FREECHROMA: GENERATIVE CHROMA KEYING FROM UNALIGNED BACKGROUND REFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

In film production, chroma keying composites an approved foreground performance into a new environment, but conventional workflows require coordinated background plates and dedicated harmonization. Recent generative methods often require pre-aligned backgrounds, focus on background-conditioned foreground relighting, or represent the environment only through an edited first frame. These restrictions limit background selection, scene composition, and camera motion while adding preparation or editing effort. We present FreeCHROMA, which adapts a pretrained video diffusion model in a parameter-efficient manner for generative video compositing from one to three unaligned background references. The model integrates the foreground into the referenced environment, preserving foreground identity, motion, and fine visual details while generating a spatiotemporally coherent background. Foreground–background harmonization occurs within the same generation process, without a separate relighting stage. Optional text instructions help disambiguate complex spatial relationships and camera behavior. To train the model, we construct pseudo-paired data from readily available real monocular videos through foreground–background separation and augmentation with existing foundation models. The original videos serve as training targets, while automated processing and quality verification enable scalable data construction. Across reference fidelity, foreground preservation, scene integration, harmonization, and overall video quality, FreeCHROMA consistently outperforms recent compositing and conditional video-generation baselines, with particularly clear gains under camera motion. FreeCHROMA relaxes the constraints of existing tools to enable end-to-end generative chroma keying from unaligned inputs, with exploratory experiments demonstrating its potential in robotics.

1 INTRODUCTION

Film and virtual-production workflows often capture an actor or object separately and composite the footage into an independently created environment. At this stage, the foreground performance—including identity, expression, motion, costume, and fine appearance—has usually been directed and approved. The goal is therefore not unconstrained video editing. It is to preserve that performance while constructing a plausible environment around it and changing only those appearance factors needed for coherent integration.

Conventional chroma keying separates foreground and background effectively, but transfers much of the difficulty to production: background plates must agree with the foreground in viewpoint, framing, camera motion, perspective, and illumination, after which matting refinement and harmonization are still required (Smith & Blinn, 1996; Sengupta et al., 2020). Recent generative systems reduce manual work but retain restrictive assumptions (Fig. 1, top left). Layer composition requires pre-aligned foreground and background videos and primarily relights an independent foreground layer against a fixed background, making it less suited to substantial camera motion or scene interaction. First/key-frame editing relies on an image editor to composite and harmonize selected frames before video generation. The environment is constrained only in those edited views, leaving newly revealed regions prone to drift; independently edited keyframes can also be inconsistent.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

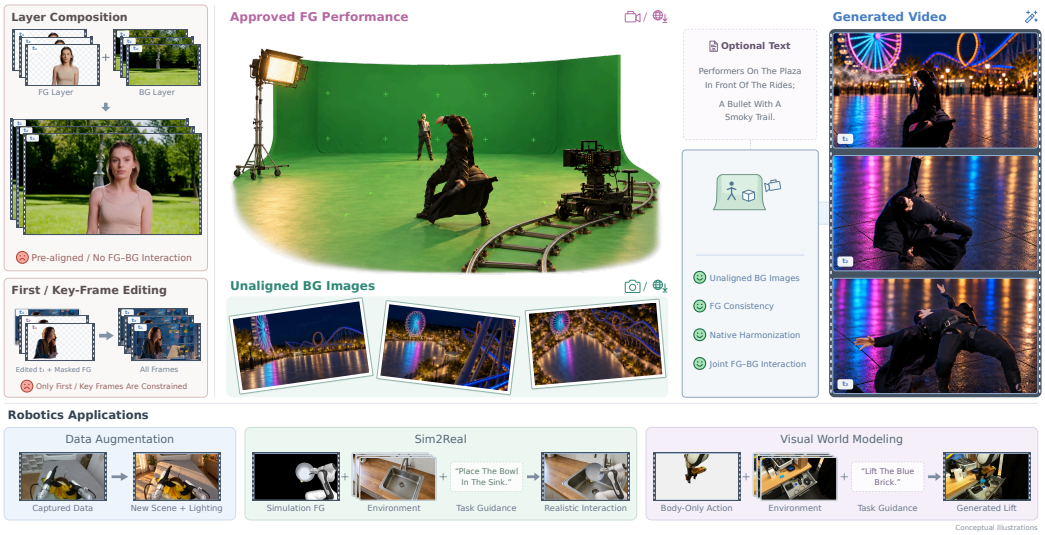

Figure 1: **FreeCHROMA and the production constraints it relaxes. Top left:** Layer composition requires pre-aligned videos and primarily relights the foreground against a fixed background. First/key-frame editing adds an image-editing step and constrains the environment only in selected views. **Top center and right:** FreeCHROMA preserves the foreground performance while using one or more unaligned background reference images throughout generation, jointly inferring composition, interactions, camera-aware background evolution, and harmonization, with optional text guidance. **Bottom:** Masked inputs retain all foreground elements for environment augmentation, the robot and interacting objects for sim-to-real transfer, or only the robot body for visual world modeling.

These restrictions are especially visible in mid-shot action sequences, such as fight choreography, parkour, or superhero-style stunts. Such shots commonly combine substantial subject motion with tracking, panning, or orbiting cameras, while retaining enough of the environment for foreground–background spatial relationships to matter. They therefore require more than relighting an approximately independent foreground layer: the environment must remain controllable as new regions enter the field of view, and the approved choreography must remain unchanged. Unified video editing models offer greater flexibility in instruction and reference conditioning, but often struggle to maintain fidelity to the supplied references.

We introduce **FreeCHROMA**, a *generative chroma keying* method that uses unaligned background references and harmonizes foreground and background during generation. It takes a masked foreground video and its mask, one to three images of the desired environment, and optional text. The reference images need not match any output frame in camera pose, crop, or layout. The model learns a joint spatiotemporal composition: where the foreground belongs in the referenced environment, how the background evolves with foreground motion and implied camera behavior, and how foreground appearance adapts for natural integration. Optional text primarily helps resolve ambiguity, guiding generation toward the intended interpretation when the visual inputs admit multiple plausible compositions.

The central challenge is the scarcity of paired supervision. We construct pseudo-pairs from ordinary videos by automatically extracting foregrounds and perturbing their illumination, while creating clean, unaligned background references through foreground removal and viewpoint transformation of selected keyframes. Reconstructing the original video from these inputs jointly teaches reference-conditioned background generation and native harmonization, without green-screen capture, 3D annotations, or a separate relighting network.

We implement this formulation through lightweight adaptation of a Wan-family video diffusion transformer (Wan et al., 2025), incorporating foreground conditioning, persistent reference context, and optional text. We train the added input-projection weights for foreground and mask conditioning, the background-reference convolution, and LoRA adapters (Hu et al., 2021). Our main

contribution is a scalable learning formulation that turns diverse natural videos into supervision for flexible composition.

Our specific contributions are:

- We formulate performance-preserving generative chroma keying with one to three spatially unaligned background references, learning composition rather than requiring it to be specified by a target plate.
- We introduce a pipeline for constructing pseudo-paired training data from readily available real monocular videos. It combines foreground–background separation with augmentation supported by existing foundation models, while retaining the original videos as training targets. Automated processing and quality verification enable scalable data construction.
- We integrate background synthesis and video-level harmonization within one parameter-efficient video diffusion model, without an explicit or post-hoc relighting stage. Across diverse foreground–background interactions and camera motions, our method outperforms state-of-the-art baselines in foreground preservation, background reference fidelity, and harmonization.

2 Background and Related Work

Matting and keying recover foreground layers (Smith & Blinn, 1996; Lin et al., 2021), while image harmonization and relighting adjust foreground appearance within an established composition (Cong et al., 2020; Zhang et al., 2025). Our setting additionally requires inferring scene composition and camera-dependent background evolution from unaligned background references.

Video relighting and background replacement. Recent open-source diffusion systems extend relighting and harmonization to video. RelightVid incorporates temporal modeling into background-conditioned foreground relighting (Fang et al., 2025), while Light-A-Video constructs a training-free video relighting pipeline from pretrained image relighting and video-generation models (Zhou et al., 2025). Lumen generates relit foreground videos conditioned on textual descriptions of the desired lighting and background (Zeng et al., 2025). IllumiCraft further introduces geometry-aware conditioning for controllable and temporally consistent video illumination (Lin et al., 2025). AnyPortal and FlowPortal combine video background replacement with temporally coherent foreground relighting in zero-shot or training-free formulations (Gao et al., 2025; 2026b). These methods demonstrate that pretrained generative models can coordinate foreground appearance and illumination over time. However, this formulation can be viewed as a simplification of the full compositing problem: it largely ignores explicit foreground–background interaction and instead treats the input background sequence—often already manually aligned or even left unaligned—as a proxy light source or global conditioning signal, rather than as a fully structured, spatially consistent environment to be jointly inferred. PAI-Studio (Gao et al., 2026a) associates background references with designated keyframes, implicitly assuming that the references are pre-aligned with the foreground at those frames. FreeCHROMA instead treats one or more spatially unaligned background images as persistent video-level references and jointly synthesizes the environment, its spatiotemporal relationship to the foreground performance, and the appearance changes required for integration.

Conditional video generation and keyframe-guided restylization. Latent diffusion models (Rombach et al., 2022) and diffusion transformers (Peebles & Xie, 2023) provide scalable backbones for conditional video synthesis. VACE (Jiang et al., 2025) and Bernini (Liu et al., 2026) unify multiple video creation and editing tasks through spatially organized visual and textual conditions. Despite their versatility, these models often struggle to achieve precise foreground preservation and faithful adherence to background references. Closely related in production motivation, ID-V2V performs identity-preserving video restylization from a source video and a stylized first frame, optionally supplemented with additional keyframes, propagating the specified scene, lighting, and appearance across the sequence while preserving the captured performance (Xu et al., 2026). This keyframe-based cascade offloads foreground–background composition to an upstream image editor, making the final video dependent on the spatial and appearance consistency of the edited keyframes and effectively using the downstream video model primarily as a controllable temporal interpolator rather than as the compositor itself. More importantly, the target environment is

represented only within the field-of-view covered by the edited keyframes; regions newly revealed by foreground or camera motion remain underconstrained and may drift away from the intended scene. FreeCHROMA addresses a distinct control regime: it composes a preserved foreground performance directly with one to three standalone background references that need not align with the source video or any output frame. These references remain available throughout generation, enabling the model to jointly infer composition, background evolution, and foreground–background harmonization without a separately edited first frame or an additional relighting stage.

3 METHOD

3.1 PROBLEM FORMULATION

Let $F \in \mathbb{R}^{T\times H\times W\times 3}$ be a foreground-only video with alpha mask M, let $\mathcal{R} = \{R_k\}_{k=1}^{K}$ contain $K \in \{1, 2, 3\}$ images of a desired environment, and let c be optional text. We seek

$$\widehat{V} = G_\theta(F, M, \mathcal{R}, c), \tag{1}$$

where $\widehat{V}$ preserves the semantic and temporal content of F while generating and harmonizing a background consistent with $\mathcal{R}$. The references are not target frames: no pixel correspondence, camera calibration, or pose alignment is assumed. The output is consequently not unique. Our objective is conditional plausibility and reference fidelity under foreground preservation, rather than reconstruction of a particular unseen background trajectory.

3.2 SCALABLE PSEUDO-PAIRS FROM NATURAL VIDEO

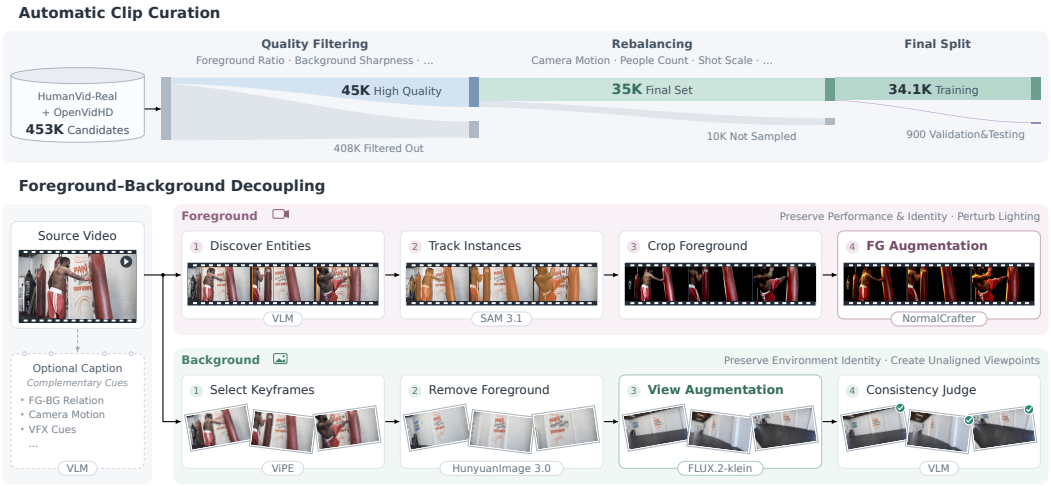

Figure 2: **Automatic pseudo-pair construction from natural videos. Clip curation:** A shot detector, conventional image-quality tests, and a vision–language model (VLM) retain clips with a visible foreground and informative background. **Foreground augmentation:** The VLM enumerates the foreground entities; SAM 3.1 tracks their masks M; and a temporally coherent photometric and normal-guided illumination transform produces $(\widetilde{F}, M)$. **Background reference construction:** A motion proxy selects one to three keyframes, a frozen image editor removes all selected entities, a frozen novel-view model changes viewpoint and crop, and a VLM quality gate retains scene-consistent references $\widetilde{\mathcal{R}}$. The conditions map to the untouched source video V, yielding scalable supervision without green-screen capture, manual compositing, or 3D annotation. Tool roles and selected models are labeled in the figure; further implementation details are provided in Appendix D. Visuals are schematic.

We derive supervision from an ordinary video V by constructing

$$\big(\widetilde{F}, M, \widetilde{\mathcal{R}}, c\big) \longrightarrow V. \tag{2}$$

The original video supplies a temporally coherent target in which foreground, background, camera motion, and illumination already agree. Fig. 2 summarizes the process; data-construction details are provided in Appendix D.

Video selection and foreground extraction. We retain short clips in which both the salient foreground and surrounding environment are sufficiently visible and sharp. A vision-language model enumerates foreground entities, including interacting objects that should remain attached to the performance. SAM 3.1, a released refinement of SAM 3 (Carion et al., 2026), segments and tracks those concepts over time. We take the union of the selected instance masks, and form $F = M \odot V$. Clips with unstable masks, excessive foreground occupancy, shot changes, or negligible background evidence are rejected.

Foreground illumination augmentation. The inexpensive construction $F = M \odot V$ introduces a degenerate shortcut: on the matte support, the conditioning foreground is identical to the regression target, so copying F is optimal regardless of the referenced environment. This shortcut fails at inference, where the foreground and target environment are generally captured under different illumination. We mitigate it by perturbing only the conditioning path while leaving the target video unchanged:

$$\widetilde{F} = T_\theta(F; N), \qquad \theta \sim p(\theta), \tag{3}$$

where N is an optional frame-aligned normal map. T_θ combines global photometric changes, low-frequency spatial variation, and normal-guided directional shading; all parameters are shared across the clip and the original matte support is restored after transformation. Because $(\widetilde{F}, \widetilde{\mathcal{R}}, c)$ must still predict the unmodified V, an exact foreground copy is no longer optimal. The model must instead preserve performance content while using the referenced environment to infer a compatible appearance. This converts natural-video supervision into an explicit harmonization signal without requiring paired relit foregrounds. The transform is deliberately bounded and designed to resemble illumination variation rather than arbitrary corruption; full operators and sampling ranges are given in Appendix D.1.

Unaligned background references. We use camera trajectories estimated by ViPE (Huang et al., 2025) to assess motion magnitude and select one to three source keyframes B_k accordingly. The VLM foreground labels and masks guide a frozen image-editing model, $\mathrm{RemoveFG}$, which removes the selected entities from B_k. A frozen viewpoint-generation model then changes the cleaned background's viewpoint and crop:

$$\widetilde{R}_k = A_{\mathrm{view}}(\mathrm{RemoveFG}(B_k)) \,. \tag{4}$$

This transformation is essential because an unmodified keyframe can be copied or aligned directly to its source frame. View augmentation forces the references to encode environment identity, appearance, and partial geometry rather than a predetermined composition. We reject transformations that change the underlying scene or introduce conspicuous artifacts using VLM-based scene-consistency and quality checks. The reference-generation models and quality checks are described in Appendix D (Background references).

3.3 Joint composition and harmonization

We initialize from Wan2.2-Fun-A14B-Control (Wan et al., 2025) and fine-tune its high- and low-noise experts separately; implementation details are provided in Appendix E. Fig. 3 summarizes the conditioning architecture and marks the frozen and trainable components. A frozen VAE encodes the target V and foreground condition $\widetilde{F}$ as z_0 and z_F, respectively. At diffusion time τ, noise gives z_τ. We concatenate the noisy-target latent, foreground latent, and explicit mask condition channel-wise and map them jointly into transformer tokens through an expanded patchify input projection:

$$h_0 = P_{\mathrm{patch}}([z_\tau \,\|\, z_F \,\|\, m]) \,, \tag{5}$$

where m denotes M on the latent grid and $\|$ denotes channel concatenation. The pretrained target-latent weights remain frozen; only the added foreground and mask weights in P_{patch} are trained. This dense interface preserves foreground–target alignment at every space–time position without a separate foreground encoder.

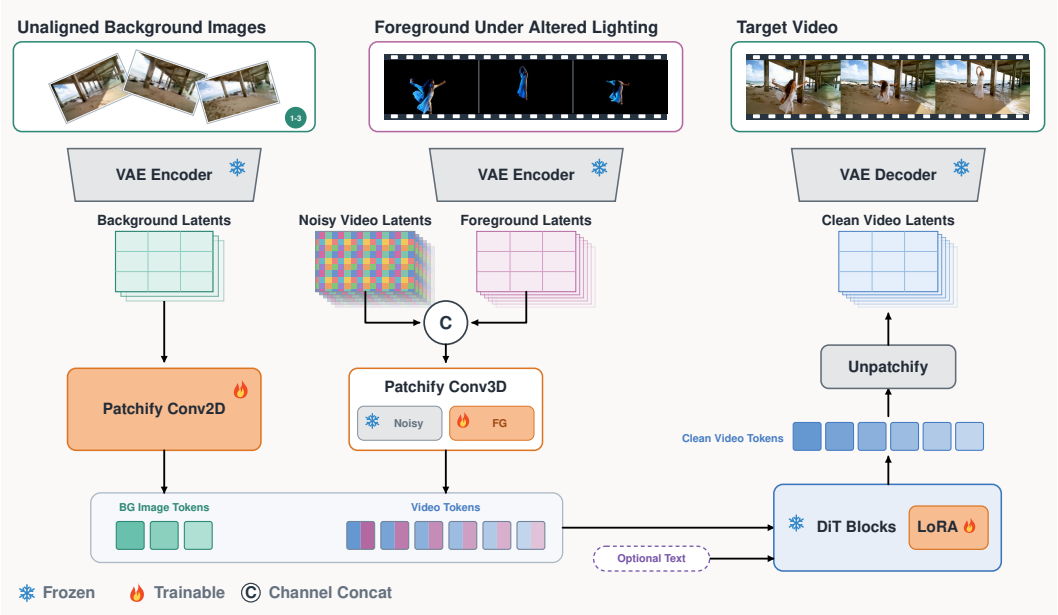

Figure 3: **Parameter-efficient FreeCHROMA architecture.** During training, a frozen VAE encodes the target and foreground streams. The noisy target latent, foreground latent, and aligned mask condition are concatenated channel-wise and enter through the expanded patchify projector in eq. (5); background reference tokens are produced by a trainable convolution and prepended to the video sequence; optional text uses the pretrained cross-attention path. Flow-matching supervision updates the added foreground and mask input weights, the background-reference convolution, and DiT LoRA adapters. At inference, the same conditioned denoiser transforms noise into a complete video, which the frozen VAE decodes. Scene synthesis, spatiotemporal composition, and foreground harmonization occur in one denoising process, without an auxiliary relighting or frame-wise harmonization network.

Each reference is encoded by the frozen VAE and mapped to tokens E_k by a trainable background-reference convolution. We concatenate $E_{1:K}$ and prepend them to the noisy video-token sequence as persistent visual context. Unlike a first-frame image condition, these tokens remain accessible to all denoising blocks and output frames. Text tokens E_c enter through the pretrained cross-attention pathway when an instruction is present; classifier-free dropout over text and references supports absent text and robust variable-cardinality conditioning.

We follow the backbone's flow-matching parameterization. For sampled τ and noise ϵ, the trainable parameters minimize

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{V,\tau,\epsilon}\left[\left\|v_\theta(z_\tau, h_0, E_{1:K}, E_c, \tau) - v^\star\right\|_2^2\right], \tag{6}$$

with the standard target velocity $v^\star$ defined by the selected Wan scheduler. The VAE, text encoder, and original DiT weights remain frozen. We optimize the added foreground and mask input-projection weights, the background-reference convolution, and DiT LoRA adapters (Hu et al., 2021). No auxiliary foreground relighting network, per-frame harmonizer, explicit depth loss, or reconstruction model is used at inference.

4 Experimental Results

4.1 Evaluation Protocol

We evaluate FreeCHROMA on 500 foreground videos (167 HumanVid-Real (Wang et al., 2024) and 333 OpenVidHD (Nan et al., 2024)) spanning static to large camera motion. Each foreground video is paired with two different sets of unaligned background references, yielding 1,000 test cases.

All methods are evaluated on the same foreground–background pairs, with inputs adapted to their supported conditioning modalities.

Baselines comprise Lumen (Zeng et al., 2025), FlowPortal (Gao et al., 2026b), IllumiCraft (Lin et al., 2025), two ID-V2V variants (Paste and FLUX.2) (Xu et al., 2026), VACE (Jiang et al., 2025), and Bernini (Liu et al., 2026). Visually conditioned methods receive the background image; Lumen receives only a deterministic Qwen caption of it. We assess video quality with VBench (Huang et al., 2024) and VBench-2.0 (Zheng et al., 2025), compositing with a vision-language evaluator, and camera-trajectory consistency with BA-Track (Chen et al., 2025).

4.2 General Video Quality

Tab. 1 shows that FreeCHROMA achieves strong overall video quality, outperforming most baselines across the evaluated metrics. These results suggest that our pipeline effectively preserves the pretrained backbone's general video generation quality while adapting it to generative compositing.

Table 1: Video quality on the evaluation benchmark: VBench (Huang et al., 2024), except multi-view consistency from VBench-2.0 (Zheng et al., 2025). Best results are bold; second-best are underlined. †Lumen receives a caption; other methods use visual background references.

	Content Consistency		Temporal Quality			Visual Quality	
Method	Subject Cons. ↑	Background Cons. ↑	Motion Smoothness ↑	Temporal Flickering ↑	Multi-view Cons. ↑	Aesthetic Quality ↑	Imaging Quality ↑
Lumen (Zeng et al., 2025)†	<u>94.26</u>	93.22	98.06	96.14	19.06	53.82	71.25
VACE (Jiang et al., 2025)	93.94	93.28	97.73	95.98	24.56	57.33	72.42
IllumiCraft (Lin et al., 2025)	91.64	92.12	96.87	95.78	15.07	53.23	70.61
FlowPortal (Gao et al., 2026b)	93.23	92.53	97.77	95.95	15.63	53.85	67.24
ID-V2V + Paste (Xu et al., 2026)	91.70	93.18	98.07	96.11	25.01	56.49	73.30
ID-V2V + FLUX.2 (Xu et al., 2026)	93.14	92.64	97.92	95.87	<u>27.36</u>	57.64	**73.68**
Bernini Liu et al. (2026)	94.08	**94.57**	<u>98.31</u>	<u>96.33</u>	**28.78**	**59.40**	73.30
FreeCHROMA	**94.59**	<u>93.65</u>	**98.55**	**97.16**	26.51	<u>57.82</u>	<u>73.63</u>

4.3 Foreground Preservation and Scene Integration

Following the evaluation criteria of PAI-Studio (Gao et al., 2026a), we assess foreground preservation, edge quality, foreground–background fusion, and illumination harmony using a vision-language model (Tab. 2); the evaluation protocol is detailed in Appendix C. FreeCHROMA achieves the highest score on all four criteria. These results suggest that joint generation preserves foreground fidelity while adapting appearance and spatial relationships to the target environment.

Table 2: Task-specific evaluation on the novel-background benchmark. VLM scores measure foreground preservation, boundary quality, scene integration, and illumination harmony. Scores are averaged over 1,000 generated videos (500 foreground videos, each paired with two different sets of background references). Best results are bold; second-best are underlined.

	Foreground Quality		Scene Integration	
Method	FG Preservation VLM ↑	Edge Quality VLM ↑	FG–BG Fusion VLM ↑	Illum. Harmony VLM ↑
Lumen (Zeng et al., 2025)†	9.778	6.722	6.470	7.040
VACE (Jiang et al., 2025)	9.852	<u>7.840</u>	<u>7.580</u>	<u>7.864</u>
IllumiCraft (Lin et al., 2025)	8.330	6.196	6.278	6.932
FlowPortal (Gao et al., 2026b)	5.624	5.842	5.802	6.450
ID-V2V + Paste (Xu et al., 2026)	9.636	6.304	6.372	6.088
ID-V2V + FLUX.2 (Xu et al., 2026)	<u>9.901</u>	7.386	7.058	7.602
Bernini (Liu et al., 2026)	9.180	6.586	6.476	7.130
FreeCHROMA	**9.904**	**8.172**	**7.728**	**7.956**

Fig. 4 further shows that FreeCHROMA achieves foreground–background coherence in spatial geometry and lighting, preserves foreground appearance, motion, and fine-grained details such as text and logos on clothing, and maintains high background fidelity and consistency across viewpoints.

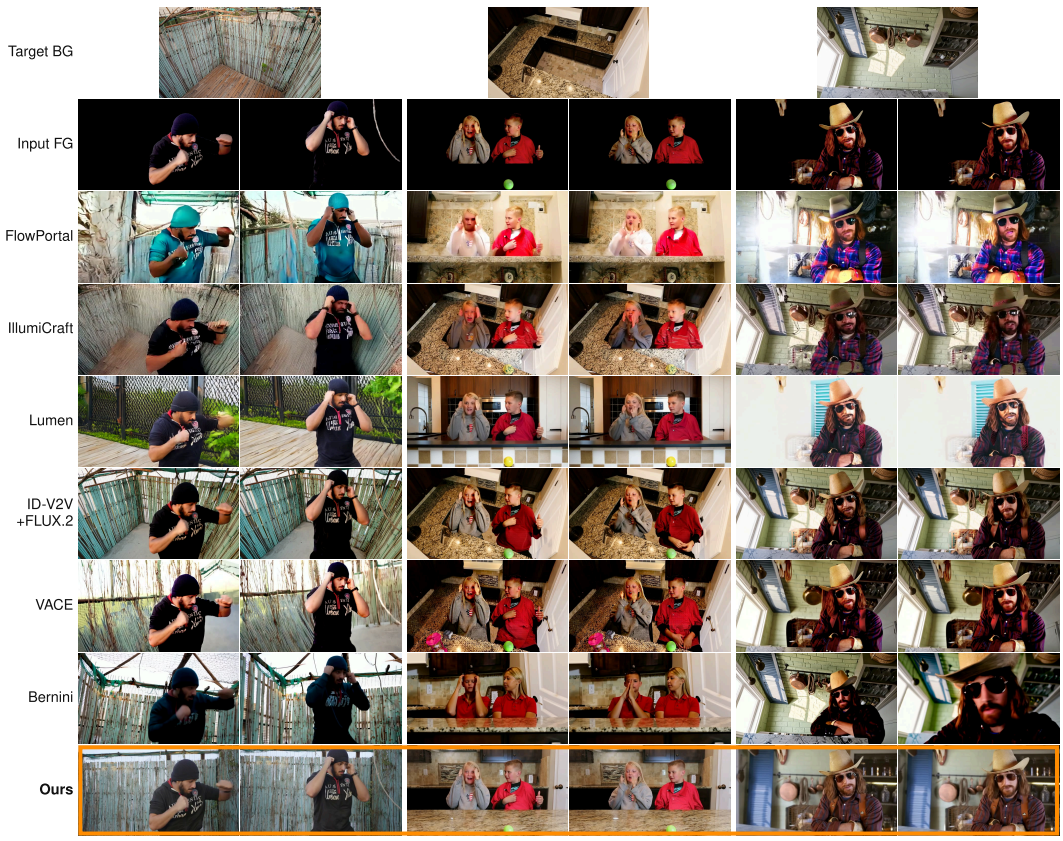

Figure 4: **Qualitative comparison of generated videos.** Given background reference images and a masked foreground sequence, each method generates an output video; two frames sampled at matched relative times are shown. FreeCHROMA faithfully preserves the foreground performance while producing more plausible foreground–background spatial relationships, including viewpoint consistency, relative scale, and physical support. It also generates more natural camera motion and harmonizes illumination and appearance across foreground and background. More video cases are presented in the HTML supplementary material.

4.4 Camera-Trajectory Consistency

A convincing composite requires the synthesized environment to remain geometrically coherent with the foreground and follow its camera motion. We therefore compare each generated video with the original source video from which its foreground input was extracted. Camera trajectories are estimated using BA-Track (Chen et al., 2025) and aligned with a per-video similarity transform.

Table 3: Camera-trajectory consistency between generated videos and the original source videos from which the foreground inputs are extracted. Best results are bold; second-best are underlined.

	Global Trajectory Error		Local Motion Error		Motion Extent
Method	ATE ↓	ATE/Path (%) ↓	RPE Translation ↓	RPE Rotation (°) ↓	Relative Path Length → 1
Lumen (Zeng et al., 2025)†	0.015	9.13	0.008	0.593	0.834
VACE (Jiang et al., 2025)	<u>0.006</u>	<u>4.50</u>	<u>0.005</u>	<u>0.176</u>	<u>0.936</u>
IllumiCraft (Lin et al., 2025)	0.017	9.25	0.009	0.653	0.823
FlowPortal (Gao et al., 2026b)	0.015	9.00	0.008	0.614	0.866
ID-V2V + Paste (Xu et al., 2026)	0.011	6.11	0.006	0.192	0.929
ID-V2V + FLUX.2 (Xu et al., 2026)	0.009	5.74	0.006	0.201	0.902
Bernini (Liu et al., 2026)	0.008	5.46	<u>0.005</u>	0.253	0.893
FreeCHROMA	**0.004**	**3.96**	**0.003**	**0.119**	**0.940**

Absolute Trajectory Error (ATE) measures global trajectory disagreement after per-video Sim(3) alignment, while ATE/Path normalizes ATE by the source trajectory length. The translational and rotational components of Relative Pose Error (RPE) measure local frame-to-frame motion disagreement. Relative Path Length is the ratio between the aligned generated and source trajectory lengths, for which 1.0 is ideal. Tab. 3 shows that FreeCHROMA achieves the lowest global and local trajectory errors and a relative path length closest to 1.0, indicating more faithful preservation of the source camera motion.

4.5 Ablation Study

We remove each training-time augmentation separately and retrain for the same number of training steps with all other settings unchanged (Tab. 4). Removing foreground illumination perturbation lowers foreground–background fusion from 7.728 to 6.384 and illumination harmony from 7.956 to 6.958, suggesting that training with mismatched foreground lighting helps the model adapt foreground appearance to the target environment. Qualitatively, this variant tends to preserve the source foreground lighting and color casts, resulting in a visible mismatch with the target environment (Fig. 5). Removing background viewpoint augmentation increases ATE from 0.004 to 0.005 and shifts relative path length from 0.940 to 0.909, indicating weaker agreement with the source camera motion. Qualitative results show that this variant tends to inherit the unaligned reference viewpoint, producing inconsistent scene geometry. The full model performs best across all metrics in this ablation, supporting the combined use of the two augmentations for harmonization and camera-trajectory consistency. Qualitative comparisons are given in Appendix A (Fig. 5), and video visualizations in the HTML supplementary material.

Table 4: Ablation results on training augmentation components.

	VLM Evaluation		VBench Evaluation		Camera Trajectory	
Method	FG–BG Fusion ↑	Illum. Harmony ↑	Subject Cons. ↑	Motion Smoothness ↑	ATE ↓	Relative Path Length → 1
w/o FG illumination perturbation	6.384	6.958	94.16	98.50	0.006	0.918
w/o BG viewpoint augmentation	7.638	7.902	94.56	98.45	0.005	0.909
FreeCHROMA	**7.728**	**7.956**	**94.59**	**98.55**	**0.004**	**0.940**

4.6 Downstream Robotics Applications

FreeCHROMA preserves selected scene elements and their motion while coherently generating the rest. Beyond background augmentation of captured manipulation videos, varying the mask supports sim-to-real transfer by retaining the robot body and interacting objects, and visual world modeling by retaining only the robot body. We explore both settings on selected episodes from RoboCasa (Nasiriany et al., 2026) and DROID (Khazatsky et al., 2024) without robot-specific training. The examples show coherent motion and plausible interactions in the reference environments, including successful and failed outcomes. These results suggest that a compositing model can supply realistic manipulation videos and potentially serve as a rough visual world model. Examples and discussion are provided in Appendix B (Figs. 6 and 7), with video visualizations in the HTML supplementary material.

5 Conclusion and Limitations

We presented **FreeCHROMA**, which preserves an existing foreground performance while jointly learning composition, background evolution, and foreground–background harmonization from one to three unaligned reference images. A scalable pseudo-pair pipeline built from natural videos and parameter-efficient adaptation of a pretrained video diffusion transformer enable flexible composition without pre-aligned plates, edited keyframes, or a separate relighting stage. Two limitations remain. The references provide context for short clips but no persistent scene memory for long-form generation. Independently generated shots also lack precise cross-shot foreground–background consistency beyond coarse textual guidance. Persistent environment representations and explicit cross-shot spatial constraints are promising directions for future work.

AI USE STATEMENT

AI tools assisted with manuscript translation and polishing, figure presentation, data construction, and VLM-based evaluation (Appendices D and C).

ETHICS STATEMENT

Our study uses existing video datasets. Generative compositing may misrepresent the context of real performances or enable non-consensual manipulation. Its use should respect privacy, consent, and dataset licenses, and generated videos should be clearly identified as synthetic.

REPRODUCIBILITY STATEMENT

We will open-source our code and release the model weights to support reproducibility. Section 3 describes the model and training objective. Appendices D, E, and C document data construction, implementation settings, and VLM evaluation, respectively. Camera evaluation metrics are defined in Tab. 3.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

Weirong Chen, Ganlin Zhang, Felix Wimbauer, Rui Wang, Nikita Araslanov, Andrea Vedaldi, and Daniel Cremers. Back on track: Bundle adjustment for dynamic scene reconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4951–4960. IEEE, 2025.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8391–8400. IEEE, 2020.

Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. Relightvid: Temporal-consistent diffusion model for video relighting. *arXiv preprint arXiv:2501.16330*, 2025.

Heyuan Gao, Bangxun Tang, Yiren Song, Guian Fang, Zijian He, Jie Yang, and Mike Zheng Shou. PAI-Studio: Cinematic video background replacement with camera-aware motion, 2026a. URL https://arxiv.org/abs/2606.01399.

Wenshuo Gao, Xicheng Lan, and Shuai Yang. Anyportal: Zero-shot consistent video background replacement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18990–18999. IEEE, 2025.

Wenshuo Gao, Junyi Fan, Jiangyue Zeng, and Shuai Yang. Flowportal: Residual-corrected flow for training-free video relighting and background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2025–2034, June 2026b.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative

models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, June 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8762–8771, 2021.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. Illumicraft: Unified geometry and illumination diffusion for controllable video generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/283066055b0256ca8e3e0c8c96019357-Abstract-Conference.html.

Chenchen Liu, Junyi Chen, Lei Li, Lu Chi, Mingzhen Sun, Zhuoying Li, Yi Fu, Ruoyu Guo, Yiheng Wu, Ge Bai, and Zehuan Yuan. Bernini: Latent semantic planning for video diffusion. *arXiv preprint arXiv:2605.22344*, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations (ICLR)*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2288–2297. IEEE, 2020.

Alvy Ray Smith and James F Blinn. Blue screen matting. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 259–268, 1996.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. Humanvid: Demystifying training data for camera-controllable human image animation. In *NeurIPS*, 2024.

Yuancheng Xu, Mingming He, Pablo Salamanca, Li Ma, Yash Kant, Emmett Steven, Paul Debevec, and Ning Yu. Id-v2v: Identity-preserving video restylization. *arXiv preprint arXiv:2607.22830*, 2026.

Jianshu Zeng, Yuxuan Liu, Yutong Feng, Chenxuan Miao, Zixiang Gao, Jiwang Qu, Jianzhang Zhang, Bin Wang, and Kun Yuan. Lumen: Consistent video relighting and harmonious background replacement with video generative models. *arXiv preprint arXiv:2508.12945*, 2025. URL https://arxiv.org/abs/2508.12945.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=u1cQYxRI1H.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. doi: 10.48550/arXiv.2503.21755.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. Light-a-video: Training-free video relighting via progressive light fusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13315–13325. IEEE, 2025.

A Qualitative Ablation

We complement the quantitative ablation in Tab. 4 with qualitative results for the two training-time augmentations that make our model robust to real inputs. Both ablated models are trained with the same data and schedule as the full model, for the same number of training steps; only the ablated component is removed.

w/o FG illumination perturbation. The foreground illumination perturbation (photometric changes and normal-guided shading) is disabled, so the foreground control seen during training always carries the illumination of the target video. The model therefore learns that the output illumination can be read off the input and, at test time, tends to copy the illumination of the input foreground instead of re-lighting it for the target environment (Fig. 5, top row): stage lighting, color

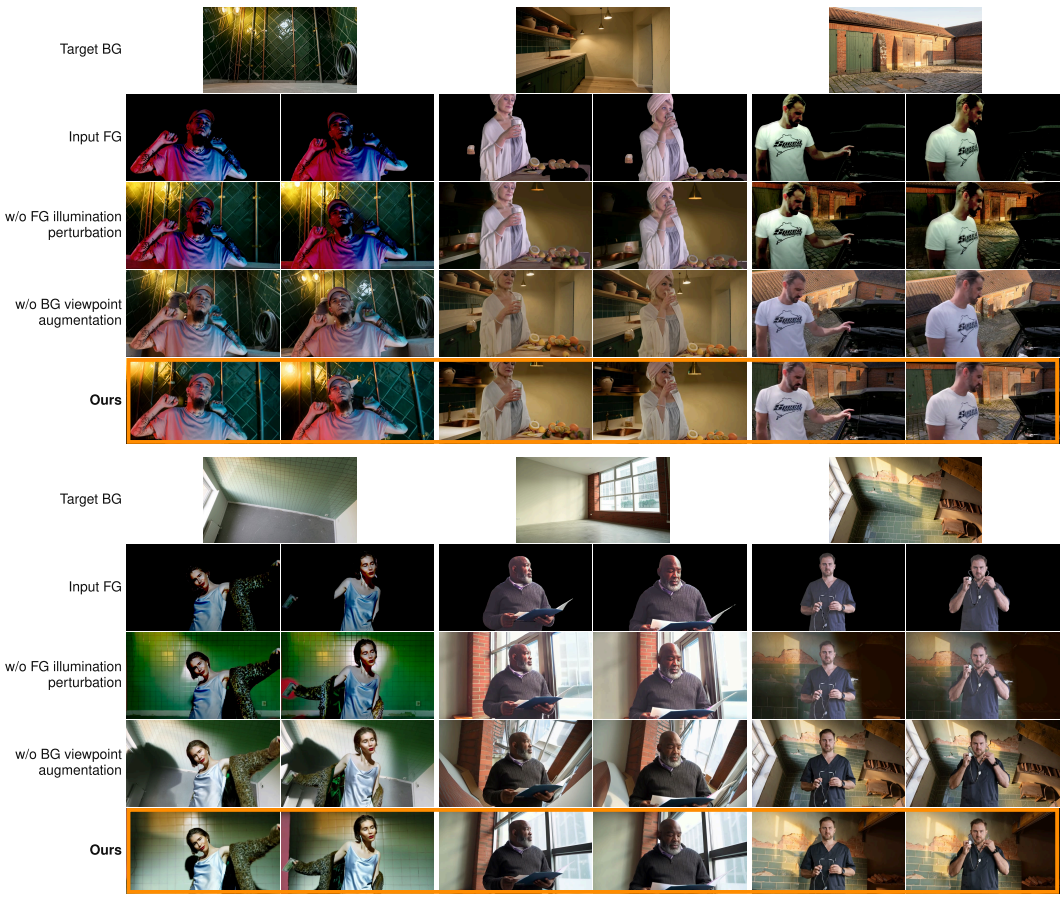

Figure 5: **Qualitative ablation of the two training-time augmentations.** For each sample we show the environment reference (*Target BG*), the masked foreground control (*Input FG*), and two frames of the videos generated by the two ablated models and by the full model (**Ours**, orange). Without the foreground illumination perturbation, the model tends to copy the illumination of the input foreground instead of re-lighting it for the target environment. Without the background viewpoint augmentation, the model is trained only on pixel-aligned background plates; given an unaligned reference at test time, it tends to force the reference's viewpoint into the output and composites a scene that is inconsistent with the foreground. The full model re-lights the foreground to match the target environment and uses the reference as an unaligned appearance cue, producing a consistent scene. More video cases are presented in the HTML supplementary material.

casts, and exposure of the source footage are carried over unchanged, and the composite does not look like it was captured in the target scene.

w/o BG viewpoint augmentation. The background references are taken from the original camera view (pixel-aligned with the target video) instead of the viewpoint-augmented, unaligned plates described in Appendix D (Background references). The model then learns to treat the reference as an aligned background layer. Given an unaligned reference at test time, it tends to force the reference's viewpoint into the output and composites a scene that is inconsistent with the foreground (Fig. 5, bottom row); the full model instead uses the reference as an unaligned appearance cue and synthesizes a background whose layout is consistent with the foreground camera.

Discussion. The two augmentations address complementary failure modes. Without the illumination perturbation the background is plausible but the foreground is not harmonized; without the viewpoint augmentation the foreground is harmonized but the scene geometry breaks.

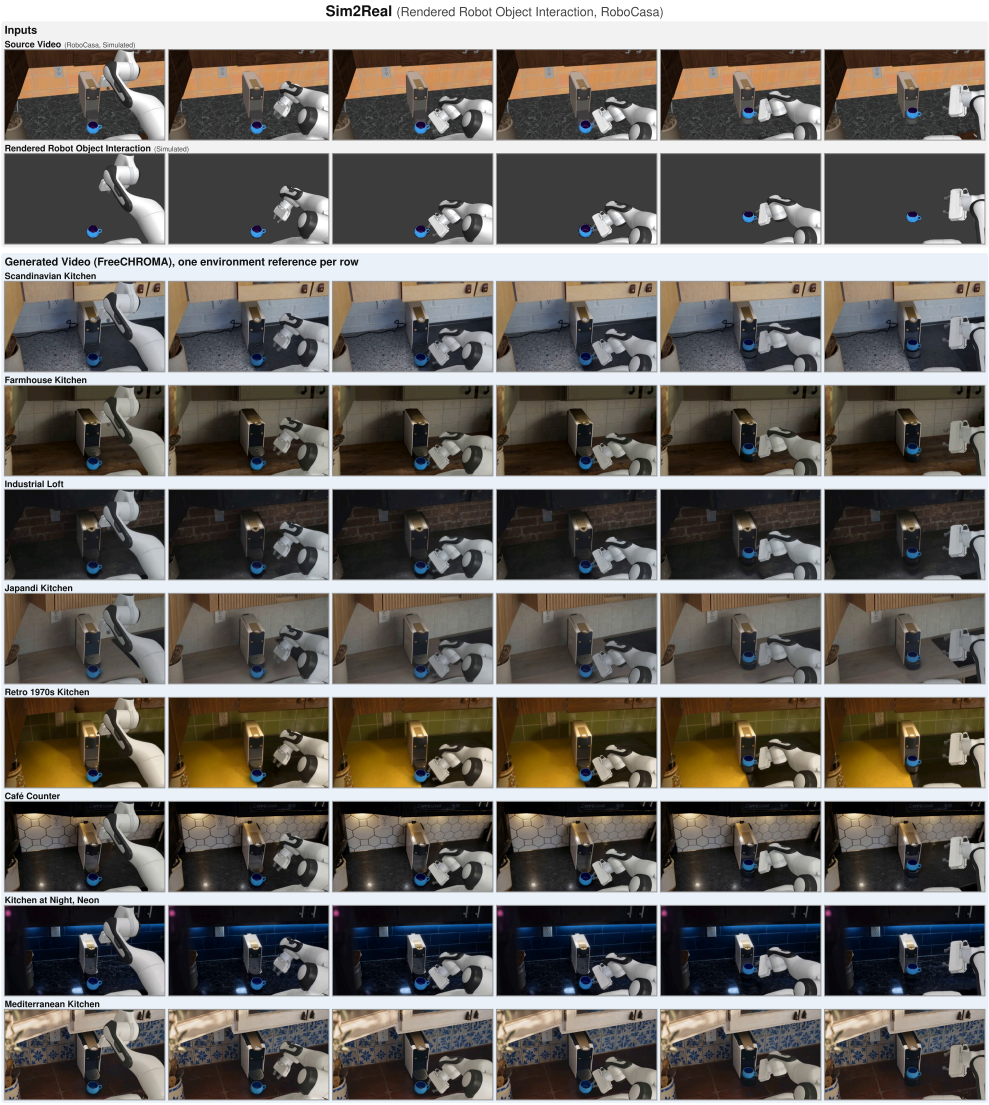

Figure 6: **Sim2Real on RoboCasa (Nasiriany et al., 2026): key frames.** Top: the simulated source video and the rendered robot-object interaction used as foreground control (the arm together with the manipulated objects). Below: the videos generated by FreeCHROMA from eight different environment reference images, shown at the same frames. The same simulated action is re-rendered into every environment while the arm motion and the object interaction are preserved. More video cases are presented in the HTML supplementary material.

B Downstream Robotics Applications

We explore robotics applications using the same generation pipeline without robot-specific training. A masked foreground sequence specifies the scene elements and motions to preserve, while background references and task prompts guide the generated environment and interactions.

Sim-to-real transfer. We use rendered sequences from RoboCasa (Nasiriany et al., 2026), retaining the robot body and interacting objects, together with their motion, as foreground conditions. A single background reference specifies the target environment, and a task prompt provides interaction guidance. FreeCHROMA synthesizes the surrounding scene and harmonizes the retained foreground with the reference environment (Fig. 6). More video cases are presented in the HTML supplementary material.

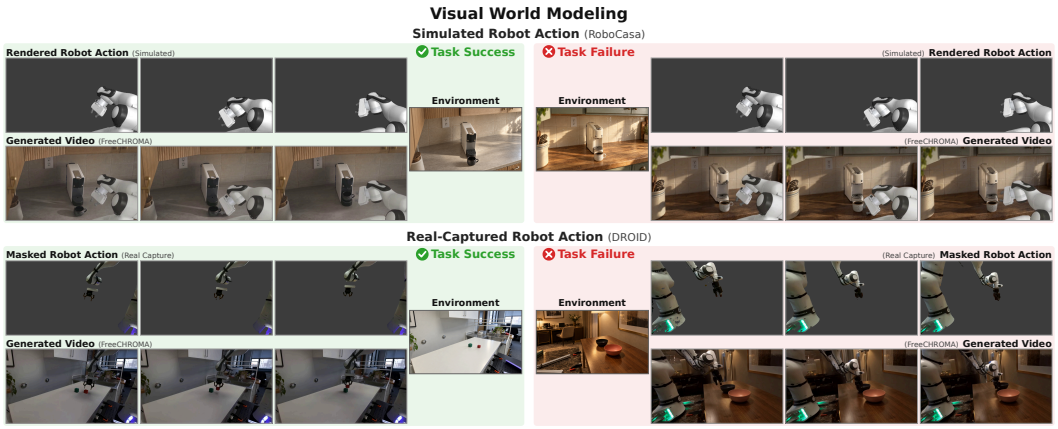

Figure 7: **Qualitative robotics applications.** Rendered RoboCasa (Nasiriany et al., 2026) (top) and masked DROID (Khazatsky et al., 2024) (bottom) arms; green/red panels show successful/failed interactions. More video cases are presented in the HTML supplementary material.

Visual world modeling. We render the robot body from RoboCasa (Nasiriany et al., 2026) sequences and segment it from DROID (Khazatsky et al., 2024) sequences, retaining only the robot action and excluding interacting objects. The input sequence specifies the robot action, while the model generates the surrounding scene and the objects' responses to that action, conditioned on the reference environment and task prompt (Fig. 7). More video cases are presented in the HTML supplementary material.

C VLM Evaluation Protocol

We use Qwen3-VL-8B-Instruct to rate how well each generated video preserves the subject and places it in the new scene. The rating criteria are adapted from PAI-Studio (Gao et al., 2026a). For each video, the model sees 12 frames spaced evenly from beginning to end. It gives separate scores for three aspects of the result: whether the subject's brightness, light direction, color, and shadows fit the background throughout the video; whether the subject has a believable size, position, and relationship to its surroundings rather than appearing to float; and whether its outlines remain clean, including difficult areas such as hair, fingers, and motion-blurred edges. We assess subject preservation separately. For that rating, the model sees 12 evenly spaced frames from the original foreground-only video and 12 from the generated result, and checks whether the subject's identity, shape, clothing, pose, and movement have been retained. The instructions explicitly allow lighting and color to change to suit the new scene. The model is not shown the target background reference. Each score is an integer from 1 to 10 and must include a brief explanation. We repeat the ratings three times with repeatable settings (temperature 0.2 and top-p 0.9), take the median of the three scores for each generated video, and report the average over 1,000 generated videos (500 foreground videos, each paired with two distinct background conditions).

D Details of Data Construction Pipeline

We construct pseudo-pairs from natural videos through the automated pipeline in section 3.2. Foreground appearance and background references are augmented, while the original video V remains the target.

Video selection and foreground extraction. We retain short clips with a clear foreground, sufficient visible background, and no shot transitions. We use Claude Opus 4.6 as the vision-language model (VLM) for foreground-entity identification and visual quality checks. SAM 3.1 (Carion et al., 2026) segments and tracks the subject and interacting objects; their masks are merged into a joint matte M to extract $F = M \odot V$. Clips with unstable masks, severe blur, or excessive foreground occupancy are rejected.

Normal estimation. We use NormalCrafter to estimate video surface normals, aligned with the foreground frames for the illumination augmentation in Appendix D.1.

Background references. We use camera trajectories estimated by ViPE to assess motion magnitude and select one to three source keyframes accordingly. HunyuanImage-3.0-Instruct removes the selected foreground entities from these keyframes. FLUX.2-klein-9B then changes the viewpoint and crop to produce unaligned background references. The VLM rejects candidates that alter scene identity or introduce conspicuous artifacts.

D.1 Foreground illumination augmentation

We perturb the foreground condition to prevent direct copying of its original illumination. The target video remains unchanged, so the model must recover a compatible appearance while preserving the performance.

Photometric changes. We sample $s \sim \mathcal{U}(0.3, 1.4)$, replacing it with a mild draw from $\mathcal{U}(0, 0.1)$ for 5% of clips. White balance, gamma, and saturation use $s_c = 0.35s$; the remaining operators use s (Tab. 5). Operators are applied in approximately linear light, $\ell = x^{2.2}$, followed by clipping and conversion back to display encoding. All random parameters are shared across the clip to avoid artificial flicker.

Table 5: Photometric augmentation, applied in the listed order. Each occurrence of u is sampled independently from $\mathcal{U}(-1, 1)$.

Operator	Sampling rule
Exposure	$\log g \sim \mathcal{N}(0, (0.40s)^2)$
White balance	Channel log gains $\sim \mathcal{N}(0, (0.25s_c)^2)$; gains normalized by their mean
Contrast	$q = 1 + 0.40su$, about linear mid-gray 0.18
Gamma	$\log \kappa_j \sim \mathcal{N}(0, (0.20s_c)^2)$ for each channel j
Saturation	$\sigma = 1 + 0.50s_c u$, about BT.601 luma

With probability 0.6, we also apply a fixed spatial field combining a directional ramp and radial vignette. The ramp direction is uniform on $[0, 2\pi)$, its amplitude lies in $[0, 0.5s]$, and the vignette strength lies in $[0, 0.4s]$. The combined gain is clipped to $[0.2, 2.0]$.

Normal-guided shading. When frame-aligned normals are available, they are resized with the foreground and renormalized. We sample $L \sim \mathcal{N}(0, I_3)$, replace L_z with $|L_z| + 0.3$, and normalize it. The shading and relative gain are

$$S = a + k \max(0, \widehat{N}^\top L), \tag{7}$$

$$\rho = \mathrm{clip}\big(1 + 0.6s\,(S/\overline{S} - 1)\,,\, 0.3,\, 2.5\big)\,, \tag{8}$$

where $a \sim \mathcal{U}(0.4, 0.8)$, $k \sim \mathcal{U}(0.3, 0.9)$, and $\overline{S}$ is the foreground mean. This normalization separates directional shading from global exposure. The light parameters remain fixed throughout the clip; unavailable normals fall back to photometric augmentation.

Foreground support. For inputs in $[0, 1]$, the augmentation uses $\widehat{M}_f(p) = \mathbf{1}[\max_j F_f(j, p) > 8/255]$ to recover foreground support. The final condition is

$$T_\theta(F; N) = \widehat{M} \odot \Psi_{\mathrm{normal}}(\Psi_{\mathrm{phot}}(F); N)\,, \tag{9}$$

where the two operators denote the photometric and optional shading branches. Multiplication by $\widehat{M}$ restores the black background. These transforms provide approximate illumination mismatches; the natural video supplies the harmonized target.

E Datasets and Implementation Details

We collect an initial pool of 453K video clips from HumanVid-Real (Wang et al., 2024) and OpenVidHD (Nan et al., 2024). Following the filtering pipeline in Fig. 2, we apply coarse filtering based

on foreground occupancy, background clarity, temporal continuity, and the presence of subtitles, logos, or black borders, retaining 45K high-quality clips. We then rebalance the pool by motion level and the numbers of foreground people and objects to obtain a final set of 35K clips, of which 34.1K are used for training and the remaining 900 clips are held out for validation and testing.

We initialize from Wan2.2-Fun-A14B-Control (Wan et al., 2025) and fine-tune its high- and low-noise experts separately, using a nominal noise boundary of 0.9 and scheduler shift 5.0. Each run uses eight NVIDIA H200 GPUs with BF16 precision and gradient checkpointing. Training clips contain 81 frames at 720×1280 resolution. We use one clip per GPU without gradient accumulation, giving a global batch size of eight. The final model uses the checkpoint saved at step 4,500 from each expert.

We apply LoRA (Hu et al., 2021) with rank 64 and $\alpha = 32$ to the attention query, key, and value projections and the linear layers of feed-forward networks. In the input projection, we freeze the weights corresponding to the noisy video latent and train only those corresponding to the foreground and mask conditions. We also train the background-reference convolution. The VAE, text encoder, and remaining backbone weights are frozen. AdamW uses learning rates of 10^{-4} for LoRA and 10^{-5} for the input layers and background-reference convolution, with weight decay 0.03 and 0, respectively. Both groups use cosine decay with 50 warmup steps and gradient-norm clipping at 0.05.

Each example provides a masked foreground video, its binary mask sequence, and up to three background references. Text dropout is 10%. Background conditioning uses 10% whole-set dropout and 10% per-reference dropout, retaining at least one valid view when the set is kept.

At inference, we use 25 sampling steps with a classifier-free guidance scale of 4.5.