Technical Appendix for DreamSwapV

Anonymous submission

1 Overview

This technical appendix includes additional details and information about our methods and experiments which cannot be fully covered in the main paper limited by space. We first strongly recommend readers to watch the video demo (Appendix 2) provided along with this appendix in the supplementary material, where we dynamically showcase our entire pipeline and present the video comparison results. Then we complement our full implementation details for reproducibility, including hyper-parameters, benchmark contents, and baseline adaptations (Appendix 3). We further provide additional experimental results which cannot be fully displayed in the main paper(Appendix 4). At last, we provide a comprehensive discussion, including the more applications, social impact, limitations and future work(Appendix 5).

2 Video Demo

Our video demo is composed of the following elements:

- The brief presentation summarizing the main paper.
- Dynamic pipeline and data flow, which brings a clearer understanding of our proposed method.
- Video comparisons of our method and four baselines across diverse subject categories and aspect ratios.
- More application examples demonstrating the strong extensibility of our method.

3 Implementation Details

3.1 Specific Hyper-parameters

The specific hyper-parameters used in our two-phase training scheme are listed in Table 1 for reproducibility.

Hyper-parameters	**Pre-training**
Iterations	15000
Learning Rate	2e-5
LR Scheduler	`constant_with_warmup`
Batch size	1
Optimizer	AdamW
Resolution	1280×720
Frame Length	65
CFG Drop Ratio	0.15
Image Mix Ratio	0.3
Flow Shift	5.0
h_1, h_2, h_3	1, 30, 20
	Quality Tuning
Iterations	10000
Learning Rate	1e-5
LR Scheduler	`constant_with_warmup`
Batch size	1
Optimizer	AdamW
Resolution	1280×720, 720×1280
Frame Length	69
CFG Drop Ratio	0.15
Image Mix Ratio	0.1
Flow Shift	5.0
h_1, h_2, h_3	1, 30, 20

Table 1: Hyper-parameters in our two-phase training.

3.2 DreamSwapV-Benchmark Contents

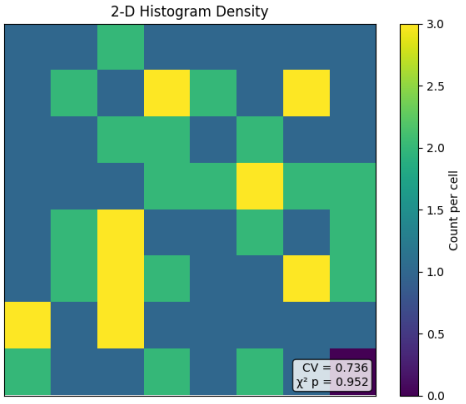
Benchmark Distribution. The subjects and their reference images for swapping from DreapSwapV-Benchmark are carefully selected to ensure a wide and balanced distribution across categories and complexities.

Figure 1(a) is the 2-D histogram density of all reference images after t-SNE (Maaten and Hinton 2008) projection: each colour block represents the number of images falling into a 8×8 grid map on the embedding plane. The cell-count coefficient of variation is 0.736 and the χ^2 goodness-of-fit test gives p = 0.952 (> 0.05), showing that only a very small fraction of grid cells is empty or over-populated—i.e., the images cover the feature space almost uniformly.

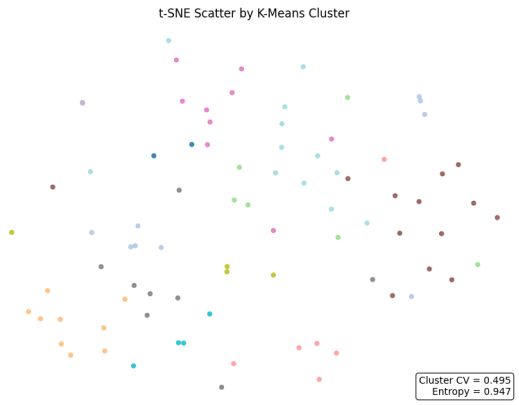
Figure 1(b) is the t-SNE scatter plot coloured by K-Means clusters (k = 12) produced in the original 2048-D feature space. The clusters have a size CV of 0.495 and a normalised Shannon entropy of 0.947 (~ 1), meaning the images are evenly split across latent semantic groups with no dominant or under-represented cluster.

Together, the two figures verify that our benchmark set is both spatially well-covered and semantically balanced, preventing bias in subsequent experiments.

Metric Calculation. We adopt five indicators inherited from VBench (Huang et al. 2024)—subject consistency, background consistency, dynamic degree, motion smoothness, and aesthetic quality—and design two other automatic metrics specifically for the subject swapping task: reference ap-



(a) 2-D histogram density of reference images (t-SNE plane).



(b) t-SNE scatter coloured by K-Means clusters (k = 12).

Figure 1: Benchmark distribution analysis. (a) shows spatial coverage and (b) shows semantic balance across clusters.

pearance and background preservation. The specific meanings and calculations of each metric are listed below:

(i) Subject consistency: The indicator measures temporal subject-identity consistency—i.e., how stable the visual appearance of the main subject remains from frame to frame within a single video. Let $d_t \in \mathbb{R}^D$ denote the ℓ_2-normalised DINO (Caron et al. 2021) feature of frame t in a T-frame video. Temporal subject identity stability is quantified by

$$S_{\text{subject}} = \frac{1}{T-1} \sum_{t=2}^{T} \frac{\langle d_1, d_t \rangle + \langle d_{t-1}, d_t \rangle}{2},$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity. The score is averaged over all test videos of a model; larger values indicate greater subject consistency.

(ii) Background consistency: This indicator measures the temporal consistency of the *background identity* throughout a video. For each of the T frames, a CLIP image encoder (Radford et al. 2021) yields a unit-normalised feature vector c_t. Temporal background consistency is then

$$S_{\text{background}} = \frac{1}{T-1} \sum_{t=2}^{T} \frac{\langle c_1, c_t \rangle + \langle c_{t-1}, c_t \rangle}{2},$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. The model-level score is obtained by averaging $S_{\text{background}}$ across all evaluation videos; larger values indicate a more consistent background.

(iii) Dynamic degree: This dimension gauges the tendency of a model to produce *non-static* videos. For each generated clip v with T frames, RAFT (Teed and Deng 2020) is employed to estimate dense optical flow between consecutive frames, yielding magnitudes $\{m_{v,t}\}_{t=1}^{T-1}$. Let $\lambda_v = \text{mean}\big(\text{top-5\%}\{\ m_{v,t}\ \}\big)$ denote the average of the largest 5% flow magnitudes, which is sensitive to small-object motion. A clip is classified as *dynamic* if $\lambda_v > \tau$, where τ is a fixed empirical threshold. The model's dynamic degree is the proportion of dynamic clips:

$$S_{\text{dynamic}} = \frac{1}{N} \sum_{v=1}^{N} \mathbf{1}\big[\lambda_v > \tau\big],$$

with N the number of evaluation videos and $\mathbf{1}[\cdot]$ the indicator function. Higher values imply a stronger propensity to generate motion.

(iv) Motion smoothness: Smooth, physically plausible motion is assessed via a frame-interpolation prior. Given a $(2n{+}1)$-frame clip $\{f_t\}_{t=0}^{2n}$, every second frame is removed to form the low–frame-rate sequence $\{f_{2k}\}_{k=0}^{n}$; a pretrained interpolator then predicts the missing frames $\{\hat{f}_{2k+1}\}_{k=0}^{n-1}$. Temporal smoothness is quantified by the complement of the normalised mean absolute error (MAE) between the predicted and original odd frames:

$$S_{\text{motion}} = 1 - \frac{1}{n} \sum_{k=0}^{n-1} \frac{\|f_{2k+1} - \hat{f}_{2k+1}\|_1}{C},$$

where C is the normalisation constant (as in Eq. 4) that bounds $S_{\text{motion}} \in [0, 1]$. Averaging S_{motion} across all test videos yields the model-level score; larger values indicate smoother motion.

(v) Aesthetic quality: Photographic composition, colour harmony, and general artistic appeal are assessed with the LAION image-aesthetics predictor (LAION-AI 2022), which returns a raw per-frame score $a_t \in [0, 10]$ for each of the T frames. Scores are linearly mapped to $[0, 1]$ and averaged:

$$S_{\text{aesthetic}} = \frac{1}{T} \sum_{t=1}^{T} \frac{a_t}{10}.$$

Averaging $S_{\text{aesthetic}}$ across all clips gives the model-level score; larger values indicate higher aesthetic quality.

(vi) Reference appearance: While the VBench *subject consistency* metric verifies that the swapped subject remains visually stable across video frames, it does not focus how faithfully that subject mirrors the *external reference image* provided by the user. We therefore introduce the reference appearance metric to measure appearance fidelity. The same DINO backbone used in *subject consistency* is employed, but the calculation is changed: each video frame is compared directly to the reference image rather than to other frames.

Let d_t be the ℓ_2-normalised DINO feature of frame t in a T-frame clip and d_r the feature of the reference image. The

metric is the mean cosine similarity

$$S_{\text{reference}} = \frac{1}{T}\sum_{t=1}^{T}\langle d_r, d_t\rangle,$$

which lies in $[0, 1]$. Higher values indicate that the swapped subject in the video retains the color, texture, and overall appearance of the reference image more closely. By complementing subject-consistency, this metric allows a comprehensive assessment of both temporal stability and appearance fidelity in video subject-swapping tasks.

(vii) Background preservation: The VBench *background-consistency* metric measures intra-video background identity but cannot reveal how much the scene diverges from the *original* footage after video subject swapping. To capture that difference, we supply with the background preservation metric directly comparing each swapped frame to its source counterpart outside the tracked object mask.

For every time step $t \in \{1, \ldots, T\}$, let U_t and V_t be the original and swapped frames, and $\mathbf{M}_t^s$ denote the background pixels (the complement of the user-provided mask). The peak signal-to-noise ratio on the background region is $P_t = \text{PSNR}(U_t \odot \mathbf{M}_t^s, V_t \odot \mathbf{M}_t^s)$ (measured in dB). A monotone rescaling maps PSNR to $(0, 1)$ and is averaged across the video frames:

$$S_{\text{preservation}} = \frac{1}{T}\sum_{t=1}^{T}\frac{P_t/50}{P_t/50 + 0.6}.$$

Higher values indicate that the background is better preserved after swapping the subject, whereas lower scores signal unintended scene alterations.

User Study Details. To further examine human preferences of our method and four baselines, we conduct a user study from three aspects:

- *Reference detail* instructs annotators to judge how faithfully the fine-grained details of the reference image are retained after subject swapping, paralleling the automatic metric *reference appearance*.
- *Subject interaction* focuses on the interaction between the inserted subject and its surroundings, assessing both the visual plausibility of that interaction and whether it remains strictly confined to the user-specified region.
- *Visual fidelity* gauges the overall perceptual quality of the video—how realistic it appears and whether any synthetic or *AI-generated* artifacts are discernible.

15 annotators are carefully selected: five specialists in video generation or editing, five practitioners or students from other areas of artificial intelligence (AI), and five individuals with no AI background. The streamlit (Khorasani, Abdou, and Fernández 2022) annotation interface (Figure 2) presents each rater with five videos at a time. For each metric, the annotators are required to produce a strict ranking of the five videos from first to fifth, with no ties permitted. The rankings are then converted into numerical scores using a weighted Borda-count (Emerson 2013) scheme: 1st = 5 points, 2nd = 3 points, 3rd = 2 points, 4th = 1 point, and 5th = 0 points. Scores are aggregated per metric and per method,

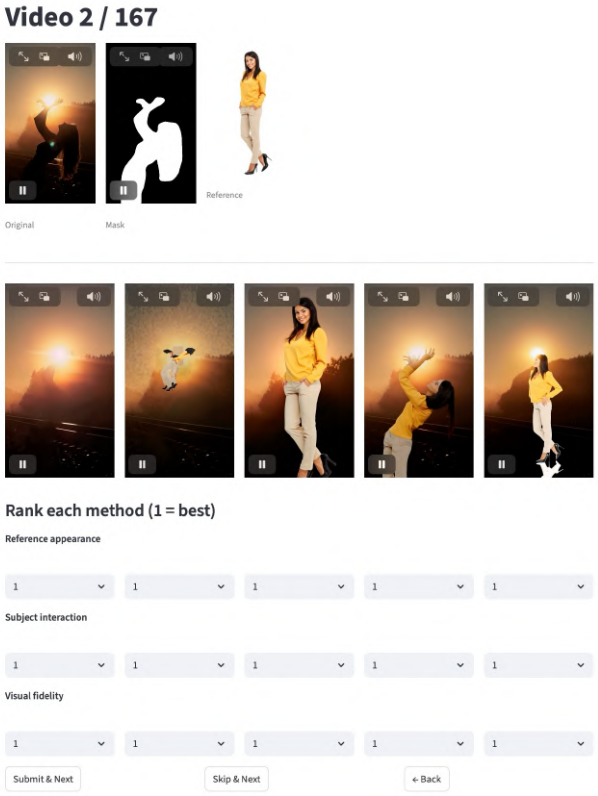


Figure 2: The streamlit annotation interface of user study.

and finally averaged over the 15 annotators to yield the overall user-study results for each method.

3.3 Baseline Adaptations

We select four baselines: AnyV2V (Ku et al. 2024), VACE (Jiang et al. 2025), HunyuanCustom (Hu et al. 2025) and Kling 1.6 Multimodal (Keling 2025) for their robust stability, high accessiblity, and close relevance to our subject swapping objective.

Specifically, for AnyV2V, we edit the first frame as required using the latest image subject swapping method InsertAnything (Song et al. 2025), and utilize AnyV2V's default I2VGen-XL pipeline to obtain subsequent edited frames. For VACE and HunyuanCustom, we preprocess the mask sequences as specified by each method to ensure their best performance. For Kling 1.6 Multimodal, we use its official in-browser masking tool to define the source subject. Notably, this tool occasionally struggle with specific selection (e.g. garment inside the human region), limiting the number of valid annotations. We process as many instances from our benchmark as possible, yielding a total of 152 Kling results. Since these four methods all support text inputs but not solely rely on them, we provide the text prompts by briefly describing of the subject-reference pair.

PikaSwaps (Pika 2025) is also a commercial model specializes in video subject swapping, but its API is consistently unavailable due to *high demand* during our whole evaluation period. If the API returns to normal operation, we will incorporate its results into benchmark at the earliest opportunity.

Ablation / Metrics	Video Quality & Video Consistency								
	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	VBench Average	Reference Appearance	Background Preservation	Total Average
w/o our reference injection									
- *channel concat*	93.87%	93.24%	99.11%	48.60%	54.17%	77.80%	37.34%	51.57%	68.27%
- *cross-attention*	95.49%	94.21%	99.29%	52.67%	56.41%	79.61%	39.23%	51.84%	69.88%
w/o our adaptive grid sizing	96.07%	94.24%	99.31%	54.37%	56.39%	80.08%	39.29%	**52.70%**	70.34%
Full (ours)	**96.41%**	**94.26%**	99.31%	**55.69%**	**56.52%**	**80.44%**	**45.22%**	52.49%	**71.41%**

Table 2: Quantitative comparison of different ablation settings (w/o reference injection or w/o adaptive grid sizing) with our full version model. The **bold** are the best results under certain metrics, and gray marks the full version model.

4 Additional Experiments and Results

4.1 Quantitative Ablation Study

Table 2 reports the ablation experiment results on our DreamSwapV-Benchmark, complementing the qualitative ablation study in the main paper. We train each ablation setting for the same iterations as ours; only the reference injection scheme or the adaptive grid sizing is altered, with all other settings held constant.

Our full version model achieves the best score on every automatic metric except *background preservation*. The slight drop in that metric is expected: the adaptive grid may deliberately enlarge the user-specified mask to improve generalizability across subject scales, thereby marginally reducing background preservation. Overall, the full configuration offers the most comprehensive and stable performance, confirming the effectiveness of both the proposed reference injection scheme of the condition fusion module and the adaptive grid sizing mechanism.

4.2 Qualitative Visual Comparisons

Figures 4-6 present extensive visual comparisons between DreamSwapV and the four baselines, spanning subject categories from *human* and *garment* to *small and large object*. Across all cases, our method delivers superior visual fidelity and stability, corroborating the state-of-the-art performance reported in the main paper.

5 Further Discussion

5.1 More Applications

As noted in the main paper, our method can be extended to numerous related or downstream tasks, which we will analyze individually below.

Image Subject Swapping. Our framework is inherently image-compatible: a single image is treated as a frame=1 video. Figure 3(a) confirms that this simple adaptation yields high-fidelity image subject swapping.

Video Inpainting or Addition. As discussed in Section 3, slight changes in the input modality transform video subject swapping into video inpainting or video addition:

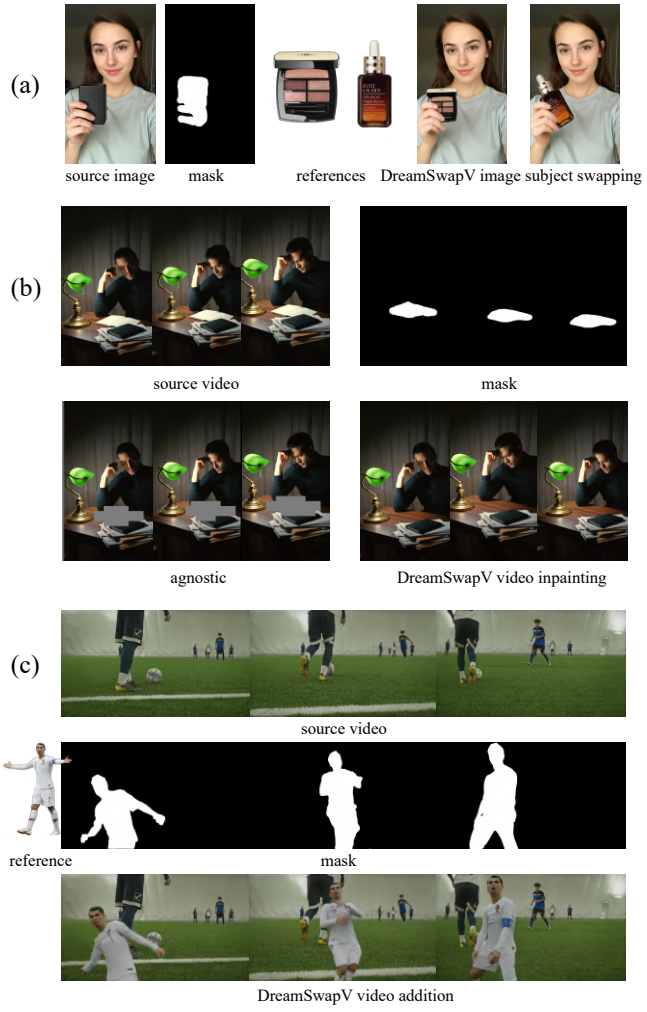


Figure 3: More applications of our DreamSwapV on (a) image subject swapping, (b) video inpainting, and (c) video addition, demonstrating the strong extensibility and transferability of our model to related and downstream tasks.

- *Inpainting.* If the reference image is omitted, the model automatically hallucinates plausible content inside the mask, restoring the damaged region (Figure 3(b)).
- *Addition.* Conversely, supplying an external mask sequence that covers empty space prompts the model to insert the reference object, effectively adding a moving subject to the clip (Figure 3(c)).

Video Try-on. When the target subject is constrained to garments, DreamSwapV acts as a video try-on system, as already illustrated in Figure 6. For more challenging scenarios like layered outfits and complex wrinkles, our model can serve as a strong prior and adapt after a brief quality tuning stage on a downstream try-on dataset, which we leave as future work.

Hand-Object Interaction. Limiting the subject to handheld items turns the task into hand–object interaction (HOI) editing (Figure 5). Unlike template-based reference-to-video(R2V) pipelines such as AnchorCrafter (Xu et al. 2024) or DreamActor-H1 (Wang et al. 2025), our approach directly swaps the product inside the original video—an avenue that remains unexplored. Same as extension to video try-on, with short, task-specific quality tuning, DreamSwapV could further bridge this gap and advance HOI video editing, a promising direction we also earmark for future exploration.

5.2 Social Impact

Positive Impacts. DreamSwapV broadens the reach of high-quality video subject swapping and related editing tasks by:

- **Lowering production barriers.** Content creators, educators and small studios can replace actors, props or garments without reshooting or green-screen setups, making professional video customization affordable and time-efficient.
- **Enabling new commercial and creative workflows.** The same pipeline supports video try-on, product insertion and damage inpainting, giving advertisers and e-commerce platforms rapid, personalized marketing assets while helping archivists restore legacy footage.
- **Research reusability.** Although DreamSwapV is initially fine-tuned from the Wan-I2V-14B backbone (Wan et al. 2025), it remains model-independent: the framework is fully plug-and-play, allowing researchers to attach it to any DiT-based I2V model and potentially achieve even stronger subject-swapping performance.

Risks and Challenges. The same ease of use can facilitate malicious deepfakes, brand counterfeits, or non-consensual impersonations; attackers need only a reference photo to graft someone into misleading footage. Copyright conflicts may arise if trademarked characters or logos are swapped into commercial videos, and highly convincing forgeries could undermine public trust in authentic media.

Ethical Concerns. Responsible use requires securing permission from anyone depicted in reference images and respecting intellectual-property rights for all source material. Clear disclosure of edited content, along with basic provenance tools (e.g., watermarking or edit logs), helps viewers distinguish synthetic footage from genuine video.

5.3 Limitations and Future Work

Although fruitful, we need to admit that DreamSwapV still has several limitations. First, a single reference image conveys only limited information about the target subject; when the object rotates or its backside is revealed, the swapping quality may degrade. Incorporating multi-view reference inputs—akin to AnchorCrafter—is a potential solution. Second, since rigid objects lack an explicit pose, fine-grained control remains challenging; introducing a spatial coordinate system (as in Orient Anything) to serve as the object's *pose* representation may yield more precise swapping. Third, the substantial training and inference costs of our full model may restrict its practicality for smaller-scale applications; we therefore plan to release a distilled version in the future for broader use. Finally, as discussed in Appendix 5.1, adapting DreamSwapV to downstream tasks such as video try-on and hand–object interaction represents a promising research direction. In summary, while our current model has limitations, these potential enhancements open new pathways for future development. We eagerly anticipate the next explorations of DreamSwapV—let's stay tuned!

References

Caron, M. et al. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Emerson, P. 2013. The original Borda count and partial voting. *Social Choice and Welfare*, 40(2): 353–358.

Hu, T. et al. 2025. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*.

Huang, Z. et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Jiang, Z. et al. 2025. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*.

Keling. 2025. Keling. https://klingai.com/cn/.

Khorasani, M. et al. 2022. Web application development with streamlit. *Software Development*, 498: 507.

Ku, M. et al. 2024. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*.

LAION-AI. 2022. aesthetic-predictor. https://github.com/LAION-AI/aesthetic-predictor.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Pika. 2025. Pika. https://pika.art/.

Radford, A. et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Song, W. et al. 2025. Insert anything: Image insertion via in-context editing in dit. *arXiv preprint arXiv:2504.15009*.

Figure 4: More qualitative comparisons between our DreamSwapV and four baselines in the *human* category (full body, half body and talking head). Please zoom in for details, and see the video demo for dynamic presentations.

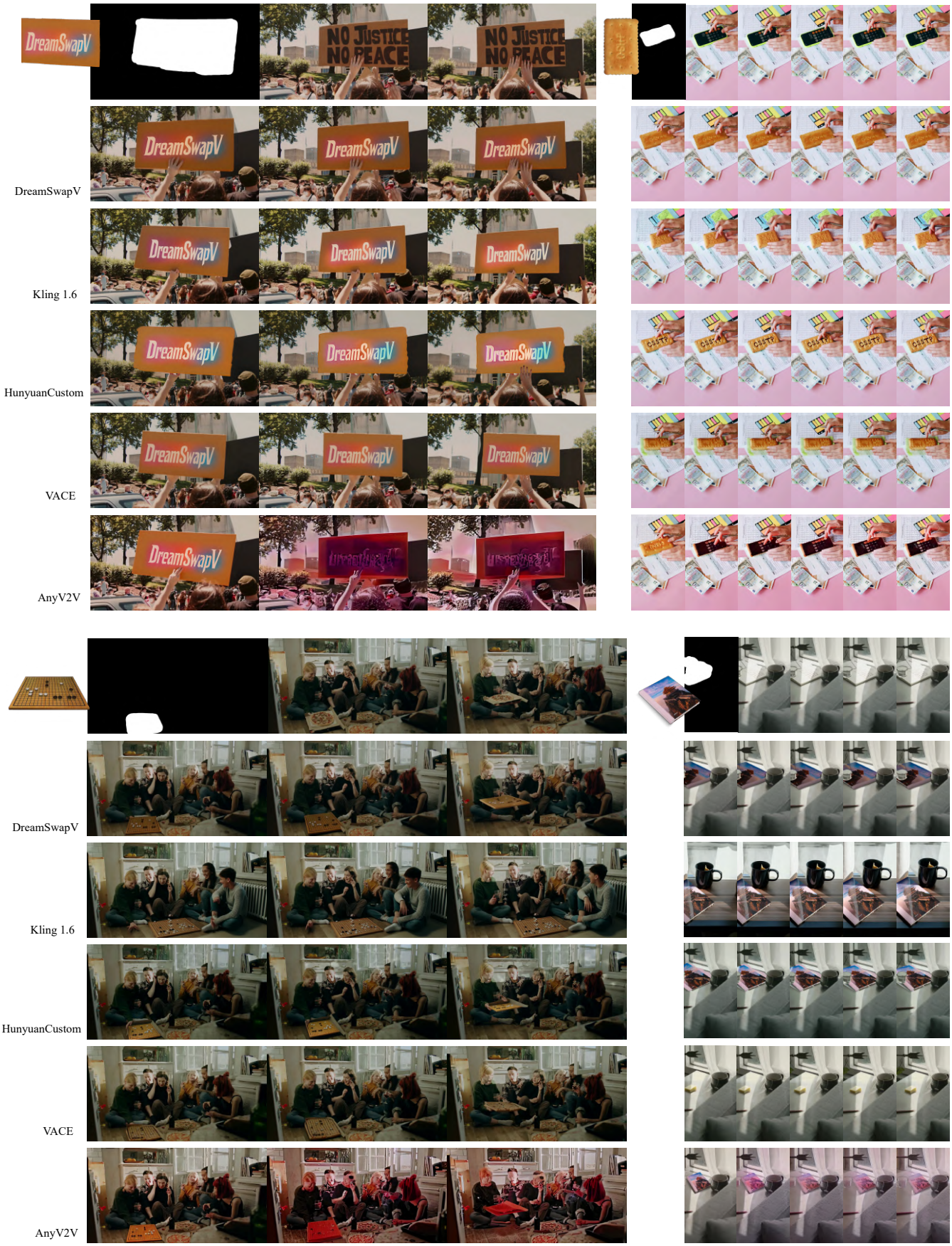


Figure 5: More qualitative comparisons between our DreamSwapV and four baselines in the *small object* category (handheld object and others). Please zoom in for details, and see the video demo for dynamic presentations.

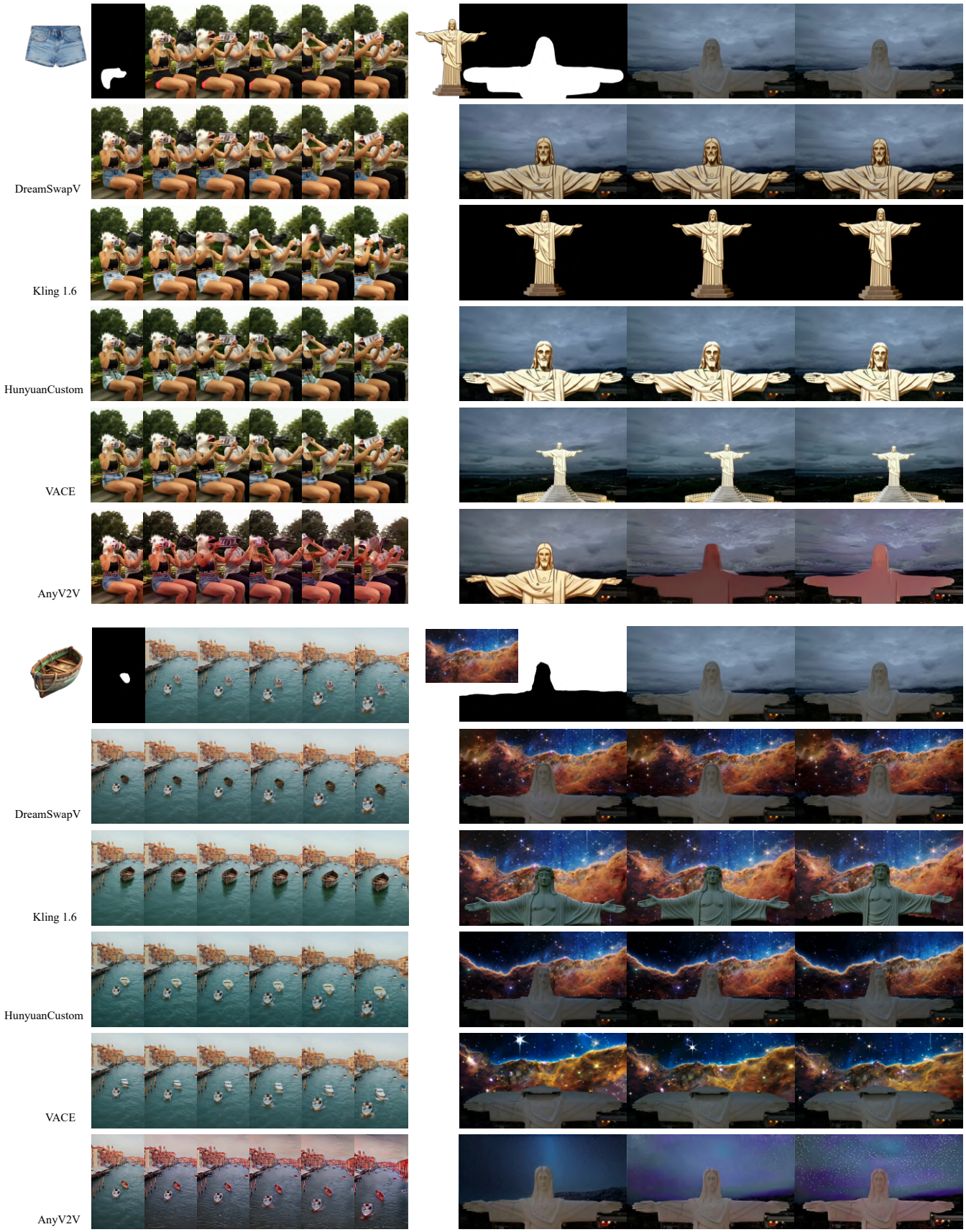


Figure 6: More qualitative comparisons between our DreamSwapV and four baselines in the *garment* and *large object* categories (vehicle, statue and sky). Please zoom in for details, and see the video demo for dynamic presentations.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow.

Wan, T. et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, L. et al. 2025. DreamActor-H1: High-Fidelity Human-Product Demonstration Video Generation via Motion-designed Diffusion Transformers. *arXiv preprint arXiv:2506.10568*.

Xu, Z. et al. 2024. Anchorcrafter: Animate cyberanchors saling your products via human-object interacting video generation. *arXiv preprint arXiv:2411.17383*.