DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Human-Centric Video Editing **Supplementary Material**

Jia-Wei Liu[1], Yan-Pei Cao[3†], Jay Zhangjie Wu[1], Weijia Mao[1], Yuchao Gu[1], Rui Zhao[1], Jussi Keppo[2], Ying Shan[3], Mike Zheng Shou[1†]

[1]Show Lab, [2]National University of Singapore [3]ARC Lab, Tencent PCG

The supplementary material is structured as follows:

- Sec. 1 presents implementation details on the network designs and optimization parameters of DynVideo-E.
- Sec. 2 summarizes additional comparisons and ablations of our DynVideo-E against SOTA approaches.

Furthermore, we provide a **supplementary video** showcasing all 24 edited video comparisons of our method against baselines, as well as 360° free-viewpoint renderings of edited dynamic scenes from our DynVideo-E.

1. Implementation Details

DynVideo-E Network Details. As shown in Fig. 1, we employ a 10-layer multilayer perceptron (MLP) as our state-conditional background network (a) and a 8-layer MLP as our state-conditional canonical human-object network (b). To edit the dynamic human, we establish a 9-layer canonical human network (c) where the parameters of its first 8 layers are initialized from the reconstructed human-object model (b). During optimization, we train the 3D background model (a) and the 3D dynamic human model (c) while freeze the reconstructed dynamic human-object model (b). During inference, for the source video that contains dynamic objects, we query the original dynamic human-object model (b) for the pixels within the object masks to keep the dynamic objects, while we query the edited dynamic human model (c) and edited background model for other pixels to obtain the colors and densities of edited contents. For the human-background videos, we only need to query the edited dynamic human model and edited background model to obtain the edited contents.

Optimization Parameters. We optimize our DynVideo-E using Adam optimizer [5]. We set the learning rate for our training process as 0.0005 with 20000 training iterations. We balance the loss terms using the following weighting factors: $\lambda_{\text{rgb}} = 5$, $\lambda_{\text{mask}} = 0.5$, $\lambda_{\text{depth}} = 0.01$, $\lambda_{\text{3D}} = 40$, $\lambda_{\text{2D}} = 1.0$, $\lambda_{\text{NNFM}} = 1.0$. The guidance scale of the 3D diffusion prior and 2D personalized diffusion prior are set to 5 and 20, respectively. We conducted all our experiments on 1 NVIDIA A100 GPU, using the PyTorch [8] deep learning framework.

Visualization of Text-guided Local Parts Super-Resolution. To improve the effective resolution during training, we utilize the text-guided local parts super-resolution to render and supervise the local parts of zoom-in humans and augment with view-conditional prompts. We provide 8 visualization examples of text-guided local parts super-resolution sampled during training in Fig. 2. As shown in Fig. 2, even though all figures are rendered in (128×128) resolutions, rendering local parts can largely improve the effective resolution and thus we can supervise the detailed geometry and textures of edited human body with diffusion priors.

2. Additional Results

More Qualitative Results. We present two more visual comparisons of our approach against all baselines in Fig. 3 and Fig. 5. As shown in the figures, our DynVideo-E achieves the best performances with photo-realistic edited videos, which clearly demonstrates the superiority of our model against other approaches on editing large-scale motion- and view-change human-centric videos. Comparing the long (a) and short (b) video editing results of Fig. 3, we find that baseline approaches perform better on short videos than long videos, but still none of them can edit the correct subject "Thanos" due to the large subject motions and viewpoint changes in videos. In contrast, our DynVideo-E produces high-quality editing results on both short and long videos. Please refer to our supplementary video for all 24 edited video comparisons of our method against baselines.

Additional Ablation Results. We conduct ablation studies on more videos from HOSNeRF dataset [6] and NeuMan dataset [3]. To evaluate the effectiveness of each proposed component in DynVideo-E, we progressively ablate

† Corresponding Authors.

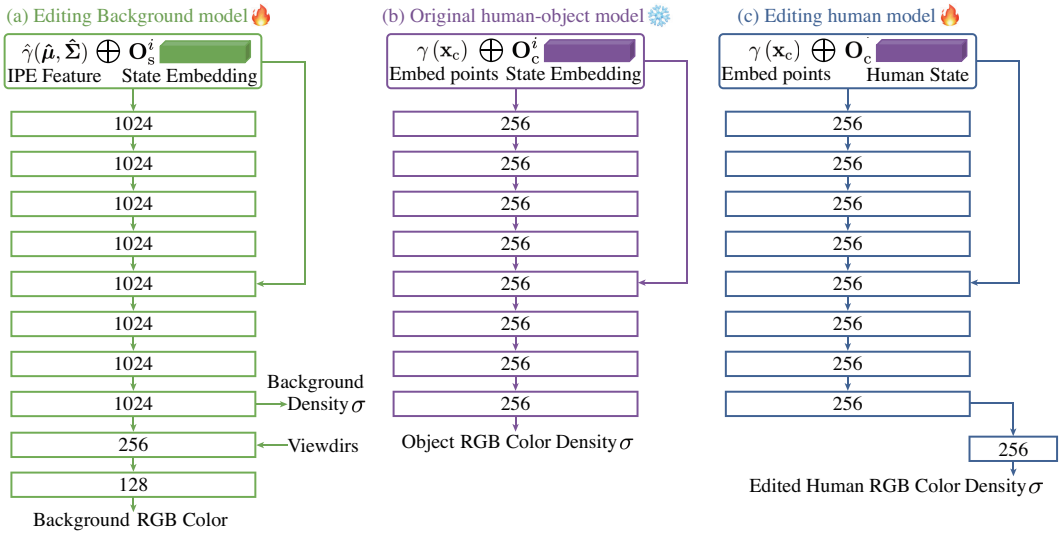

Figure 1. DynVideo-E network designs: (a) Editing Background model, (b) Original human-object model, (c) Editing human model.

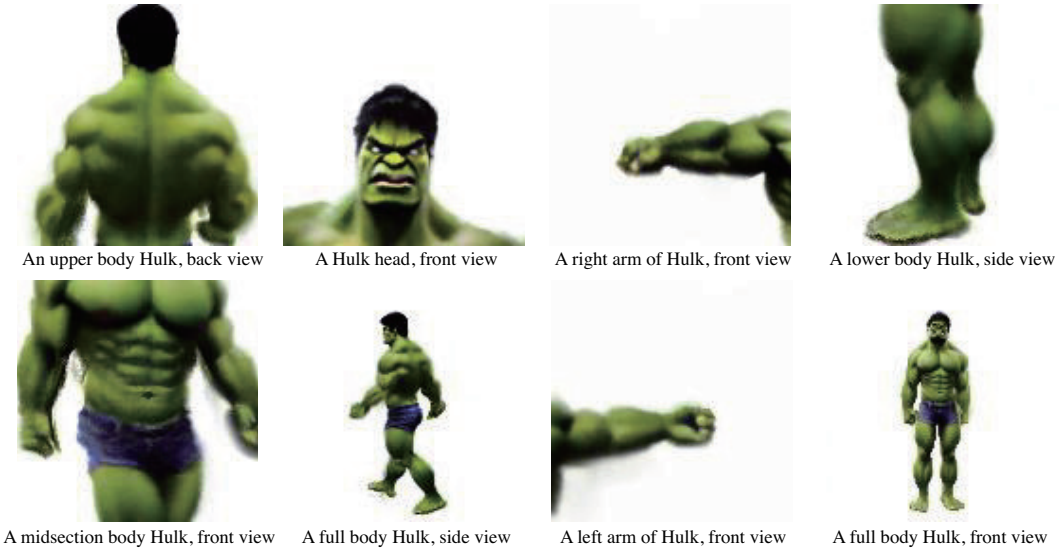

Figure 2. Visualization examples of text-guided local parts super-resolution sampled during training.

Ablation components	Average CLIP Score
Full model	**0.674**
w/o Super-solution	0.659
w/o Super-solution, Rec	0.650
w/o Super-solution, Rec, 2D SDS	0.572
w/o Super-solution, Rec, 3D SDS	0.641
w/o Super-solution, Rec, 3D SDS, 2D LoRA	0.593

Table 1. Averaged quantitative ablation results of our method.

each component from local parts super-resolution, reconstruction loss, 2D personalized SDS, 3D SDS, and 2D personalization LoRA. We observe that the model even fails to converge on some videos when we disable several components of our model. We compute the average CLIP score of all successfully edited videos in Tab. 1, where the CLIP score progressively drops with the disabling of each component, with the full model achieving the best performances, which clearly demonstrates the effectiveness of our designs.

Visualization of Canonical Images from CoDeF [7] and Atlas from Text2LIVE [1] and StableVideo [2]. For challenging videos with large-scale motions and viewpoint changes, CoDeF [7], Text2LIVE [1], and StableVideo [2] largely overfit to input video frames and learn meaningless canonical images or neural atlas, and thus cannot generate meaningful editing results. We show several examples of their learned canonical images [7] and neural atlas [1, 2] in Fig. 6, where Text2LIVE [1] and StableVideo [2] utilizes the same foreground and background atlas during editing. As shown in Fig. 6, canonical images and atlas all fail to represent the challenging large-scale motion- and view-change videos, and thus they cannot generate satisfactory editing results. In addition, the atlas performs better for short videos in NeuMan dataset [3] than

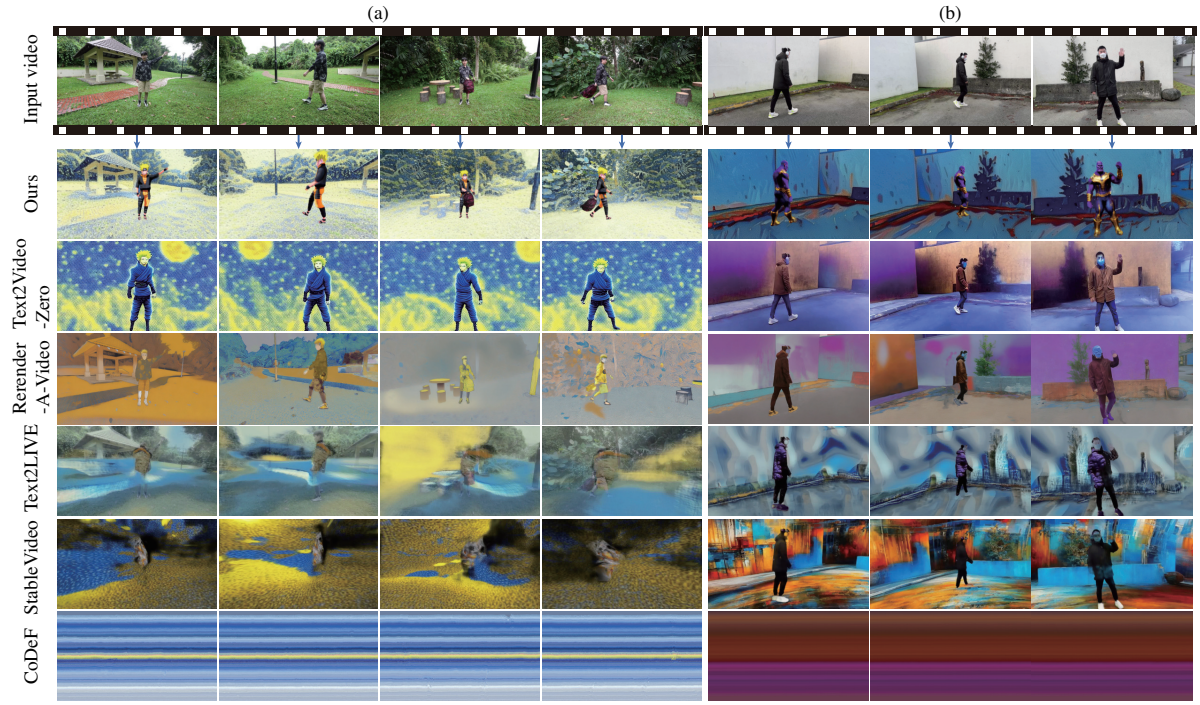

Figure 3. More qualitative comparisons of DynVideo-E against SOTA approaches on the Backpack scene (a) and Parkinglot scene (b).

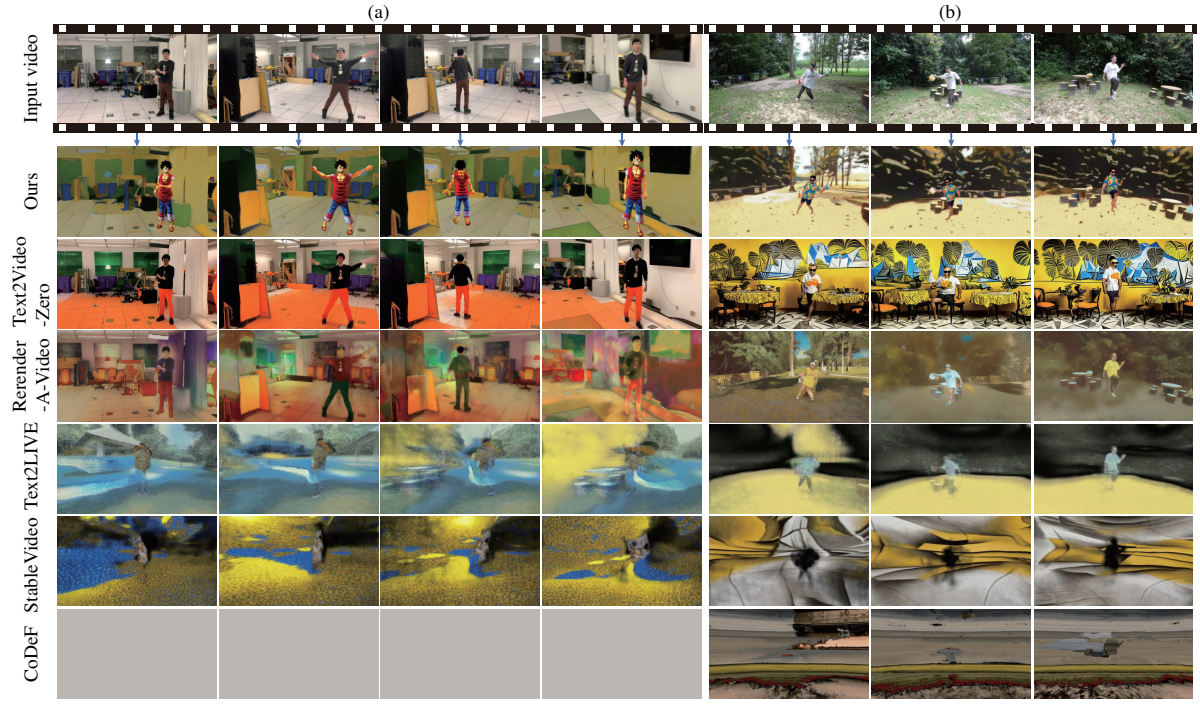

Figure 4. More qualitative comparisons of DynVideo-E against SOTA approaches on the Lab scene (a) and Dance scene (b).

long videos with a better background atlas, but the foreground atlas still cannot represent the humans with large motions. In contrast, our DynVideo-E represents videos with the dynamic NeRFs to effectively aggregate the large-scale motion- and view-change video information into a 3D dynamic human space and a 3D background space, and achieves high-quality video editing results by editing the 3D dynamic spaces.

Editing Operation Time Comparison. We compare the editing operation time of our DynVideo-E against other approaches on a long video of the HOSNeRF dataset ($[300, 400]$ frames) using a single A100 GPU in Tab. 2.

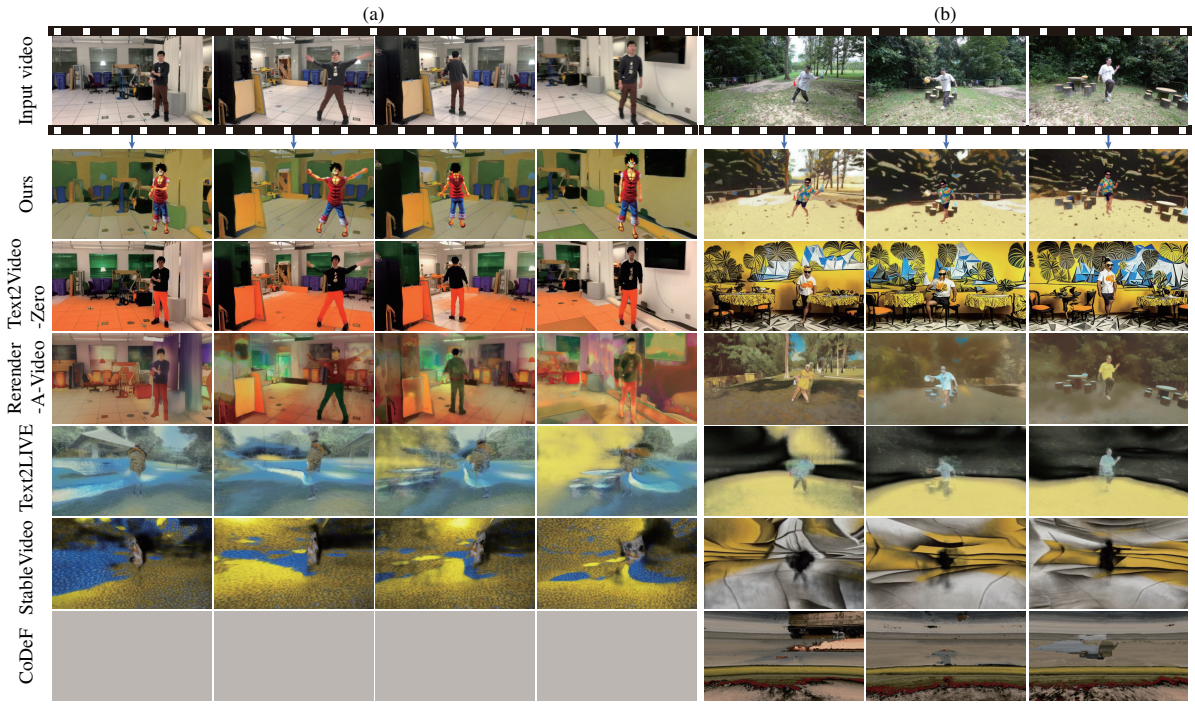

Figure 5. More qualitative comparisons of DynVideo-E against SOTA approaches on the Lab scene (a) and Dance scene (b).

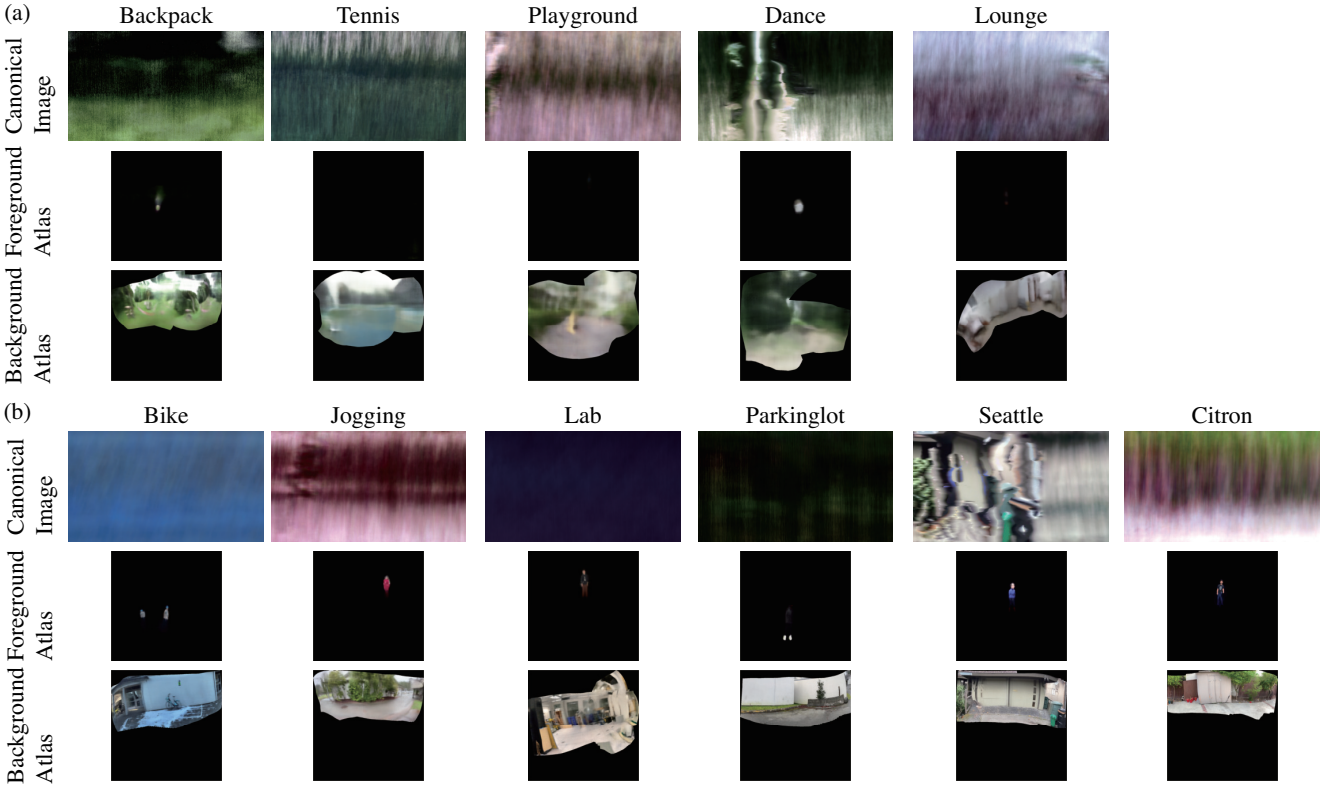

Figure 6. Visualization of canonical images from CoDeF [7], and foreground and background atlas from Text2LIVE [1] and StableVideo [2] on (a) HOSNeRF dataset [6] and (b) NeuMan dataset [3].

Although other approaches are faster than ours, 2D-video representation-based methods such as CoDeF [7], StableVideo [2], and Text2LIVE [1] cannot accurately reconstruct large-scale motion- and view-change videos and thus fail to

Method	CoDeF [7]	Text2Video-Zero [4]	Rerender-A-Video [9]	StableVideo [2]	Text2LIVE [1]	DynVideo-E (Ours)
Time	$\sim$ 1 mins	15 mins	1.2 hrs	$\sim$ 1 mins	$\sim$ 2 hrs	7.3 hrs

Table 2. Editing operation time comparison of our method against other approaches.

generate meaningful editing results, as validated in Fig. 6. Text2Video-Zero [4] and Rerender-A-Video [9] fail to edit the challenging human-centric videos with large-scale motion and viewpoint changes and their editing results are highly inconsistent. Therefore, previous approaches cannot handle the challenging human-centric videos no matter how many computation resources are provided. In contrast, our method is the first work to achieve highly consistent long-term video editing that outperforms previous approaches by a large margin of $50\% \sim 95\%$ in terms of human preference, and we leave accelerating our model with voxel or hash grid representation as a faithful future direction.

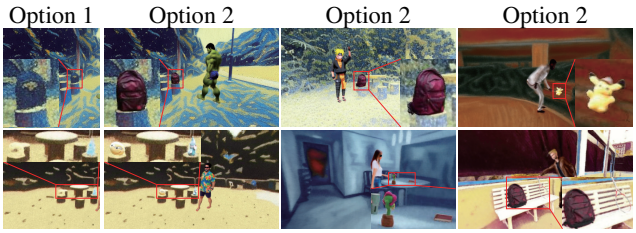

Figure 7. 2 options to render interacted objects in background.

Interacted Objects in Background. As shown in Fig. 7, we provide two options to render the interacted objects in background: using the edited background model (Option 1); or using the original background model with the object masks obtained during the camera pose calibration process (Option 2). Since we integrate the edited model and original model into a single model, we can conveniently switch between these two options. We show results from Option 1 in the main paper.

Example of Human Preference Questionnaire. We utilize Amazon MTurk * to recruit raters to rate our pairwise comparing videos. For each comparison, we show our result and one baseline result (shuffled order in questionnaires), together with textual descriptions to raters and ask their preferences. In total, we collected 1140 comparisons over all pairwise results from 32 different raters. Fig. 8 illustrate one comparison example in our questionnaires.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 2, 4, 5

[2] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *arXiv preprint arXiv:2308.09592*, 2023. 2, 4, 5

[3] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022. 1, 2, 4

[4] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 5

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Eric Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. *arXiv preprint arXiv:2304.12281*, 2023. 1, 4

[7] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 2, 4, 5

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[9] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 5

* https://requester.mturk.com/

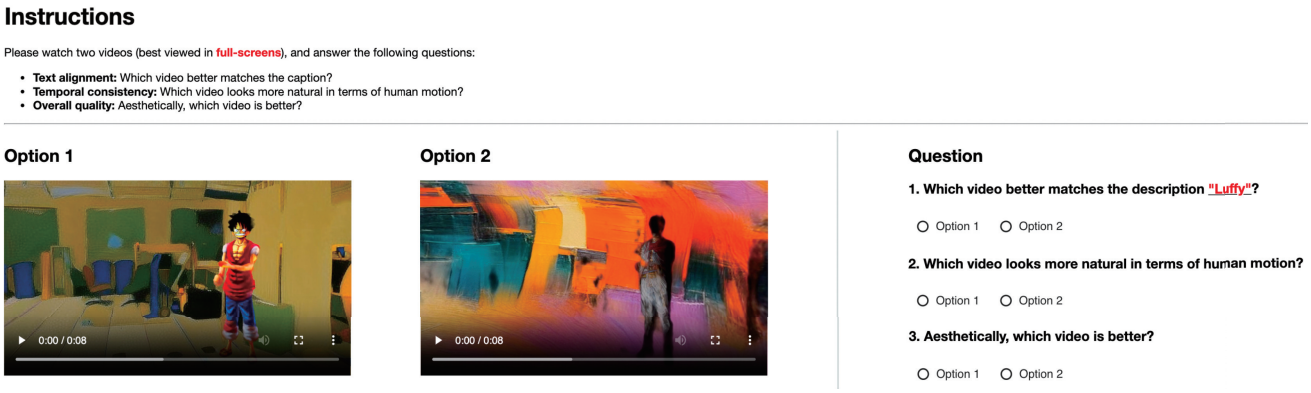

Figure 8. One comparison example from our questionnaires.