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.

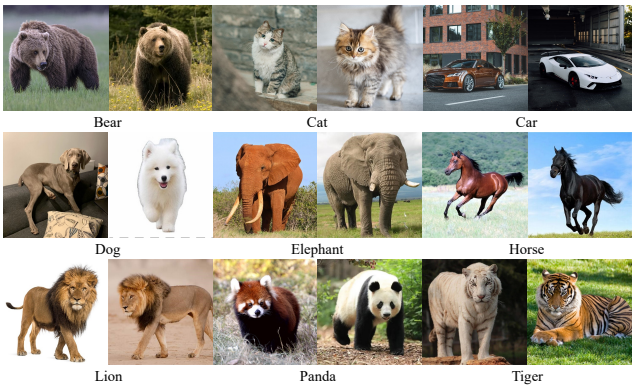

Figure 2. The overview of the dataset we use to test customized object video generation.

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

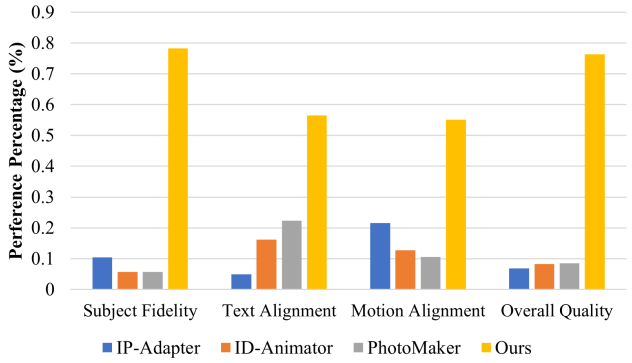

Figure 4. User Study for Customized Human Video Generation.

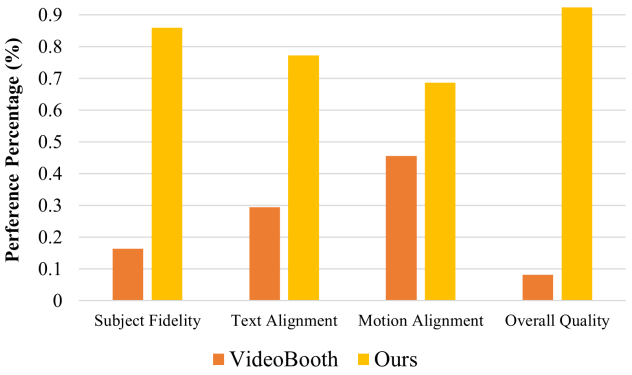

Figure 5. User Study for Customized Object Video Generation.

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.

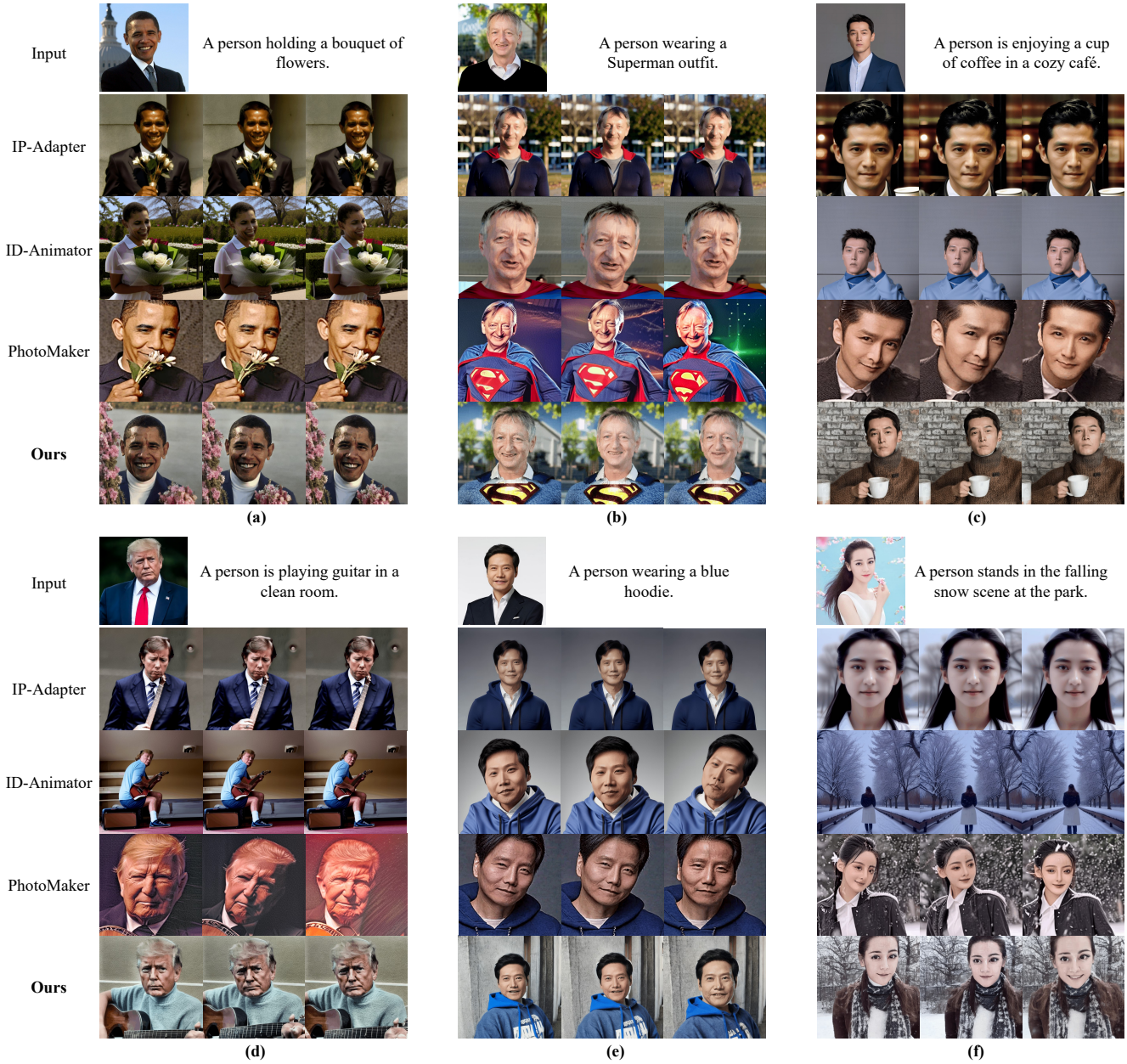

Figure 6. More Qualitative comparison for customized human video generation on celebrity dataset.

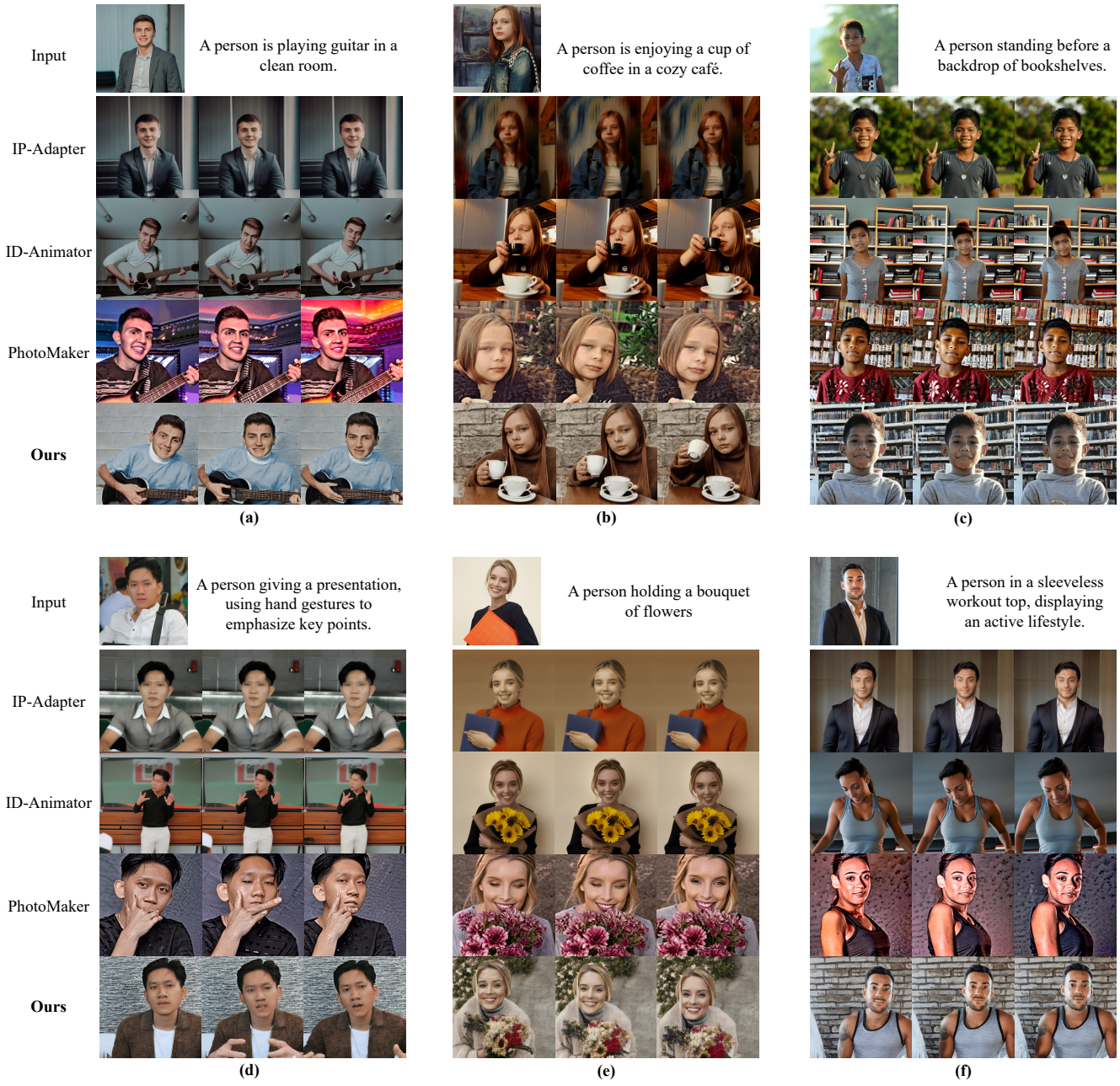

Figure 7. More Qualitative comparison for customized human video generation on non-celebrity dataset.

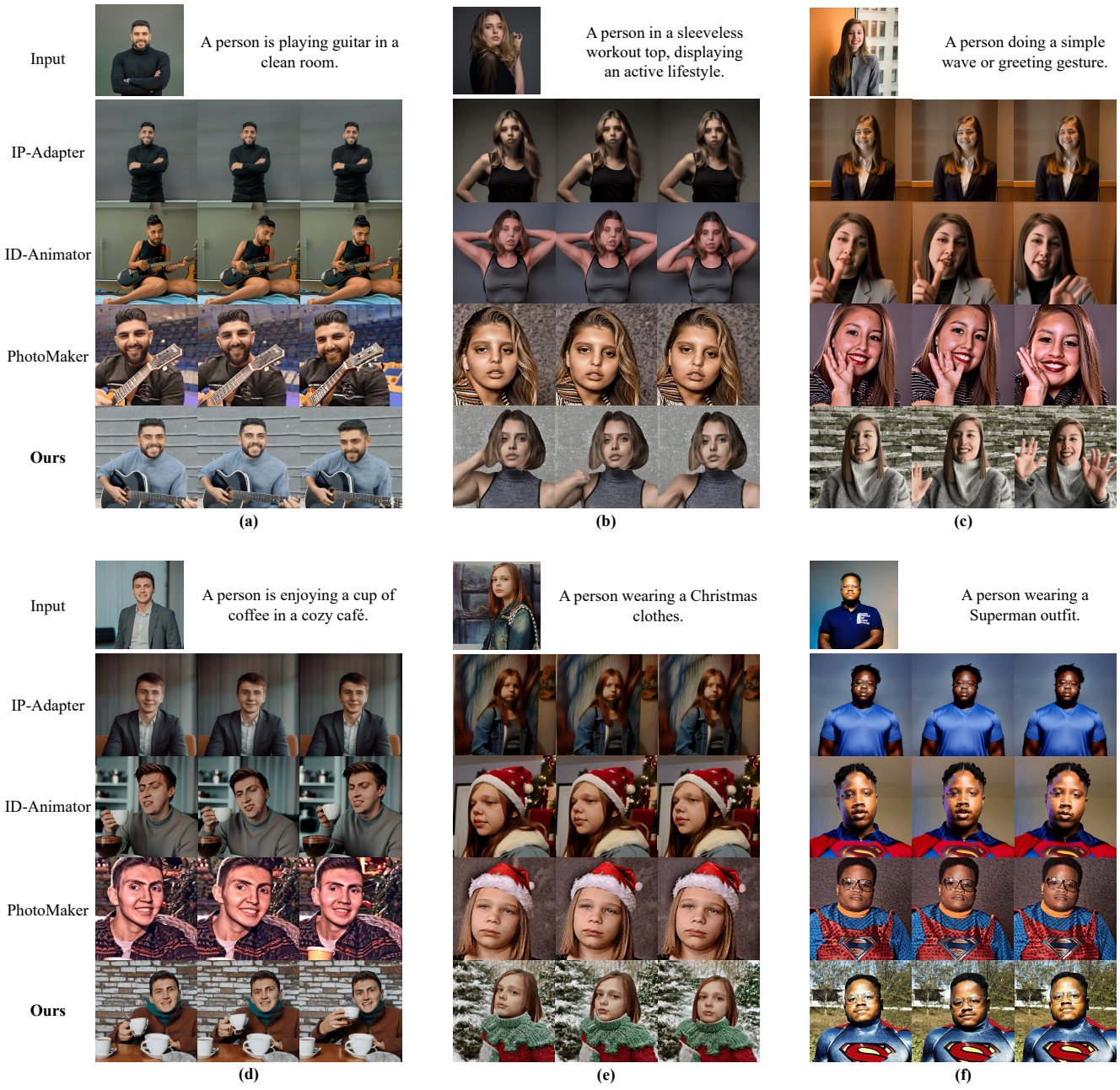

Figure 8. More Qualitative comparison for customized human video generation on non-celebrity dataset.

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