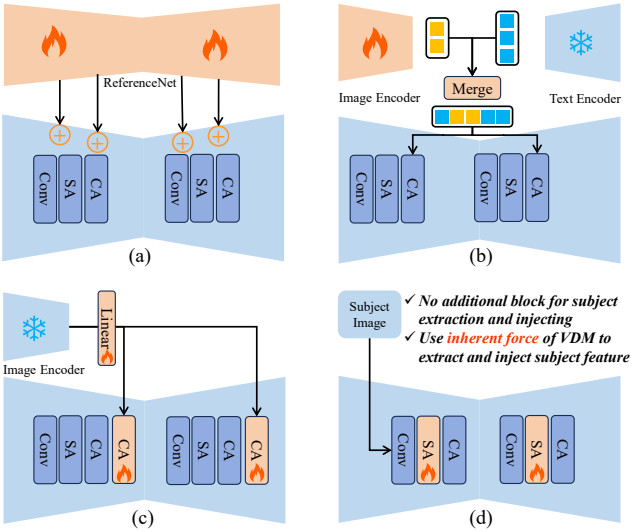

Figure 2. Compared with the existing zero-shot customized generation framework. Our framework does not require any additional modules to extract or inject subject features. It only needs simple concatenation of the reference image and generated video, and VDM's inherent force is used to generate custom video.

- Our framework outperforms existing methods and achieves high-quality zero-shot customized video generation by only fine-tuning some parameters.

2. Related Work

2.1. Text-to-video diffusion models

With the progress in diffusion models and image generation [12, 28, 29, 46–49, 51, 52, 66, 76, 82], there have been significant advancements in text-to-video (T2V) generation. Given the limited availability of high-quality video-text datasets [4, 33], numerous researchers have tried to develop T2V models by leveraging existing text-to-image (T2I) generation frameworks. Some studies [3, 13, 20, 60, 61, 75, 77, 80, 83] have focused on improving traditional T2I models by incorporating temporal blocks and training these new components to convert T2I models into T2V models. Notable examples include AnimateDiff [19], Emu video [17], PYoCo [16], and Align your Latents [4]. Furthermore, approaches such as LVDM [24], VideoCrafter [7, 9], ModelScope [57], LAVIE [62], and VideoFactory [59] have utilized similar architectures, initializing with T2I models, and fine-tuning both spatial and temporal blocks to achieve enhanced visual outcomes. Besides, Sora [5], CogVideoX [70], Latte [44] and Allegro [84] have made notable strides in video generation by integrating Transformer-based backbones [44, 73] and employing 3D-VAE technology. The development of these foundational models lays a solid foundation for customized video generation.

2.2. Customized Image/Video Generation

Similar to the development history of foundational models, the rapid advancement of text-to-image technology has spurred significant progress in customized generation within the image domain. Customized image generation, which adapts to user preferences, has attracted increasing attention [8, 10, 21, 22, 27, 30, 37, 39, 41, 42, 54, 55, 58, 64]. These works can be broadly categorized into two types based on whether the entire model needs to be retrained when changing the subject. The first category includes methods such as Textual Inversion [14], DreamBooth [53], Custom Diffusion [35], and Mix-of-Show [18]. These approaches achieve full customization by learning a text token or directly fine-tuning all or part of the model's parameters. Although these methods often produce content with high visual fidelity to the specified subject, they require retraining when the subject changes. The second category includes methods like IP-Adapter [71], InstantID [58], and PhotoMaker [38]. These approaches employ various information injection techniques and leverage large-scale training to eliminate the need for parameter retraining when the subject changes. Building on these methods, customized video generation has also evolved with advancements in foundational models. DreamVideo [65], CustomVideo [63], Animate-A-Story [25], Still-Moving [6], CustomCrafter [67], and VideoAssembler [81] achieve customization by fine-tuning parts of the Video Diffusion Model. However, this entails a higher training cost for users than customized image generation, resulting in significant inconvenience. Some works, such as VideoBooth [32] and ID-Animator [23], attempt to adopt training methods similar to IP-Adapter. However, they have not yet achieved the same level of success as customized image generation.

3. Preliminary

Video diffusion models (VDMs) [9, 19, 24, 57] are designed for video generation tasks by extending image diffusion models to adapt to video data. VDMs learn a video data distribution by the gradual denoising of a variable sampled from a Gaussian distribution. First, a learnable autoencoder (consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$) is trained to compress the video into a smaller latent space representation. Then, a latent representation $z = \mathcal{E}(x)$ is trained instead of a video x. Specifically, the diffusion model ϵ_θ aims to predict the added noise ϵ at each timestep t based on the text condition c_{text}, where $t \in \mathcal{U}(0, 1)$. The training objective can be simplified as a reconstruction loss:

$$\mathcal{L}_{video} = \mathbb{E}_{z,c,\epsilon\sim\mathcal{N}(0,\mathrm{I}),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, c_{text}, t\right)\|_2^2\right], \quad (1)$$

where $z \in \mathbb{R}^{F\times H\times W\times C}$ is the latent code of video data with F, H, W, C being frame, height, width, and channel, respectively. c_{text} is the text prompt for the input video. A noise-corrupted latent code z_t from the ground truth z_0 is formulated as $z_t = \lambda_t z_0 + \sigma_t \epsilon$, where $\sigma_t = \sqrt{1-\lambda_t^2}$, λ_t and σ_t are hyperparameters to control the diffusion process. In this work, we selected the AnimateDiff [19] as our base video diffusion model.

4. Method

Given a photo of a subject, our goal is to train a model that can extract the subject's appearance and generate a video of the same subject. Besides, the model does not require retraining when changing the subject. We discuss the key ideas of our method in Section 4.1, and detail how we utilize the inherent force of VDM to extract subject features and enable VDM to learn the subject in Section 4.2. In Section 4.3, we introduce our proposed training strategy to better distinguish between reference information and generated content. Furthermore, we add some details about the training and inference in Section 4.4.

4.1. Explore Video Diffusion Model

To achieve customized video generation, two core problems must be addressed: subject feature extraction and feature injection. For subject feature extraction, some works employ

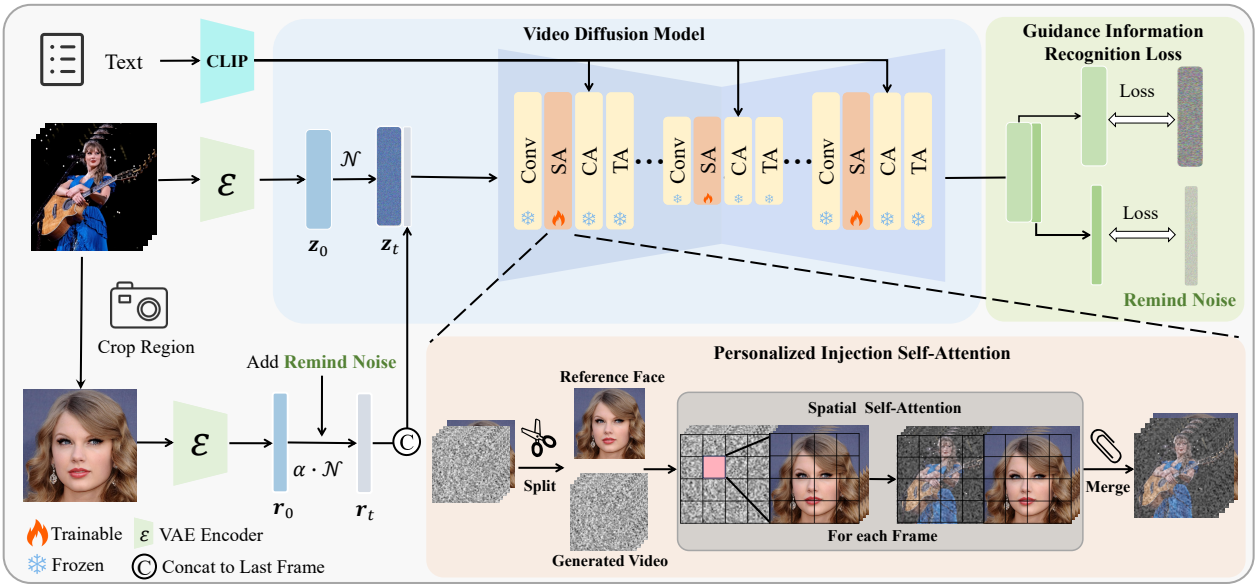

Figure 3. Overall pipeline of VideoMaker. We directly input the reference image into VDM and use VDM's modules for fine-grained feature extraction. We modified the computation of spatial self-attention to enable feature injection. Additionally, to distinguish between reference features and generated content, we designed the Guidance Information Recognition Loss to optimize the training strategy.

cross-model alignment models, such as CLIP [50]. However, because of their training tasks, these models produce coarse-grained features that fail to capture the subject's appearance in detail. Some studies attempt to train a ReferenceNet but significantly increase training overhead. We propose a new method leveraging the pre-trained VDM for subject feature extraction. When the subject reference image is input directly into the VDM without added noise, this can be considered as a special case of VDM at $t = 0$. Therefore, the VDM can accurately process and extract the features of the noise-free reference image. This approach allows for extracting fine-grained subject features without additional training overhead while reducing the domain gap between the extracted features and the VDM's inherent knowledge.

Regarding feature injection, spatial cross-attention is used for VDM's cross-modal interaction between image and text. Influenced by this design, existing methods employ cross-attention heuristically to inject subject features. However, spatial self-attention in VDM is responsible for modeling the relationships between pixels within a frame. In customized video generation, a key objective is to ensure the subject "appears" in the frame. So, injecting subject features when constructing pixel relationships within the frame is a more direct method. Moreover, spatial self-attention can selectively interact with these features, which helps promote diversity in the generated videos. Benefiting from the feature extraction performed by the VDM itself, we can directly use the VDM's inherent spatial self-attention modeling capability for more direct information interaction.

4.2. Personalized Injection Self Attention

Subject feature extraction. Unlike previous approaches, we leverage the existing network structure of the VDM to achieve this, i.e. Resblock in unit-based VDM. As illustrated in Figure 3, given a video x that is encoded into the latent space and then noised to obtain $z_t \in \mathbb{R}^{F\times H\times W\times C}$ through a VAE, along with a reference image R of the specified subject, we first encode the reference image R using the VAE to obtain r without adding noise. We then concatenate the encoded reference image latent space r with z_t along the frame dimension, resulting in $z'_t \in \mathbb{R}^{(F+1)\times H\times W\times C}$ as the actual input to the model. Next, we use the Resblock as a feature extractor to extract features from z'_t, obtaining the input $f \in \mathbb{R}^{(F+1)\times h\times w\times c}$ for the spatial self-attention layer. We then separate the features f to obtain the noise part corresponding to the video to be generated $f_z \in \mathbb{R}^{F\times h\times w\times c}$ and the part corresponding to the reference information $f_r \in \mathbb{R}^{1\times h\times w\times c}$. We have completed the feature extraction for the specified subject at this stage.

Subject feature injection. After extracting the specified subject features, injecting these features into the VDM is next. For each frame f_z^i in f_z, it is transformed into $h \times w$ tokens before computing spatial self-attention. We concatenate f_r with f_z^i so that the input to the spatial self-attention layer for each frame becomes $2 \times h \times w$ tokens. We denote these tokens as $\mathbf{X}$. Then, we fuse the information through spatial self-attention:

$$\mathbf{X}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \quad (2)$$

where $\mathbf{X}'$ represents the output attention features, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices, respectively. Specifically, $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, and $\mathbf{V} = \mathbf{X}\mathbf{W}_V$. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are the corresponding projection matrices, and d is the dimension of the key features. After computing the attention, we separate the output attention features $\mathbf{X}'$ to obtain f'_z and f'_r. Since f'_r is repeated F times, we take the average of the F corresponding results as the final f'_r. Finally, we concatenate the obtained f'_r with f'_z to obtain the updated f', which is then fed into the subsequent model layers for further processing.

4.3. Guidance Information Recognition Loss

Since the actual input z'_t to our framework includes an additional frame compared to the input in Equation 1, the output $\epsilon_\theta \in \mathbb{R}^{(F+1)\times H\times W\times C}$ also has an extra frame relative to the output in Equation 1. A straightforward training objective would be to eliminate the output corresponding to the reference information r and compute the loss only for the remaining frames. This approach encourages the model to focus on learning customized video generation with the specified subject. However, our observations of the final results revealed that without supervision on the reference information, the model struggles to accurately recognize that the reference information r is an image without adding noise, leading to instability in the generated results. To address this, we introduced a Guidance Information Recognition Loss to supervise the reference information, enabling the model to accurately distinguish between the reference information and the generated content, thereby improving the quality of the customized generation. Specifically, during the training process at timestep t, we add a remind noise to the reference information r:

$$r_t = \lambda_{t'} r + \sqrt{1-\lambda_{t'}^2}\epsilon, \tag{3}$$

where $t' = \alpha \cdot t$, and α is a manually set hyperparameter. To prevent the added noise from heavily degrading the reference information, α is set to a small value to ensure that the reminder noise remains minimal. When computing the loss function, we also calculate the loss same as Equation (1) for the reference information r:

$$\mathcal{L}_{reg} = \mathbb{E}_{r,c,\epsilon\sim\mathcal{N}(0,\mathrm{I}),t}\left[\|\epsilon - \epsilon_\theta\left(r_t, c_{text}, t\right)\|_2^2\right]. \tag{4}$$

We use $\mathcal{L}_{reg}$ as an auxiliary optimization objective, combined with the main objective, to guide the model's training:

$$\mathcal{L} = \mathcal{L}_{video} + \beta \cdot \mathcal{L}_{reg}, \tag{5}$$

where β is a hyperparameter. To avoid interfering with the optimization of the main customized video generation task, β is chosen as a relatively small value.

4.4. Training and Inference Paradigm

Training. Our framework's straightforward design allows us to avoid the need for additional subject feature extractors during training. Since we only adjust the number of input tokens to the model's original spatial self-attention layer, injecting subject information into the VDM does not increase the parameter count. We assume that the ResBlock in pretrained VDM is already sufficient to extract the feature information from the reference image. Therefore, our model needs to fine-tune the original VDM's spatial self-attention layer while freezing the parameters of the remaining parts during training. In addition, To enable temporal attention todistinguish well between reference information and generated videos, we recommend fine-tuning the parameters of the motion block synchronously during training. It can also achieve customized video generation without fine-tuning the motion blocks. We also randomly drop image conditions in the training stage to enable classifier-free guidance in the inference stage:

$$\hat{\boldsymbol{\epsilon}}_\theta\left(\boldsymbol{z}_t, \boldsymbol{c}_t, r, t\right) = w\boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, \boldsymbol{c}_t, r, t\right) + (1-w)\boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, t\right). \tag{6}$$

Inference. During the inference process, for the output of the model, the output corresponding to the reference information is discarded directly. Additionally, although we added light noise to the subject's reference image during training to explicitly help the model distinguish the guidance information, we chose to remove the noise addition to the reference image during inference. This ensures that the generated video is not affected by noise, thus maintaining the quality and stability of the output.

5. Experiments

5.1. Experimental Setup

Datasets. Due to the lack of large-scale video customization datasets, we selected CelebV-Text [72] and VideoBooth datasets [32] for our experiments on customized human video generation and customized object generation, respectively. The CelebV-Text dataset contains 70,000 facial video clips with a resolution of at least 512×512 and semi-automatically annotated text prompts. To ensure that the model focuses on learning the subject and avoids overfitting to the background of the subject's reference image, we use subject highlight preprocessing same as VideoBooth. We use Grounding DINO [40] and [34] to remove the background and get the main subject of a frame, which we then use as the reference image. The VideoBooth dataset consists of 48,724 video clips selected from a subset of WebVid [2], covering nine categories of objects: bear, car, cat, dog, elephant, horse, lion, panda, tiger. Since each video in this dataset is accompanied by a reference image, we use the provided reference images directly as input.

Implementation details. To facilitate comparison with other existing methods, we followed the setup of [23, 71] and used the Stable Diffusion 1.5 version of AnimateDiff as the base model for our experiments. All experiments are conducted using four NVIDIA A100 GPUs, with a batch size of 1 per GPU. We fixed the frame stride for each video at 8 and set the output video resolution to 512×512. We used the AdamW optimizer, setting the learning rate to 1×10^{-5} and the weight decay to 1×10^{-2}. For the customized human video generation experiments on the CelebV-Text dataset, we trained for 150,000 steps. For the VideoBooth dataset, due to the smaller data scale, we adjusted the training to 100,000 steps. For the hyperparameter settings in our method, we set α in Equation (3) to 0.01, and β in Equation (5) to 0.1. During the inference generation process, we use DDIM [56] for 30-step sampling and a classifier-free guidance scale of 8 to generate videos.

Baselines. Since customized human video generation is significantly more challenging than customized object generation and better highlights the capability of customized generation, we primarily focus on comparing customized human video generation to demonstrate the effectiveness of our method. We selected IP-Adapter [71], IP-Adapter-Plus, IP-Adapter-FaceID and ID-Animator [23] for a fair comparison. IP-Adapter-Plus represents an enhanced version of IP-Adapter that utilizes Q-Former [36] to extract features from CLIP image embeddings, while IP-Adapter-FaceID substitutes CLIP with a dedicated face recognition model. Since PhotoMaker [38] only has pretrained weights for the SDXL [49] version, we used the results generated with AnimateDiff SDXL at a resolution of 512×512 for comparison. For customized object video generation, we use VideoBooth [32] as the baseline for comparison.

Evaluation metrics. Following the [23, 63, 65], we evaluate generated video quality from two perspectives: overall consistency and subject fidelity. We employ three metrics for overall consistency: CLIP-T, Temporal Consistency (T. Cons.), and Dynamic Degree (DD). CLIP-T measures the average cosine similarity between the CLIP [50] image embeddings of all generated frames and their text embeddings. T. Cons. calculates the average cosine similarity between the CLIP image embeddings of consecutive frames. DD [31] utilizes optical flow to quantify motion dynamics. To evaluate subject fidelity, we use CLIP-I and DINO-I for both customized human and object video generation tasks. CLIP-I assesses the visual similarity between the generated frames and the target subjects by computing the average cosine similarity between the CLIP image embeddings of all generated frames and the reference images. DINO-I [53] is another metric for visual similarity, using ViT-S/16 DINO [78]. Additionally, for customized human video generation, since CLIP-I captures relatively coarse visual features, we also incorporate Face Similarity [11] for a more fine-grained and precise comparison, enhancing the accuracy of our subject fidelity assessment.

Method	CLIP-T	Face Sim.	CLIP-I	DINO-I	T.Cons.	DD
IP-Adapter	0.2064	0.1994	0.7772	0.6825	<u>0.9980</u>	0.1025
IP-Adapter-Plus	0.2109	0.2204	<u>0.7784</u>	<u>0.6856</u>	**0.9981**	0.1000
IP-Adapter-Faceid	0.2477	<u>0.5610</u>	0.5852	0.4410	0.9945	0.1200
ID-Animator	0.2236	0.3224	0.4719	0.3872	0.9891	0.2825
Photomaker(SDXL)	**0.2627**	0.3545	0.7323	0.4579	0.9777	<u>0.3675</u>
Ours	<u>0.2586</u>	**0.8047**	**0.8285**	**0.7119**	0.9818	**0.3725**

Table 1. Comparison with the existing methods for customized human video generation. The best and the second-best results are denoted in bold and underlined, respectively.

Method	CLIP-T	CLIP-I	DINO-I	T.Cons.	DD
VideoBooth	0.266	0.7637	0.6658	0.9564	0.5091
Ous	**0.284**	**0.8071**	**0.7326**	**0.9848**	**0.5132**

Table 2. Comparison with the existing methods for customized object video generation

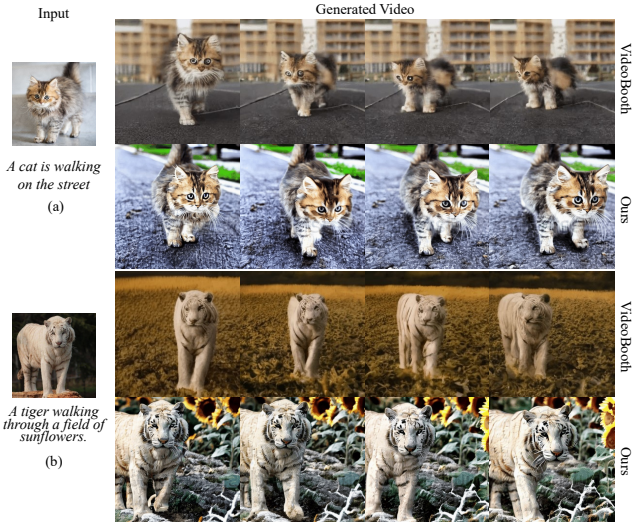

Figure 4. Qualitative comparison for customized object video generation. Compared with the blurry videos generated by VideoBooth [32], our generated videos have more details.

5.2. Quantitative Comparison

Customized object video generation. We selected two examples for each of the nine object categories included in the VideoBooth dataset, totaling 18 subjects. For each subject, we generated 10 prompts tailored to their category using ChatGPT [1]. As shown in Table 2, our method outperforms VideoBooth across all evaluation metrics, demonstrating the effectiveness of our approach.

Customized human video generation. Following [23, 38, 58], we created a testing benchmark comprising 16 different individuals. we generated 25 prompts using ChatGPT, addressing five aspects: expressions, attributes, decorations,

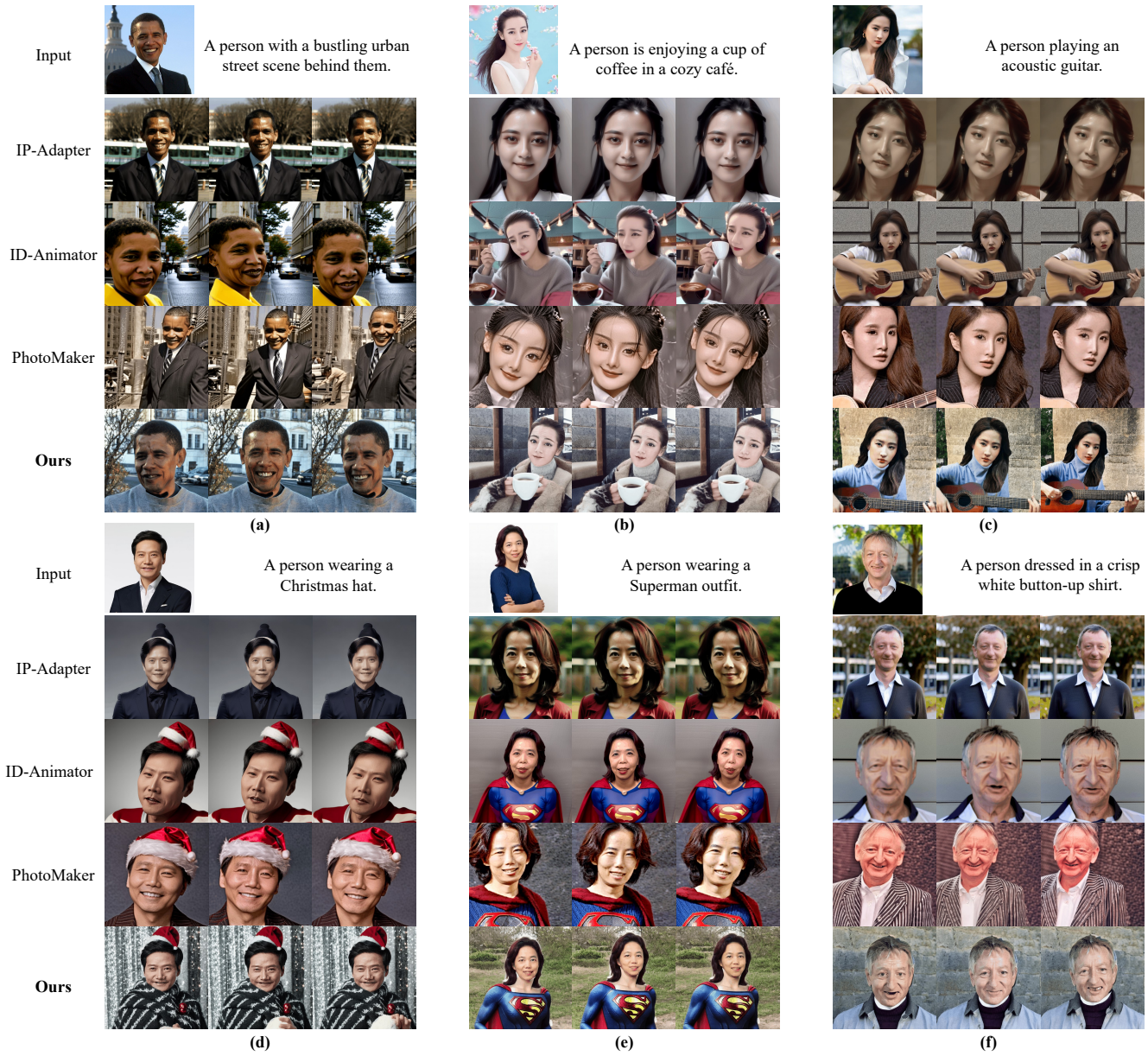

Figure 5. Qualitative comparison for customized human video generation. We compare our method with IP-Adapter [71], ID-Animator [23] and PhotoMaker [38]. We observe that our method achieves high-quality generation, promising editability, and subject fidelity.

PISA	GIRL	W/O Cross	Update Motion	SHP	CLIP-T	Face Sim.	CLIP-I	DINO-I	T.Cons.	DD
✓					0.2206	0.7928	0.7966	0.6694	0.9671	0.2725
✓	✓				0.2258	0.8184	0.8484	0.7536	0.9855	0.2750
✓	✓	✓			0.2291	0.8454	0.8469	0.7351	0.9747	0.2915
✓	✓	✓	✓		0.2302	0.8563	0.8674	0.7635	0.9823	0.3575
✓	✓	✓	✓	✓	0.2586	0.8047	0.8285	0.7119	0.9818	0.3725

Table 3. Quantitative results of each component. “PISA” is our Personalized Injection Self Attention, GIRL is our Guidance Information Recognition Loss, “W/O Cross” refers to whether our reference frame interacts with the text prompt, “Update Motion” refers to whether to update the motion block, “SHP” is our subject highlight preprocessing for datasets,

actions, and backgrounds, to enable a comprehensive evaluation. All images and prompts are provided in the Appendix. As shown in Table 1, our method significantly outperforms existing methods in Face Similarity, CLIP-I, and DINO-I. Our method demonstrates strong performance, especially for Face Similarity, a fine-grained metric that measures subject fidelity. This proves that our approach of using the model's inherent force for customized generation can better extract the subject's features and inject these features into the VDM. For text alignment, our method achieved the best results on the same base model. While PhotoMaker's base model, AnimateDiff SDXL, possesses stronger generative capabilities at the base model level, our method achieves comparable results, demonstrating improved text alignment while maintaining high subject fidelity. The main reason for our T. Cons. behind the IP-Adapter series models is that the videos generated by these models tend to remain static, leading to higher cross-frame consistency. The DD metric illustrates this issue. Moreover, our method has a better degree of dynamicity.

5.3. Qualitative Comparison

To further validate the effectiveness of our method, we compared the visual results generated by our method with existing methods. For the human generation, as shown in Figure 5, our method significantly improves subject fidelity compared to other methods while ensuring text alignment. Notably, our generated humans exhibit more refined facial details, underscoring the advantages of using VDM for feature extraction and information injection, contributing to enhanced consistency in subject appearance. For object generation, as illustrated in Figure 4, our method produces videos that faithfully capture the subject's texture details, whereas VideoBooth often loses facial details when generating animals like cats and tigers. Additionally, in Figure 4(b), VideoBooth failed to generate a sunflower according to the prompt, whereas our method successfully rendered it. Note that the watermark on our generated results comes from the dataset itself, while VideoBooth removes it using an additional external model. We provide more qualitative comparison results and their videos, which can be found in the supplementary materials.

5.4. Ablation Studies

We conducted a series of ablation experiments on the CelebV-Text dataset to evaluate the effectiveness of each component in our framework and validate specific design choices. In our model, the reference frame participates in cross-attention and temporal attention computations after participating in self-attention. Our ablations examine the influence of each processing step on model performance. Since part of the methods we are comparing (e.g. IP-Adapter) do not perform additional data processing, for fairness, we conducted experiments using the original data in the first four rows of Table 3, using random frames of the video as the reference images.

Effect of Personalized Injection Self Attention. In the initial experiment, we applied feature injection exclusively through self-attention while preventing the reference frame from influencing cross-attention calculations. This setup supervised only the frames corresponding to the generated video. As shown in the first line of Table 3, compared to existing methods, just by modifying the subject extraction and injection, we can significantly improve the subject fidelity.

Effect of Guidance Information Recognition Loss. To improve appearance consistency, we introduce Guidance Information Recognition Loss, designed to help the model accurately distinguish reference frame from other frames. This component further stabilizes the subject's appearance across frames, as demonstrated in the second row of Table 3.

Whether to participate in cross-attention. Allowing the reference frame to participate in cross-attention computation introduces two potential outcomes: (1) Altering the subject features, which could negatively impact subject fidelity, or (2) Enabling interaction with textual information, thereby improving text alignment in the generated video and enhancing the coherence of the reference features. To assess this, we conducted experiments where the reference frame was included in cross-attention. The results in row three indicate that allowing this interaction enhances text alignment without compromising subject fidelity.

Whether Update Motion Blocks. As we mentioned in Section 4.4, we found that incorporating motion block training during fine-tuning helps the model better distinguish between reference information and generated video. This approach also enhances video dynamics without compromising subject fidelity, as shown in row four of Table 3.

Effect of Subject Highlight Preprocessing. Finally, we evaluated the impact of subject-specific data preprocessing. Compared to using a random video frame as the subject reference, our preprocessing approach reduces the model's tendency to overfit to irrelevant background details, directing focus onto the subject's appearance features. This improved alignment between the generated video and text prompts, as seen in the last row of Table 3.

6. Conclusion

In this paper, we propose VideoMaker, a novel framework that uses the inherent force of VDM to achieve high-quality zero-shot customized generation. Compared with the heuristic external model to extract and inject subject features, we discover and use the force of inherent VDM to complete the fine-grained subject feature extraction and injection required for customized generation. Experimental results confirm the efficacy of our approach across both customized human and object video generation tasks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 6, 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In Proc. ICCV, pages 1728–1738, 2021. 5, 3

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945, 2024. 3

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In Proc. CVPR, 2023. 3

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2, 3

[6] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. arXiv preprint arXiv:2407.08674, 2024. 2, 3

[7] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512, 2023. 3

[8] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374, 2023. 3

[9] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In Proc. CVPR, pages 7310–7320, 2024. 2, 3

[10] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In Proc. CVPR, pages 6593–6602, 2024. 3

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proc. CVPR, pages 4690–4699, 2019. 6

[12] Huanzhang Dou, Ruixiang Li, Wei Su, and Xi Li. Gvdiff: Grounded text-to-video generation with diffusion models, 2024. 3

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In Proc. ICCV, pages 7346–7356, 2023. 3

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022. 3

[15] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization, 2024. 2

[16] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In Proc. ICCV, 2023. 3

[17] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. arXiv preprint arXiv:2311.10709, 2023. 3

[18] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In Proc. NeurIPS, 2024. 3

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

[20] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In Proc. ECCV, pages 393–411. Springer, 2024. 3

[21] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In Proc. ICCV, pages 7323–7334, 2023. 3

[22] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face adapter for pre-trained diffusion models with fine-grained id and attribute control. arXiv preprint arXiv:2405.12970, 2024. 3

[23] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. arXiv preprint arXiv:2404.15275, 2024. 2, 3, 6, 7

[24] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221, 2022. 3

[25] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. arXiv preprint arXiv:2307.06940, 2023. 3, 2

[26] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In Proc. CVPR, pages 8153–8163, 2024. 1, 2

[27] Miao Hua, Jiawei Liu, Fei Ding, Wei Liu, Jie Wu, and Qian He. Dreamtuner: Single image is enough for subject-driven generation. arXiv preprint arXiv:2312.13691, 2023. 3

[28] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. Group diffusion transformers are unsupervised multitask learners, 2024. 3

[29] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers, 2024. 3

[30] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Chen Liang, Tong Shen, Han Zhang, Huanzhang Dou, Yu Liu, and Jingren Zhou. Chatdit: A training-free baseline for task-agnostic free-form chatting with diffusion transformers, 2024. 3

[31] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proc. CVPR, pages 21807–21818, 2024. 6

[32] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In Proc. CVPR, pages 6689–6700, 2024. 2, 3, 5, 6, 1

[33] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. arXiv preprint arXiv:2407.06358, 2024. 3

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proc. ICCV, pages 4015–4026, 2023. 5, 1

[35] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In Proc. CVPR, pages 1931–1941, 2023. 3

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proc. ICML, pages 19730–19742. PMLR, 2023. 6

[37] Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In Proc. CVPR, pages 2187–2196, 2024. 3

[38] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In Proc. CVPR, pages 8640–8650, 2024. 2, 3, 6, 7, 1

[39] Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. Idea-bench: How far are generative models from professional designing?, 2024. 3

[40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023. 5, 1, 2

[41] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. arXiv preprint arXiv:2303.05125, 2023. 3

[42] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In Proc. NeurIPS, pages 57500–57519, 2023. 3

[43] Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li, Luke Zettlemoyer, Shih-Fu Chang, Wen-Tau Yih, and Hu Xu. Mode: Clip data experts via clustering. In Proc. CVPR, pages 26354–26363, 2024. 2

[44] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048, 2024. 3

[45] Zehong Ma, Shiliang Zhang, Longhui Wei, and Qi Tian. Ovmr: Open-vocabulary recognition with multi-modal references. In Proc. CVPR, pages 16571–16581, 2024. 2

[46] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In Proc. AAAI, pages 4296–4304, 2024. 3

[47] 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. In ICML, 2022.

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proc. ICCV, pages 4195–4205, 2023.

[49] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023. 3, 6

[50] 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 Proc. ICML, pages 8748–8763. PMLR, 2021. 2, 4, 6

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022. 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proc. CVPR, 2022. 3

[53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proc. CVPR, pages 22500–22510, 2023. 2, 3, 6

[54] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In Proc. CVPR, pages 6527–6536, 2024. 3

[55] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In Proc. CVPR, pages 8543–8552, 2024. 3

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020. 6

[57] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571, 2023. 2, 3

[58] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519, 2024. 2, 3, 6, 1

[59] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. arXiv preprint arXiv:2305.10874, 2023. 3

[60] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. Proc. NeurIPS, 36, 2024. 3

[61] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In Proc. CVPR, pages 6572–6582, 2024. 3

[62] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103, 2023. 3

[63] Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. arXiv preprint arXiv:2401.09962, 2024. 3, 6

[64] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In Proc. ICCV, pages 15943–15953, 2023. 3

[65] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In Proc. CVPR, pages 6537–6549, 2024. 2, 3, 6

[66] Tao Wu, Xuewei Li, Zhongang Qi, Di Hu, Xintao Wang, Ying Shan, and Xi Li. Spherediffusion: Spherical geometry-aware distortion resilient diffusion model. In Proc. AAAI, pages 6126–6134, 2024. 3

[67] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. arXiv preprint arXiv:2408.13239, 2024. 2, 3

[68] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In ACM SIGGRAPH 2024 Conference Papers, pages 1–11, 2024. 2

[69] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In Proc. ICLR. 2

[70] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024. 2, 3

[71] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023. 2, 3, 6, 7

[72] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In Proc. CVPR, pages 14805–14814, 2023. 5, 3

[73] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. arXiv preprint arXiv:2403.14148, 2024. 3

[74] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. Proc. NeurIPS, 36, 2024. 2

[75] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: instructing video diffusion models with human feedback. In Proc. CVPR, pages 6463–6474, 2024. 3

[76] Yike Yuan, Huanzhang Dou, Fengjun Guo, and Xi Li. Semanticmim: Marring masked image modeling with semantics compression for general visual representation, 2024. 3

[77] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. Int. J. Comput. Vis., pages 1–15, 2024. 3

[78] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605, 2022. 6

[79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proc. ICCV, pages 3836–3847, 2023. 2

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

[81] Haoyu Zhao, Tianyi Lu, Jiaxi Gu, Xing Zhang, Zuxuan Wu, Hang Xu, and Yu-Gang Jiang. Videoassembler: Identity-consistent video generation with reference entities using diffusion model. arXiv, 2023. 3

[82] Ruisi Zhao, Mingming Li, Zheng Yang, Binbin Lin, Xiaohui Zhong, Xiaobo Ren, Deng Cai, and Boxi Wu. Towards fine-grained hboe with rendered orientation set and laplace smoothing. In Proc. AAAI, pages 7505–7513, 2024. 3

[83] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018, 2022. 3

[84] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. arXiv preprint arXiv:2410.15458, 2024. 3

VideoMaker: Zero-shot Customized Video Generation with the Inherent Force of Video Diffusion Models

Supplementary Material

Category	Prompt
Clothing	A person dressed in a crisp white button-up shirt.
	A person in a sleeveless workout top, displaying an active lifestyle.
	A person wearing a sequined top that sparkles under the light, ready for a festive occasion.
	A person wearing a Superman outfit.
	A person wearing a blue hoodie.
Action	A person holding a book open, reading a book, sitting on a park bench.
	A person playing an acoustic guitar.
	A person laughing with their head tilted back, eyes sparkling with mirth.
	A person is enjoying a cup of coffee in a cozy café.
	A person watching a laptop, focused on the task at hand.
Accessory	A person wearing a headphones, engaged in a hands-free conversation.
	A person with a pair of trendy headphones around their neck, a music lover's staple.
	A person with a beanie hat and round-framed glasses, portraying a hipster look.
	A person wearing sunglasses.
	A person wearing a Christmas hat.
View	A person captured in a close-up, their eyes conveying a depth of emotion.
	A person framed against the sky, creating an open and airy feel.
	A person through a rain-streaked window, adding a layer of introspection.
	A person holding a bottle of red wine.
	A person riding a horse.
Background	A person is standing in front of the Eiffel Tower.
	A person with a bustling urban street scene behind them, capturing the energy of the city.
	A person standing before a backdrop of bookshelves, indicating a love for literature.
	A person swimming in the pool
	A person stands in the falling snow scene at the park.

Table 1. Evaluation text prompts for customized human video generation.

A. Dataset Details

Training dataset. As mentioned in Section 5.1 of the main text, we employed subject highlight preprocessing to process the dataset. Specifically, we first use Grounding DINO [40] with the prompt "head" to process a randomly sampled frame from each video. This provides the bounding box corresponding to the person in each video. We then integrate the SAM [34] model to obtain the subject mask and set the area outside the mask to white, which serves as the reference image for each video. During training, we randomly select any one of the four frames as the actual input reference image. Additionally, we removed videos containing multiple people or those where the proportion of the face is too small. After processing, the CelebV-Text dataset contains 40,600 videos. Furthermore, during training, we applied *RandomHorizontalFlip* and *RandomAffine* transformations to the reference images as data augmentation.

Evaluation dataset. Here we present the test dataset used in Section 5.2. For customized human video generation, we followed the works of [38, 58] and collected 20 different individuals as the test set, as shown in Figure 1. For the text prompts, we considered five factors: clothing, accessories, actions, views, and background, which make up 25 prompts listed in Table 1 for testing. During inference, we processed the reference images using subject highlight preprocessing. For customized object video generation, since

Figure 1. The overview of the celebrity dataset we use to test customized human video generation.

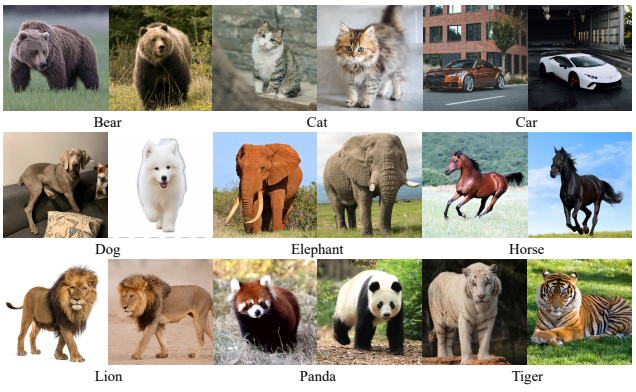

Figure 2. The overview of the dataset we use to test customized object video generation.

VideoBooth [32] did not publicly release their test samples, we collected two samples from each of the nine categories that were not present in the training data for testing. The

Figure 3. The overview of the non-celebrity dataset we used for testing customized human video generation.

prompts used for testing were generated using ChatGPT [1] based on the object categories, as detailed in Table 3. During inference, we processed the reference images using subject highlight preprocessing and set the prompt for Grounding DINO [40] to "`<class word>.` " where `<class word>` represents the category of the object used, such as dog, cat.

B. Quantitative Comparison Results on Non-Celebrity Dataset

Some studies [74] have pointed out that pre-trained text-to-image diffusion models can directly generate photos of certain celebrities. Therefore, in addition to following works such as [38, 58] by selecting some celebrities for testing, we also selected some non-celebrity data for testing. As shown in Figure 3, we followed the Unsplash50 dataset from [15] and collected a small set of 16 recently uploaded images with permissive licenses from https://unsplash.com/ as our non-celebrity dataset to ensure that these images have never appeared in the pre-training data. For the text prompts, we used the same prompts as those for celebrities.

The quantitative comparison results are shown in Table 2. Our method still demonstrates good performance on the non-celebrity dataset. All methods show a slight decrease in metrics on the non-celebrity dataset due to the loss of certain prior knowledge, but the conclusions from the quantitative comparison are largely consistent with those using the celebrity dataset. Our method continues to lead

Method	CLIP-T	Face Sim.	CLIP-I	DINO-I	T.Cons.	DD
IP-Adapter	0.2347	0.1298	0.6364	0.5178	<u>0.9929</u>	0.0825
IP-Adapter-Plus	0.2140	0.2017	<u>0.6558</u>	<u>0.5488</u>	0.9920	0.0815
IP-Adapter-Faceid	0.2457	<u>0.4651</u>	0.6401	0.4108	0.9930	0.0950
ID-Animator	0.2303	0.1294	0.4993	0.0947	**0.9999**	0.2645
Photomaker*	**0.2803**	0.2294	0.6558	0.3209	0.9768	<u>0.3335</u>
Ours	<u>0.2773</u>	**0.6974**	**0.6882**	**0.5937**	0.9797	**0.3590**

Table 2. Comparison with the existing methods for customized human video generation on our non-celebrity dataset. The best and the second-best results are denoted in bold and underlined, respectively. Besides, PhotoMaker [38] is base on AnimateDiff [25] SDXL version.

significantly in the three metrics measuring subject fidelity: Face Similarity, CLIP-I, and DINO-I. For text alignment, our method achieves the best results among those using the AnimateDiff SD1.5 version as the base model. PhotoMaker uses the AnimateDiff SDXL version as its base model, which has a more powerful generative capability at the base model level. However, our method achieves comparable results, indicating that our approach of injecting subject information using the model's native capabilities can ensure high-fidelity subject appearance consistency while maintaining alignment between the generated video and the given prompt. Additionally, our method exhibits better dynamism.

C. User Study

To further validate the effectiveness of our method, we conducted a human evaluation comparison of our method and existing methods. For customized human video generation, we selected 10 celebrities and 10 non-celebrities as the test benchmark. For each individual, we used two prompts to generate videos. We invited 10 professionals to evaluate the methods. We evaluated the quality of the generated videos from four dimensions: Text Alignment, Subject Fidelity, Motion Alignment, and Overall Quality. Text Alignment evaluates whether the generated video matches the text prompt. Subject Fidelity measures whether the generated object is close to the reference image. Motion Alignment is used to evaluate the quality of the motions in the generated video. Overall Quality is used to measure whether the quality of the generated video overall meets user expectations. As shown in Figure 4, our method received significantly more user preference across various evaluation metrics. Additionally, it demonstrated a notable improvement in subject fidelity, thereby proving the effectiveness of our framework.

For customized object video generation, we conducted subjective evaluations on the 9 categories of objects included in the VideoBooth dataset. Each category provided one subject, and two prompts generated by ChatGPT [1] were used for testing. We similarly invited 10 professionals

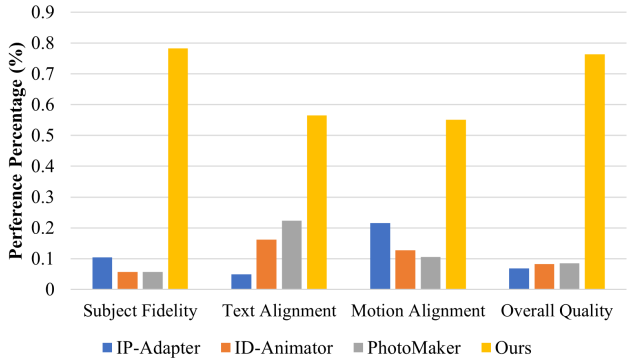

Figure 4. User Study for Customized Human Video Generation.

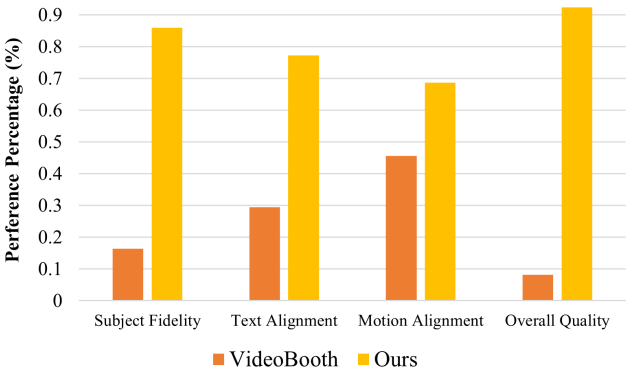

Figure 5. User Study for Customized Object Video Generation.

to evaluate the methods. As shown in Figure 5, our method received more favorable evaluations in all aspects compared to VideoBooth.

D. Limitations and Future Work

Our method only focuses on maintaining a single subject in the generated videos, and cannot control multiple subjects of generated persons in one video simultaneously. In addition, our method, which is based on AnimateDiff and the dataset we utilized, inherits certain biases and limitations from these sources.

Limitations of the base model. Our method is based on the SD1.5 version of AnimateDiff, and thus is limited by the generative capabilities of the base model. This can result in issues such as abnormal rendering of hands and limbs in the generated videos. Besides, since AnimateDiff inserts and fine-tunes Motion Blocks on the original image model, the base model's generated videos may exhibit poor dynamic effects, which in turn limits the dynamism of our method. Additionally, the base model has issues with facial clarity when the face is small in the generated images, affecting our customized portrait generation by failing to inject facial details well when the face occupies a smaller portion of the image. However, to ensure fair comparison with other methods and due to the limitations of our experimental equipment, we have not yet conducted experiments on better open-source models such as VideoCrafter [7, 9], CogVideoX [70], and Latte [44]. In the future, we will attempt to use more powerful base models to achieve better generative effects.

Limitations of the training datasets. For customized human video generation: The CelebV-Text [72] dataset mainly consists of half-body videos, resulting in the model we trained on this dataset performing poorly in generating full-body videos. Our method excels at generating half-body portrait videos but is relatively less proficient at generating full-body portrait videos. Additionally, due to the coarse-grained captions in the training data, fine-grained control is not achievable. For customized object video generation: The VideoBooth [32] dataset contains only a limited set of nine categories, so the model trained on this dataset cannot achieve truly universal generation of all objects. Furthermore, since the training videos for VideoBooth dataset are sampled from the WebVid [2] dataset, which contains watermarks, our customized object generation model trained on this dataset also results in generated videos with watermarks. In the future, we can attempt to train on better high-quality datasets to achieve truly universal zero-shot customized generation.

E. More Qualitative Comparison Results.

To further demonstrate the effectiveness of our method, we have supplemented additional visualizations for qualitative comparison. For customized human video generation, we first added some customized generation results for celebrities. As shown in Figure 6, our method exhibits stronger subject fidelity compared to existing zero-shot customization methods while ensuring text alignment. The videos generated by our method contain more facial details. For example, in Figure 6 (c), our method not only accurately depicts the action of "enjoying a cup of coffee" compared to other methods but also achieves high subject fidelity, maintaining the subject's appearance consistency where other methods fail to do so. Additionally, we further demonstrate the generation effects of our method on the non-celebrity dataset. As shown in Figures 7 and 8, our method can still achieve high-fidelity zero-shot customized generation on non-celebrity data, with better subject fidelity compared to existing methods. For example, in Figure 6 (f), our method accurately generates a video of the specified subject based on the reference image and text prompt, demonstrating a clear advantage over other methods.

For customized object video generation, the VideoBooth dataset we used for training contains nine categories of ob-

jects. Therefore, we supplemented qualitative comparisons for all nine categories. As shown in Figure 9, our method achieves significant improvements in both text alignment and subject fidelity compared to VideoBooth. As illustrated in Figure 9 (a, g), our method correctly generates the 'snowy' scene, whereas VideoBooth fails to generate the corresponding scene accurately. Additionally, in Figure 9 (i), our method correctly generates the scene of 'a field of wildflowers,' which VideoBooth does not. In terms of subject fidelity, our method shows significant improvements over VideoBooth. As shown in Figure 9 (a, c, d, e, f, g, h, i), for these animals, our method can accurately depict the texture details of the reference subject in large scenes, which VideoBooth fails to achieve.

F. Potential Societal Impacts

In this paper, we present VideoMaker, a novel framework that leverages the inherent force of VDM to achieve zero-shot customized generation. Compared to heuristic external models for subject feature extraction and injection, we cleverly use VDM to accomplish the extraction and injection of subject features required for customized generation, resulting in high-quality customized video generation.

In practical applications, our method can be used in the film or video game industry to directly generate some required film clips through customized video generation. It can also be applied in virtual reality to provide a more immersive and personalized experience.

However, we acknowledge the ethical considerations that come with the ability to generate high-fidelity videos of humans or objects. The proliferation of this technology could lead to the misuse of generated videos, infringing on personal privacy rights, and potentially causing a surge in maliciously altered videos and the spread of false information. Therefore, we emphasize the importance of establishing and adhering to ethical guidelines and using this technology responsibly.

Category	Prompt	Category	Prompt
bear	A bear walking through a snowy landscape. A bear walking in a sunny meadow. A bear resting in the shade of a large tree. A bear walking along a beach. A bear fishing in a rushing river. A bear running in the forest. A bear walking along a rocky shoreline. A bear drinking from a clear mountain stream. A bear standing on its hind legs to look around. A bear running on the grass.	car	A car cruising down a scenic coastal highway at sunset. A car silently gliding through a quiet residential area. A car smoothly merging onto a highway. A car driving along a desert road. A car speeding through a muddy forest trail. A car drifting around a sharp corner on a mountain road. A car navigating through a snow-covered road. A car driving through a tunnel with bright lights. A car driving through a beach. A car driving through a foggy forest road.
cat	A cat is perched on a bookshelf, silently observing the room below. A cat is sitting in a cardboard box, perfectly content in its makeshift fortress. A cat is curled up in a human's lap, purring softly as it enjoys being petted. A cat is circle around a food bowl in a room, patiently waiting for mealtime. A cat is lying on a windowsill, its silhouette framed by the setting sun. A cat is running on the grass. A cat is walking on a street. There are many buildings on both sides of the street. A cat is sitting in a window, watching the raindrops race down the glass. A cat is playing with a ball of wool on a child bed. A cat is playing in the snow, rolling and rolling, snowflakes flying.	dog	A dog is lying on a fluffy rug, its tail curled neatly around its body. A dog is walking on a street. A dog is swimming. A dog is sitting in a window, watching the raindrops race down the glass. A dog is running. A dog, a golden retriever, is seen bounding joyfully towards the camera. A dog is seen leaping into a sparkling blue lake, creating a splash. A dog is seen in a snowy backyard. A dog is seen napping on a cozy rug. A dog is seen playing tug-of-war with a rope toy against a small child.
elephant	An elephant walking through the jungle. An elephant crossing a river. An elephant walking on the grass. An elephant walking on a road. An elephant walking along a dirt road. An elephant playing in a mud pit. An elephant walking through a dense jungle. An elephant walking along a sandy beach. An elephant running through a meadow of wildflowers. An elephant running across a desert landscape.	horse	A horse walking through a dense forest. A horse running across a grassy meadow. A horse walking along a sandy beach. A horse running through a shallow stream. A horse walking on a mountain trail. A horse running across a desert landscape. A horse walking through a quiet village. A horse running in an open field. A horse walking along a forest path. A horse running through tall grass.
lion	A lion running along a savannah at dawn. A lion walking through a dense jungle. A lion running on a snowy plain. A lion running along a rocky coastline. A lion walking through a field of sunflowers. A lion running across a grassy hilltop. A lion walking through a grassland. A lion running along a riverbank. A lion walking on a savannah during sunrise. A lion running on a plain.	panda	A panda walking through a bamboo forest. A panda running on a grassy meadow. A panda running through a field of wildflowers. A panda walking through a snowy landscape. A panda walking through a city park. A panda walking in front of the Eiffel Tower. A panda wandering through a dense jungle. A panda running along a sandy beach. A panda exploring a cave. A panda is eating bamboo.
tiger	A tiger running along a savannah at dawn. A tiger walking through a dense jungle. A tiger running on a snowy plain. A tiger running along a rocky coastline. A tiger walking through a field of sunflowers.	tiger	A tiger running across a grassy hilltop. A tiger walking through a grassland. A tiger running along a riverbank. A tiger walking on a savannah during sunrise. A tiger running on a plain.

Table 3. Evaluation text prompts for customized object video generation.

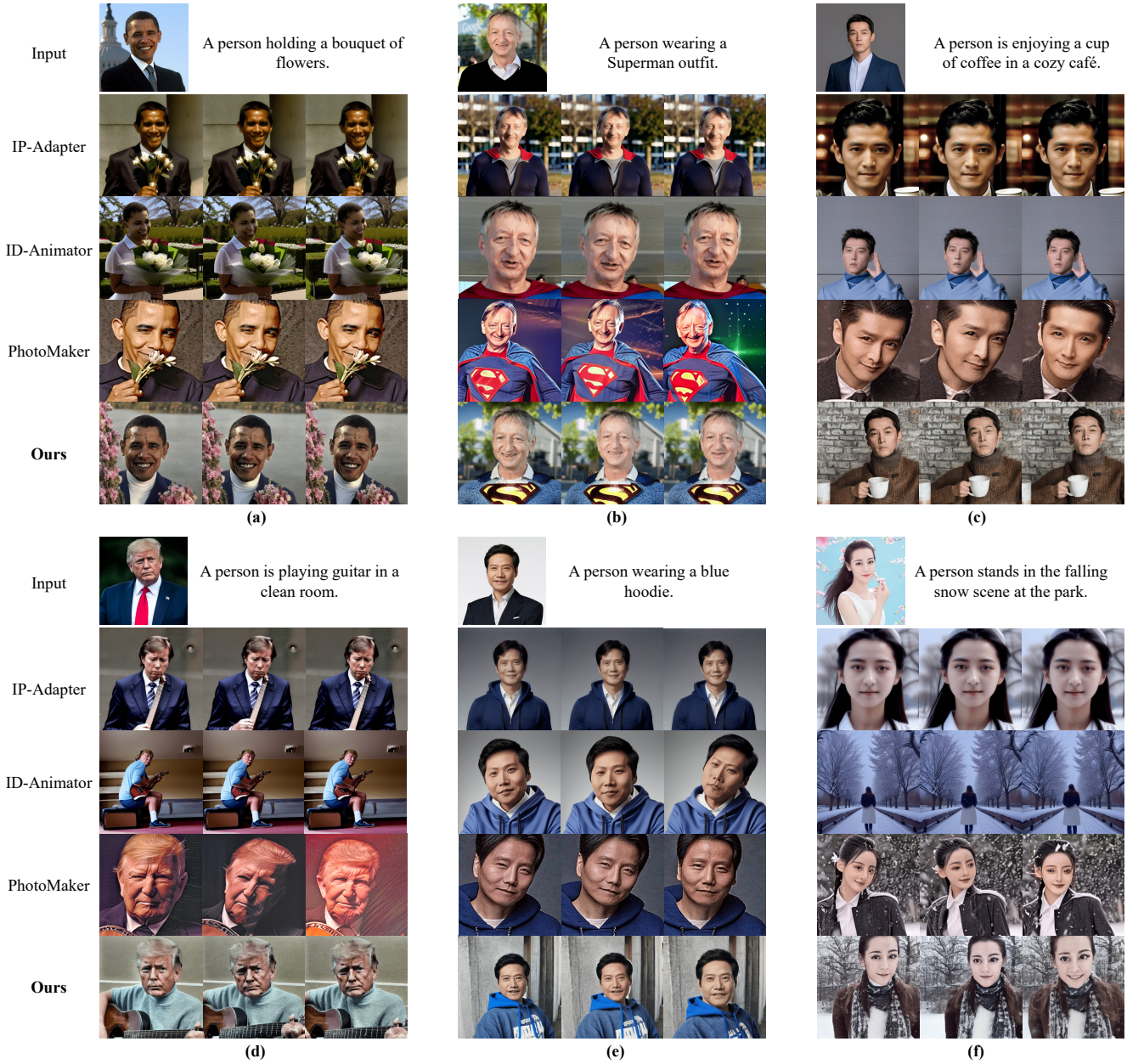

Figure 6. More Qualitative comparison for customized human video generation on celebrity dataset.

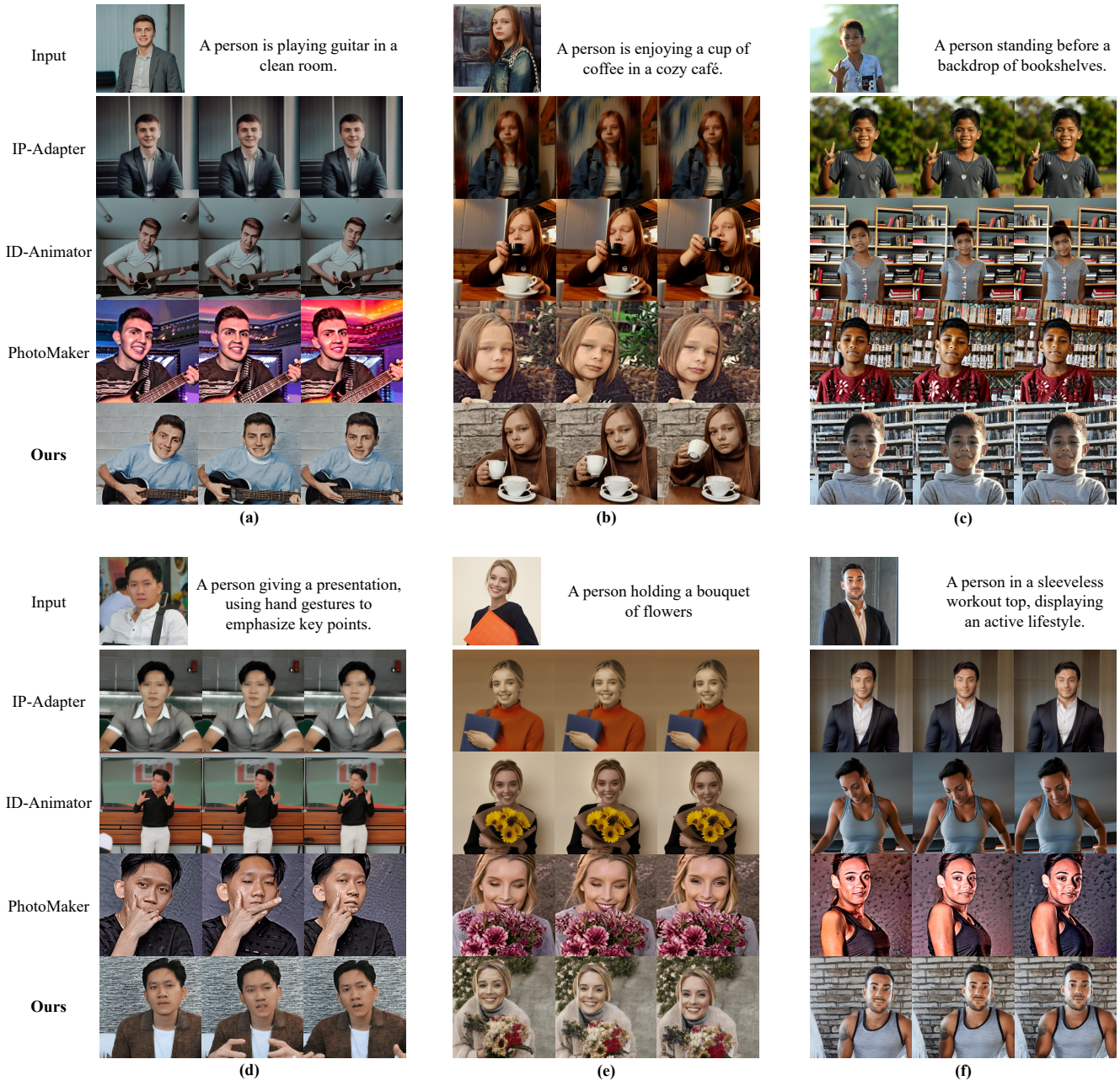

Figure 7. More Qualitative comparison for customized human video generation on non-celebrity dataset.

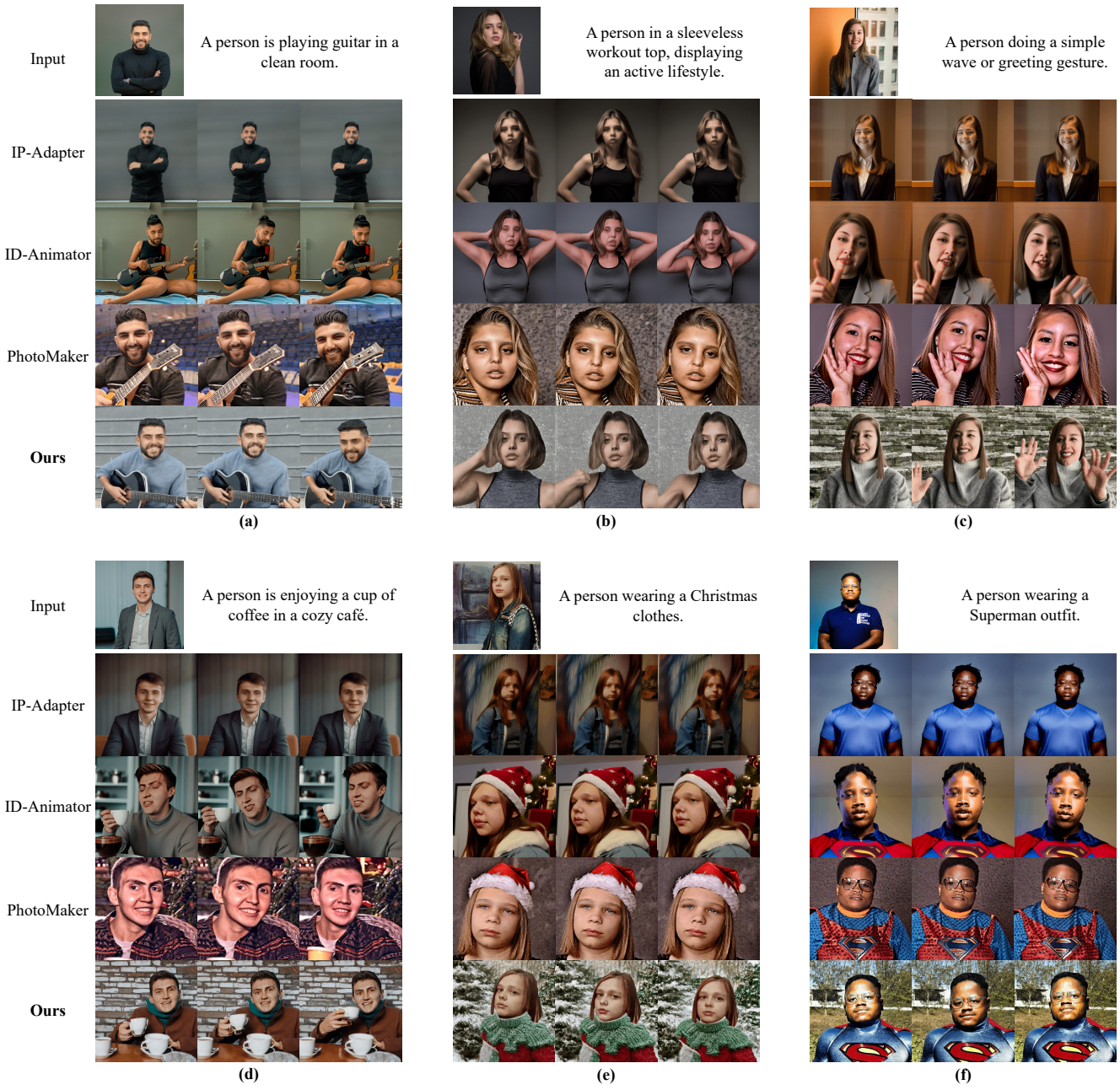

Figure 8. More Qualitative comparison for customized human video generation on non-celebrity dataset.

Figure 9. More Qualitative comparison for customized object video generation.