

# Virtually Being: Customizing Camera-Controllable Video Diffusion Models with Multi-View Performance Captures

YUANCHENG XU, Eyeline Labs, United States of America  
 WENQI XIAN, Eyeline Labs, United States of America  
 LI MA, Eyeline Labs, United States of America  
 JULIEN PHILIP, Eyeline Labs, United Kingdom  
 AHMET LEVENT TAŞEL, Eyeline Labs, Canada  
 YIWEI ZHAO, Netflix, United States of America  
 RYAN BURGERT, Eyeline Labs, United States of America  
 MINGMING HE, Eyeline Labs, United States of America  
 OLIVER HERMANN, Eyeline Labs, Germany  
 OLIVER PILARSKI, Eyeline Labs, Germany  
 RAHUL GARG, Netflix, United States of America  
 PAUL DEBEVEC, Eyeline Labs, United States of America  
 NING YU, Eyeline Labs, United States of America

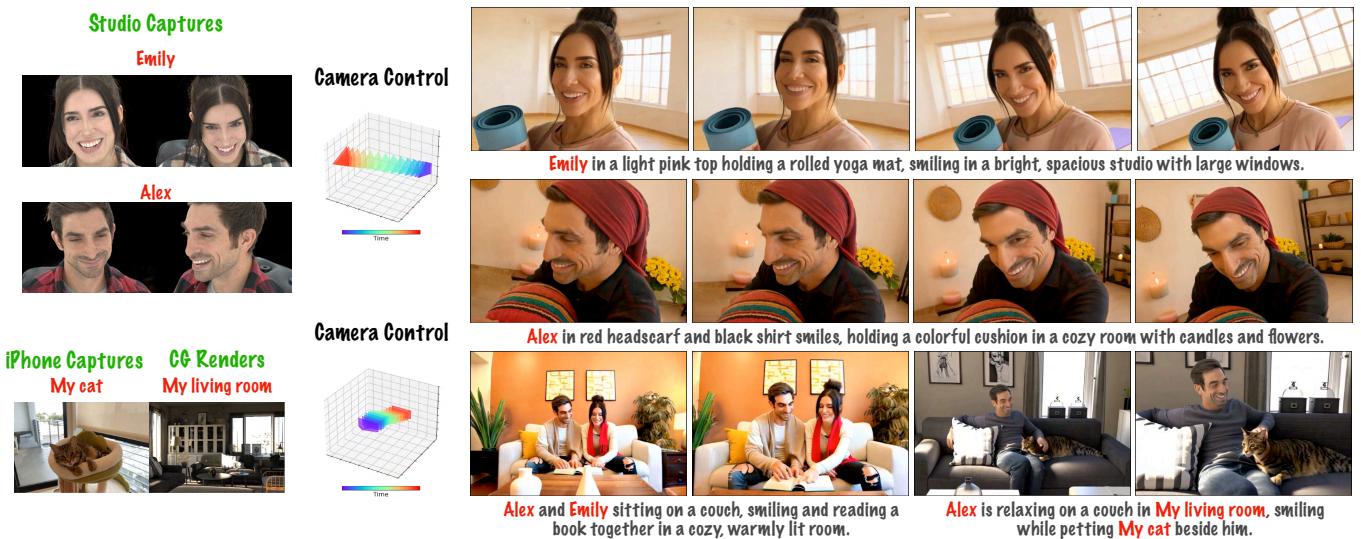


Fig. 1. Leveraging multi-view data from 4D reconstructions from studio captures, CG renders, and real-life videos, our method enables video generation with strong multi-view identity preservation and camera control, while also supporting multi-subject composition and subject–scene interactions.

Authors' Contact Information: Yuancheng Xu, Eyeline Labs, United States of America, [xuyuancheng0@gmail.com](mailto:xuyuancheng0@gmail.com); Wenqi Xian, Eyeline Labs, United States of America, [wenqixian3@gmail.com](mailto:wenqixian3@gmail.com); Li Ma, Eyeline Labs, United States of America, [lmaag@connect.ust.hk](mailto:lmaag@connect.ust.hk); Julien Philip, Eyeline Labs, United Kingdom, [julien.philip@scanlinevfx.com](mailto:julien.philip@scanlinevfx.com); Ahmet Levent Taşel, Eyeline Labs, Canada, [leventtasel@gmail.com](mailto:leventtasel@gmail.com); Yiwei Zhao, Netflix, United States of America, [yiweiz@netflix.com](mailto:yiweiz@netflix.com); Ryan Burgert, Eyeline Labs, United States of America, [ryancentralorg@gmail.com](mailto:ryancentralorg@gmail.com); Mingming He, Eyeline Labs, United States of America, [hmm.lillian@gmail.com](mailto:hmm.lillian@gmail.com); Oliver Hermann, Eyeline Labs, Germany, [oliver.hermann@scanlinevfx.com](mailto:oliver.hermann@scanlinevfx.com); Oliver Pilarski, Eyeline Labs, Germany, [oliver.pilarski@scanlinevfx.com](mailto:oliver.pilarski@scanlinevfx.com); Rahul Garg, Netflix, United States of America, [rahulgarg@netflix.com](mailto:rahulgarg@netflix.com); Paul Debevec, Eyeline Labs, United States of America, [debevec@gmail.com](mailto:debevec@gmail.com); Ning Yu, Eyeline Labs, United States of America, [ningyu.hust@gmail.com](mailto:ningyu.hust@gmail.com).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation

We introduce a framework that enables both multi-view character consistency and 3D camera control in video diffusion models through a novel customization data pipeline. We train the character consistency component with recorded volumetric capture performances re-rendered with diverse camera trajectories via 4D Gaussian Splatting (4DGS), lighting variability

on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

SA Conference Papers '25, Hong Kong, Hong Kong

© 2025 Copyright held by the owner/authors. Publication rights licensed to ACM.

ACM ISBN 979-8-4007-2137-3/2025/12

<https://doi.org/10.1145/3757377.3763888>

obtained with a video relighting model. We fine-tune state-of-the-art open-source video diffusion models on this data to provide strong multi-view identity preservation, precise camera control, and lighting adaptability. Our framework also supports core capabilities for virtual production, including multi-subject generation using two approaches: joint training and noise blending, the latter enabling efficient composition of independently customized models at inference time; it also achieves scene and real-life video customization as well as control over motion and spatial layout during customization. Extensive experiments show improved video quality, higher personalization accuracy, and enhanced camera control and lighting adaptability, advancing the integration of video generation into virtual production. Our project page is available at: <https://eyeline-labs.github.io/Virtually-Being/>.

**CCS Concepts:** • Computing methodologies → Computer vision; Biometrics; Motion capture; Image representations.

**Additional Key Words and Phrases:** Video generation model, camera control, customization, identity preservation, 4d reconstruction

#### ACM Reference Format:

Yuancheng Xu, Wenqi Xian, Li Ma, Julien Philip, Ahmet Levent Taşel, Yawei Zhao, Ryan Burgett, Mingming He, Oliver Hermann, Oliver Pilarski, Rahul Garg, Paul Debevec, and Ning Yu. 2025. Virtually Being: Customizing Camera-Controllable Video Diffusion Models with Multi-View Performance Captures. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3757377.3763888>

## 1 Introduction

Virtual production is reshaping filmmaking by combining live-action footage with computer-generated elements through technologies such as LED volumes, camera tracking, and performance capture. This integration allows filmmakers to interact with virtual environments during production, offering direct control over scene layout, lighting, character performance, and camera motion. Increasingly, video generation models [Blattmann et al. 2023a,b; Guo et al. 2024; Team 2025; Weijie Kong and Jie Jiang 2024; Yang et al. 2024a] are being incorporated into this pipeline to further expand creative possibilities and streamline content creation. A key application is subject-specific customization, enabling the generation of novel scenes featuring particular characters. Additionally, in cinematic storytelling, camera motion is essential for conveying perspective and emotion, but it also exposes the subject from multiple angles, posing a significant challenge for preserving identity. As a result, multi-view identity preservation becomes crucial for generating coherent and realistic customized videos with camera control.

However, despite growing interest in customized video generation, explicit multi-view identity preservation under camera motion remains largely unexplored. Recent methods often customize subjects using a single reference image [Chen et al. 2025; Jiang et al. 2024b; Yuan et al. 2025a], which lacks view diversity and leads to identity inconsistencies when the subject is observed from novel angles. Moreover, existing approaches offer limited camera control during customization. For example, MotionBooth [Wu et al. 2024b] supports only 2D camera translations, falling short for 3D camera movements common in filmmaking.

In this work, we introduce an approach that enables both multi-view subject customization and 3D camera control by repurposing 4D Gaussian Splatting (4DGS) from volumetric captures [Yang et al. 2024b] — originally designed for 4D reconstruction — as a data

generator for video diffusion models. This data pipeline bridges 4DGS’s accurate multi-view rendering with the generative flexibility of video generation models. We begin by capturing dynamic human performances using two professional volumetric capture rigs: a 75-camera facial setup and a 160-camera full-body system. From these captures, we apply 4DGS to reconstruct dynamic human motion and render videos with diverse and precisely annotated camera trajectories. To reflect the central role of lighting in cinematography, we further augment the data using a generalizable relighting model [Mei et al. 2025], producing HDR-based lighting variations. This pipeline provides rich multi-view character supervision, accurate camera conditioning, and lighting diversity essential for high-fidelity generation. Building on this customization dataset, we adopt a two-stage training strategy: first pretraining on general camera-annotated datasets for camera-conditioned synthesis, then fine-tuning on our customized data to achieve multi-view identity preservation under 3D camera motion.

Our framework supports a range of features essential for virtual production, enabling flexible control over subjects, scenes, and actions in generated videos. (1) Multi-subject Generation: We enable multi-subject video generation by jointly customizing the model on subject-specific datasets. In addition to joint training, we incorporate a noise blending approach [Kong et al. 2024] that combines independently customized models at inference time by blending LoRA features using segmentation masks from GroundingDINO [Liu et al. 2024b] and SAMv2 [Ravi et al. 2024], enabling efficient subject composition. (2) Scene Customization: Our method supports scene-specific generation using high-quality CG videos curated by professional artists, allowing novel subject–scene interactions under diverse camera motions. (3) Real-life Customization Data: Beyond 4DGS data, we validate our pipeline on real-life videos with estimated camera parameters [Wang et al. 2025], demonstrating its effectiveness in real-world settings. (4) Action and Spatial Layout Control: By fine-tuning the Go-with-the-Flow model [Burgett et al. 2025] on our customized dataset, we enable control over subject motion and layout, preserving spatial arrangement and movement patterns from source videos.

In summary, our key contributions are as follows:

- (1) We present the first framework to explicitly preserve multi-view identity under precise 3D camera control, enabled by a novel customization data pipeline that integrates professional volumetric captures, 4DGS reconstruction, and relightable rendering — combined with a two-stage training strategy that first learns general camera-conditioned video generation and then customizes to specific subjects.
- (2) Our method supports a broad range of generative capabilities for filmmaking, including multi-subject generation via both joint training and a noise blending scheme for combining independently customized models, scene customization, real-life video-based customization, and control over subject motion and spatial layout.
- (3) We conduct extensive benchmarking, ablations, and user studies, demonstrating clear improvements in multi-view identity preservation, camera control accuracy and lighting control—highlighting the method’s utility for virtual production applications.

## 2 Related work

### 2.1 Scene and character customization

Early video diffusion models, while capable of generating content from text prompts, often struggled with temporal consistency and multi-view coherence. More recent approaches address these issues by incorporating structured scene representations. Methods like ReconX [Liu et al. 2024a], ViewCrafter [Yu et al. 2024], and Wonderland [Liang et al. 2024] combine diffusion models with 3D representations like point clouds or 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023]. However, these approaches often struggle with fast scene motion, and typically focus only on scene generation with less explicit control over characters in the scene.

Character customization in video diffusion models can be approached through finetuning-free or finetuning-based methods. Fine-tuning-free methods, such as Videobooth [Jiang et al. 2024b] and ConsisID [Huang et al. 2024a], generate subject-consistent videos from a single reference image, while ConceptMaster [Huang et al. 2025], VideoAlchemist [Chen et al. [n. d.]], and Phantom [Liu et al. 2025] extend these ideas to multi-subject video generation. However, in scenarios involving subject motion and camera motion, where viewpoints can vary significantly, a single reference image often fails to capture the full multi-view characteristics of the subject. Fine-tuning-based methods, including DreamVideo [Wei et al. 2024], VideoStudio [Long et al. 2024], Magic-Me [Ma et al. 2024b], and Yuan et al. [2025a], build on DreamBooth [Ruiz et al. 2023] to adapt pre-trained models to specific characters using small sets of images. Chen et al. [2025] further extends fine-tuning approaches to multiple identities. These methods are often constrained by limited input data, reducing viewpoint diversity, and struggle to integrate customized characters into dynamic, consistently lit scenes with controlled camera motion. We address this with a fine-tuning-based strategy for both character and scene customization with controllable camera motions.

### 2.2 Camera and lighting control

Controlling camera movement is crucial for cinematic storytelling. Early approaches used fine-tuning techniques, like AnimateDiff's [Guo et al. 2024] use of LoRAs [Hu et al. 2022], to manage specific motion types. However, this offers limited precision. More recent methods, such as MotionCtrl [Wang et al. 2024b] and its successors [He et al. 2024b, 2025; Kuang et al. 2024; Xu et al. 2024], condition video generation directly on camera extrinsics, often using Plücker embeddings [Sitzmann et al. 2021]. Training-free strategies [Hou et al. 2024; Ling et al. 2024; Xiao et al. 2024] exist but can be difficult to integrate with other control aspects. 4D scene generation methods [Sun et al. 2024; Watson et al. 2024; Wu et al. 2024a] offer inherent camera control, but their synthesis quality currently lags behind dedicated video generation models. Our approach builds upon AC3D [Bahmani et al. 2025a] and VD3D [Bahmani et al. 2025b], which effectively combine ControlNet [Zhang et al. 2023] and Plücker embeddings for high-quality, but extends this by adapting the camera conditioned model for other controls. Other works focus on redirecting camera trajectories in monocular videos [Bai et al. 2025; Yu et al. 2025], while we aim to generate novel customized videos with controllable camera motion. Most relevant to our work,

MotionBooth [Wu et al. 2024b] supports customized video generation with camera control but is limited to 2D translations, while our method handles full 3D camera motion.

Lighting control is another vital aspect of cinematography. While some methods allow for lighting adjustments via text prompts [Zhang et al. 2025], this lacks the precision required for professional cinematography. DiffRelight [He et al. 2024a] focuses on portrait relighting, which is limited in scope. *Lux Post Facto* [Mei et al. 2025] provides state-of-the-art video-based relighting via HDR map encoding, which we use out of the box to augment our customization dataset with diverse lighting conditions.

### 2.3 Controllable video generation with volumetric captures

While the above methods control video generation by deploying the data prior in the video diffusion model, which is trained on millions of in-the-wild videos, another line of work focuses on acquiring high dimensional data with volumetric light stages and uses complex data processing pipeline to achieve fine-grained control for the final video. With synchronized multi camera settings, one can reconstruct 3D or 4D representations such as mesh [Beeler et al. 2011; Cagniart et al. 2010; Fyffe et al. 2011], NeRF [Isik et al. 2023; Lombardi et al. 2019] or Gaussian Splatting [He et al. 2024a; Jiang et al. 2024a; Luiten et al. 2024], which achieves camera controls. Further more, by designing 3D deformable model [Ma et al. 2024a; Qian et al. 2024] or skinned model [Peng et al. 2023, 2021], one can achieve expression and pose control by providing detailed expression code and pose parameters. With data that captures the performance under different lighting conditions such as one-light-at-a-time (OLAT), photometric relighting can be achieved [Mei et al. 2025]. In this work, we leverage volumetric captures for customizing video generation models.

## 3 Method



Fig. 2. Overview of our training pipeline (right) for camera-controlled customized video generation, consisting of a camera pretraining stage and a customization stage. The data pipeline (left) generates customization data by capturing multi-view performances, applying 4D Gaussian Splatting, and rendering videos with diverse viewpoints, camera motions, and lighting.

**Two-stage Training.** We aim to generate subject-specific videos that preserve multi-view identity across novel contexts with controllable camera motion. As viewpoint variation becomes more pronounced with subject and camera movement, maintaining multi-view consistency is critical. To address this, we adopt a two-stage training pipeline for both text-to-video and image-to-video models. In the pretraining stage, we train the model on diverse general datasets to learn camera-conditioned video generation across varying scenes and motions. In the customization stage, we fine-tune the

model on subject-specific multi-view data to capture identity details and view-dependent variations. This design enables the model, at inference time, to generate identity-consistent, camera-controllable videos of customized subjects in diverse and unseen contexts.

*Camera Pretraining Stage.* To enable camera-controlled video generation of general content, we adopt a ControlNet architecture inspired by AC3D [Bahmani et al. 2025a], where camera information is represented using Plücker coordinates [He et al. 2024b; Kuang et al. 2024]. The camera representations are processed through a fully convolutional encoder, concatenated with video tokens from the main DiT, and passed through the ControlNet’s DiT blocks. We integrate the camera information into the main DiT via token-wise summation before each block. Following AC3D, camera conditioning is applied only during the first 40% of denoising timesteps and injected into the first 25% of DiT blocks to improve controllability and visual quality. During training, we freeze the main DiT and train only the ControlNet on RealEstate10K [Zhou et al. 2018] for static scenes and HumanVid [Wang et al. 2024a] for dynamic, human-centric videos. The resulting model generates diverse general-content videos while accurately following specified camera trajectories.

*Customization Stage.* In this stage, we customize the camera-conditioned model to generate the target subject while preserving multi-view identity consistency, which is particularly crucial under camera motion. We adapt the DreamBooth framework [Ruiz et al. 2023], fine-tuning the model on a subject-specific customization dataset while simultaneously using a regularization dataset—sampled from the pretraining data—to maintain general video generation and camera control capabilities. During customization, each subject is associated with a unique token embedded in the text prompts, allowing the model to learn to generate that specific subject when conditioned on the token. In the next section we detail how to construct effective customization datasets.

*Image-to-video Generation.* In addition to customizing the video generation model, image-to-video (I2V) generation requires a subject-specific initial frame, which we generate using text-to-image (T2I) methods. For single-subject generation, we fine-tune FLUX.1-dev using DreamBooth [Ruiz et al. 2023] on subject-specific images from our customization datasets. For multi-subject generation, we use MuDI [Jang et al. 2024], which extends FLUX.1-dev to support identity-consistent multi-subject synthesis. These customized images are then converted into videos using a camera-conditioned I2V model, which is also fine-tuned on the same customization dataset to preserve identity and support controllable motion.

### 3.1 Constructing customization datasets

A key challenge in the customization stage is constructing an effective customization dataset: it must provide high-quality multi-view captures of the subject and precise camera annotations to support accurate camera control during generation. In the following, we discuss three sources of data: 4DGS from professional volumetric captures for human subject customization, CG scenes for scene customization and real-life videos.

*Professional Volumetric Captures.* To obtain high-quality 4D reconstruction of a human subject, we first capture multi-view data in a controlled studio environment, using a volumetric face rig equipped with 75 synchronized cameras arranged on a cylindrical structure measuring 2.5 meters in height and 2.7 meters in diameter. Full-body performances are recorded using a larger body rig with 160 synchronized cameras mounted on a 4-meter-wide cylinder. Subjects are illuminated with multiple strobe lights to ensure flat and diffuse lighting. Each subject performs 3–6 multi-view sequences, each lasting approximately 50 to 180 frames at 24 frames per second. The captured subjects, referred to as “Alex” and “Emily,” serve as reference identities throughout our study.

*4D Reconstruction as a Data Source.* To obtain sufficient customization data with multi-view identity information and camera information, we propose a novel approach: re-purposing 4D Gaussian Splatting (4DGS) based on [Duan et al. 2024; Yang et al. 2024b], originally designed for 4D reconstruction, as a data generator for video generation tasks. While 4DGS excels at producing high-fidelity multi-view renderings along diverse camera trajectories, it lacks the capability to synthesize novel content beyond the captured scenes. Conversely, video generation models can create new content and contexts, but often struggle with maintaining precise subject identity and view-dependent consistency. By leveraging 4DGS-generated data to customize the video generation model, we combine the strengths of both paradigms: precise subject modeling from 4DGS and creative generalization from video generation. Specifically, we reconstruct each captured sequence using 4DGS and render videos along diverse camera trajectories, generated by randomly sampling starting and ending positions within a 2–10 meter radius and linearly interpolating between them to create smooth motion paths. To further enrich the data with lighting diversity, we apply a generalizable video relighting model [Mei et al. 2025] using HDRI maps from Poly Haven [Haven 2025].

*CG and Real-Life Videos.* To support application in subject–scene composition and evaluate our method in both controlled and real-world settings, we incorporate two additional data sources: CG-rendered scenes and real-life videos. For CG data, we use Blender Cycles to render photorealistic 3D indoor environments with diverse camera trajectories, providing precise camera annotations for training scene-customized models. For real-life data, we capture handheld videos of scenes and dynamic subjects using an iPhone, introducing natural camera motion. We estimate camera poses and intrinsics using CUT3R [Wang et al. 2025], enabling subject and scene customization from accessible, real-life video. These sources enable our framework to generalize beyond studio captures, supporting novel camera paths and real-world contexts.

### 3.2 Multi-subject generation

*Joint Training.* To enable the generation of multiple entities—such as two subjects or a subject and a scene—within the same video under novel contexts, we adopt a joint training strategy. Specifically, we fine-tune the model on separate single-entity customization datasets, where each video contains either a single subject or a single scene. Despite training on disjoint examples, the model

learns to compose multiple customized entities during inference. For multi-subject generation, we further incorporate joint-subject data—videos featuring both subjects together—to improve the realism and coherence of inter-subject interactions. Each customized entity is associated with a unique token during training, and these tokens are combined at inference time to generate videos featuring multiple customized components.

*Customization via Independently Customized Models with Noise Blending.* We propose an alternative to joint training for multi-subject video generation by composing independently customized models at inference time, avoiding the need to retrain for every subject combination. Inspired by OMG [Kong et al. 2024], originally designed for multi-subject image generation, our method leverages independently customized text-to-video (T2V) models—each fine-tuned separately for a specific subject—and integrates their outputs through a noise-blending strategy. The method comprises two stages: The first step is focused on **Spatio-Temporal Layout**. We first generate a coarse layout video without identity-specific customization, using a generic prompt (e.g., “a man and a woman in a coffee shop”) that excludes subject-specific tokens. This video establishes plausible spatial and temporal arrangements of subjects. We then use SA2VA [Yuan et al. 2025b], based on SAM2 [Ravi et al. 2024], to segment each subject, producing spatio-temporal masks  $M_i$ . These masks are extended across the entire video by assigning each pixel to the nearest segmented region, ensuring complete spatial-temporal coverage. The second step is **personalized generation via Noise Blending**. To retain layout while incorporating subject-specific identities, we adopt a two-phase denoising procedure. Beginning from the same initial noise seed as Step 1, we first perform the initial 10% of denoising steps without customization to preserve the coarse spatial-temporal layout. This threshold provides a balance: setting it too high reduces identity fidelity, while setting it too low destabilizes the scene layout. Subsequently, at each remaining timestep  $t$ , we predict the next latent state  $z_{t-1}^i$  separately for each subject  $i$  using its corresponding customized model  $T2V^i$ :  $z_{t-1}^i = T2V^i(z_t, p^i, t)$ , where  $p^i$  is a modified version of the original prompt in which only the target subject is replaced by its specific identity token. We then blend these predictions according to the segmentation masks:  $z_{t-1} = \sum_i M_i * z_{t-1}^i$ . This approach accurately customizes each subject region while ensuring coherent global spatio-temporal consistency.

## 4 Experiments

### 4.1 Datasets

*4DGS from Volumetric Captures.* To validate our method in a controlled studio setting, we capture facial and full-body performances of two subjects, “Alex” and “Emily.” Each sequence is reconstructed using 4DGS [Yang et al. 2024b], and rendered with randomly generated moving camera trajectories to create diverse multi-view videos. In total, each subject has 256 videos across 8 performance sequences. To introduce lighting diversity, we apply a generalizable video relighting model [Mei et al. 2025] using HDRI maps from Poly Haven [Haven 2025], generating an additional 128 relit videos per subject. For multi-subject generation, we render 27 joint-subject videos featuring both subjects across 3 sequences.

*CG Scenes.* We construct a dataset of 10 artist-designed indoor scenes rendered with Blender Cycles, each featuring 24 diverse camera trajectories. After filtering out invalid paths, 16 valid trajectories per scene are retained for subject–scene composition. Each rendered video includes ground-truth camera annotations.

*Real-Life Videos.* We capture two indoor environments and one dynamic subject (a cat) using handheld iPhone videos, each about one minute long and recorded while walking to introduce natural camera motion and viewpoint diversity. Each video is split into 20 two-second clips, with camera poses and intrinsics estimated using CUT3R [Wang et al. 2025]. This setup supports effective subject and scene customization under real-world conditions.

### 4.2 Baselines

*Camera Control.* We select 2 contemporary camera control models: AC3D [Bahmani et al. 2025a] and CameraCtrl [He et al. 2024b]. We choose AC3D’s implementation based on CogVideoX [Yang et al. 2024a], a diffusion transformer based model, whereas CameraCtrl uses a UNet-based video diffusion backbone [Guo et al. 2024].

*Personalization.* To validate our method can generate character in consistent with the identity source, we compare our method with ConsisID [Huang et al. 2024a], Videobooth [Jiang et al. 2024b], Magic-Me [Ma et al. 2024b], Dreamvideo [Wei et al. 2024] and Motionbooth [Wu et al. 2024b]. Notably, MotionBooth also enables camera control during personalization. However, its camera motion is limited to 2D movements, such as panning, which restricts its applicability. In contrast, our method can generate 3D camera motions for personalized subjects, offering greater flexibility and realism.

### 4.3 Evaluation metrics

*Multi-view Identity Preservation.* We collect 10 reference images per subject, capturing a variety of viewpoints, including frontal and profile views. For all methods, we first generate videos using the same set of 100 text prompts. Next, on the generated frames, we run SCRFD face detection [Guo et al. 2021] and compare the similarity between the cropped face and the reference faces using AdaFace [Kim et al. 2022]. We then take the maximum similarity score across multiple reference faces from the same identity source. We discard frames where no face is detected.

*Camera Control.* We use rotation and normalized translation errors [He et al. 2024b] estimated by CUT3R [Wang et al. 2025] assess camera steerability. We evaluate all methods on the same 200 test text prompts and camera trajectories from the RealEstate10K dataset.

*General Video Quality.* We assess text controllability by computing the average CLIP [Radford et al. 2021] similarity between the prompt and generated frames. Temporal consistency is measured via average CLIP image similarity between consecutive frames. Additionally, we benchmark our method using four metrics from VBench [Huang et al. 2024b], specifically evaluating subject consistency, background consistency, and temporal flickering to provide a comprehensive assessment of video quality.



Fig. 3. Customization results of T2V baselines and our method, demonstrating superior multi-view identity preservation by our approach.

#### 4.4 Identity preservation

*Comparison with Baselines.* The quantitative comparison of identity preservation and general video generation quality is shown in table 1, with illustrative examples from each method presented in fig. 3. Notably, for multi-view identity preservation, our model achieves the highest AdaFace score among all baselines, surpassing ConsisID, which relies solely on a single facial image during inference and consequently struggles to maintain consistent identities across multiple views. This underscores the importance of utilizing multi-view datasets, as employed by our method, for enhancing identity consistency in personalized multi-view video generation. Although the fine-tuning-based MagicMe method achieves superior scores in terms of subject consistency, background consistency, and reduced temporal flickering, we observe that it suffers from substantial identity degradation and generates notably less motion compared to our approach. Also, MotionBooth, which enables 2D camera control, fails to generate identity-preserving videos.

*User Study on Baseline Comparison.* We conducted a user study with 19 participants to compare video generation methods on multi-view identity preservation, facial realism, and text alignment across 60 prompts featuring two reference identities (Emily or Alex). Participants viewed multi-view reference images of each identity and selected the best video per criterion. Our method was preferred in 81.3% of cases for identity preservation, 70.6% for facial realism, and 74.1% for text alignment.

*Effects of Multi-view Data.* To further investigate the impact of multi-view training data, we performed an ablation study using only frontal-view training images. Quantitative results are presented in table 1, showing a noticeable decrease in AdaFace scores compared to models trained on the complete multi-view dataset. Qualitative examples (fig. 4) also illustrate poorer identity preservation from side-view angles when trained exclusively on frontal-view data. These results underscore the importance of multi-view data for effective multi-view identity preservation.

*Effects of Relit Data.* To evaluate the impact of relit 4DGS data on lighting realism, we conducted an ablation study comparing models trained with and without it. In a user study across 60 prompts, 18 participants preferred the relit-data model in 83.9% of cases. As



Fig. 4. Generated videos with multi-view data (top) and frontal-view-only data (bottom). Multi-view training yields markedly better identity preservation across viewpoints.

shown in fig. 5, relit data significantly enhances lighting realism, while its absence results in flatter illumination.



Fig. 5. Generated videos with additional relit data (top) and without (bottom). Relit data significantly enhances lighting realism and diversity.

#### 4.5 Camera control

A quantitative comparison of camera controllability is presented in table 2. Our pretrained camera-conditioned model achieves the lowest translation and rotation errors, establishing a strong foundation for further camera-conditioned customization.

*Multi-view Customization with Camera Control.* Figure 8 shows qualitative examples of videos generated after subject-specific customization with camera control, along with the corresponding input camera trajectories. Our method faithfully follows the input camera path, generating temporally coherent videos where the subject's appearance is consistently maintained across different viewpoints. It

Table 1. Quantitative results for customization. User study columns report the percentage of responses favoring each method for each evaluation criterion.  $\uparrow/\downarrow$  indicates a higher/lower value is better. **Bold** indicates the best results. Gray-shaded cells indicate values that are significantly lower than the others.

	AdaFace $\uparrow$	CLIP-T $\uparrow$	CLIP-I $\uparrow$	Subject Consistency	VBench Background Consistency	Temporal Flickering	Dynamic Degree	Multi-view Identity	UserStudy $\uparrow$ Facial Realism	Text Alignment
<b>Text-to-video customization</b>										
MagicMe	0.280	0.303	<b>0.991</b>	<b>0.978</b>	<b>0.967</b>	<b>0.984</b>	0.15	3.18%	10.13%	1.85%
DreamVideo	0.194	0.318	0.961	0.893	0.918	0.958	0.44	0.98%	0.99%	1.16%
VideoBooth	0.279	0.274	0.966	0.909	0.941	0.967	0.55	1.54%	0.99%	1.62%
MotionBooth	0.191	0.324	0.954	0.905	0.904	0.927	<b>1.0</b>	2.20%	0.66%	1.16%
ConsisID	0.301	0.355	0.981	0.882	0.892	0.963	0.36	12.96%	17.30%	21.29%
Ours	<b>0.351</b>	<b>0.356</b>	<b>0.991</b>	0.933	0.946	0.975	0.72	<b>81.34%</b>	<b>70.59%</b>	<b>74.07%</b>
Ours (frontal-only)	0.327	0.353	0.989	0.929	0.952	0.980	0.59	-	-	-
<b>Image-to-video customization</b>										
Non-customized	0.324	0.343	0.980	0.898	0.925	0.953	0.78	34.57%	-	-
Ours-I2V	<b>0.350</b>	<b>0.345</b>	<b>0.984</b>	<b>0.908</b>	<b>0.932</b>	<b>0.960</b>	0.72	<b>65.43%</b>	-	-

also handles background changes smoothly and ensures that the subject, background, and lighting evolve together in a visually plausible manner as the camera moves. This coherence is supported by our multi-view-aware customization training, which enables the model to maintain consistent subject identity and spatial relationships under dynamic camera motion.

*Effects of Moving Camera during Customization.* To study the effect of camera motion during customization, we fine-tune the model on two dataset variants: one with dynamic trajectories (Ours-customized) and one with static cameras (Ours-customized-static). As shown in table 2, removing camera motion increases translation and rotation errors, highlighting its importance for maintaining camera controllability.

Table 2. Quantitative comparisons for camera control.  $\uparrow/\downarrow$  indicates a higher/lower value is better. **Bold** indicates the best results.

	TransErr $\downarrow$	RotErr $\downarrow$	CLIP-T $\uparrow$	CLIP-I $\uparrow$
CameraCtrl	0.522	0.163	0.301	0.965
AC3D	0.310	0.112	0.329	0.990
Ours	<b>0.267</b>	<b>0.047</b>	<b>0.332</b>	<b>0.991</b>
Ours-customized	0.324	0.086	0.321	0.993
Ours-customized-static	0.482	0.125	0.330	0.991

*Real-life Videos.* Beyond 4DGS human data, we demonstrate the effectiveness of our method on a real-world customization dataset. Figure 9 showcases a dataset featuring a cat captured under diverse camera angles, viewpoints, poses, and environments, along with generated videos of the same cat in novel contexts. The results show that our method successfully preserves the cat’s identity across views and supports controllable camera motion, highlighting its applicability to in-the-wild scenarios.

#### 4.6 Multi-subject generation

*Multi-subject Interaction.* Figure 10 presents qualitative examples of videos generated by our customized model with camera control,

which was trained on separate single-subject customization datasets (each video containing either subject) as well as a small joint-subject dataset featuring both subjects together. The results demonstrate that our model can accurately generate both subjects in the same scene, maintaining strong multi-view identity consistency for each individual. Moreover, the interactions appear natural and coherent, showing the model captures both individual traits and their spatial and behavioral relationships.



Fig. 6. Generated videos with (top) and without (bottom) joint-subject data. Including joint-subject data improves multi-subject interaction quality.

*Effect of joint-subject data.* To assess the impact of joint-subject data—videos featuring both subjects—we conducted an ablation study comparing models trained with and without it. In the ablated setting, the model was trained only on single-subject datasets, where each video contained either subject but not both. As shown in fig. 6, adding joint-subject data improves spatial relationships and interaction realism. A user study across 60 prompts with 18 participants further confirmed this: 72.9% preferred videos from the model trained with joint-subject data, underscoring its importance for realistic multi-subject interaction.

*Noise Blending Enables Synergy between Independently Customized Models.* We use a noise blending technique to combine independently customized models at inference time for multi-subject video generation. As shown in fig. 11, this approach successfully generates scenes featuring both subjects and captures plausible interactions

between them—even though no model was trained on data containing both individuals. The noise blending technique achieves an AdaFace score of 0.320, slightly lower than the 0.337 obtained with joint training. However, unlike joint training, it enables flexible and modular multi-subject generation without requiring retraining on combined datasets.

#### 4.7 Image-to-video customization

*Effect of Image-to-Video Customization.* As shown in fig. 12, the customized I2V model generates identity-consistent videos for both single and multiple subjects while accurately following camera trajectories. To assess the need for I2V customization—even when the initial frame from the T2I model is accurate—we compare a customized I2V model fine-tuned on subject-specific multi-view data with a pre-trained, non-customized model. (1) Qualitative results in fig. 7 show that the non-customized model exhibits significant identity drift, while the customized model maintains subject’s identity. (2) Quantitative results in table 1 show higher AdaFace scores and improved subject consistency, background consistency, and temporal stability, confirming that I2V customization yields more robust results for personalized subjects. (3) In a user study with 18 participants across 60 prompts, 65.4% preferred the customized model in terms of identity preservation. Together, these results highlight the importance of customizing the I2V model for achieving high-quality, identity-consistent video generation.



Fig. 7. Generated videos with (top) and without (bottom) image-to-video customization. Customization improves multi-view identity preservation.

#### 4.8 Other applications

*Scene Customization.* Beyond subject customization, our method supports scene customization to enable subject–environment interactions. We fine-tune on scene-specific datasets with varying camera trajectories. Figure 13 shows a single-subject case with natural scene interaction, while Figure 14 illustrates realistic multi-subject interactions within the customized scene.

*Customization with Motion and Layout Control.* Beyond camera control, we enable control over subject motion and spatial layout by fine-tuning the Go-with-the-Flow T2V model [Burgert et al. 2025] on our customization dataset. This model uses optical flow from a source video as a control signal, allowing synchronized control over camera and object movement. Given a source video with humans, we aim to preserve its motion and layout while replacing the subjects with customized ones. We extract optical flow and pair it with the source prompt as input to the customized model. As shown in fig. 15,

the generated videos largely maintain the original motion and layout while generating the target subject’s appearance.

### 5 Conclusion

We have introduced a framework that addresses two key challenges in video generation for filmmaking: customization for multi-view identity preservation and precise camera control. Central to our approach is a novel customization data pipeline that combines volumetric capture, 4DGS-based re-rendering for diverse, accurately annotated camera trajectories, and relightable augmentation. Fine-tuning on this dataset improves identity fidelity, camera conditioning, and lighting adaptability. Additionally, our framework supports core virtual production capabilities, including multi-subject generation via joint training and noise blending, scene and real-life customization, and motion-aware spatial layout control. Together, these components offer a scalable and flexible solution for controllable, high-fidelity video generation in virtual production. Our limitations include the need for fine-tuning to fully leverage high-quality multi-view 4DGS data and the low resolution of the CogVideoX backbone, which underutilizes these higher-resolution inputs.

### Acknowledgments

We would like to express our gratitude to Stephan Trojansky and Jeffrey Shapiro for their initial and ongoing executive support; Sebastian Sylwan, Daniel Heckenberg, Jitendra Agarwal, Matheus Leao, and Sungmin Lee for their IT support; Xueming Yu and David George for their hardware support; Jennifer Lao and Lianette Alnaber for their operational support; and Winnie Lin, Lukas Lepicovsky, Ashish Rastogi, Ritwik Kumar, Cornelia Carapcea, and Girish Balakrishnan for their insightful technical discussions.



Fig. 8. Text-to-video (T2V) generation results with camera control for a single subject.

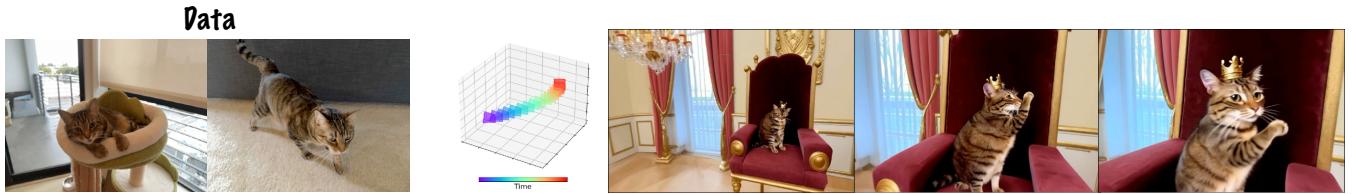


Fig. 9. Customization results from real-life videos of a cat.

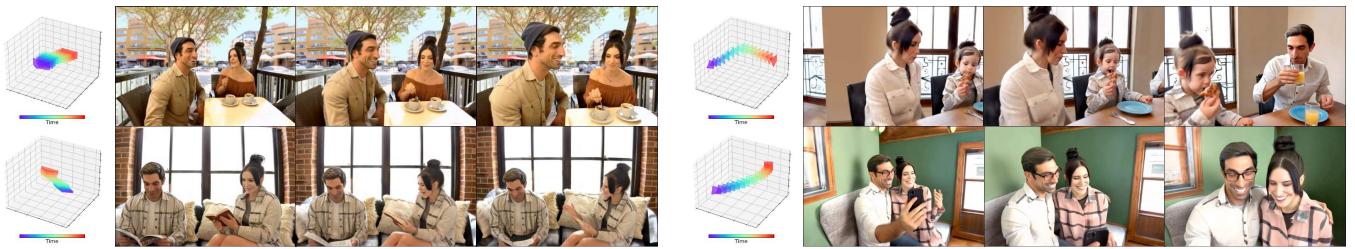


Fig. 10. Text-to-video (T2V) generation results with camera control for multiple subjects.

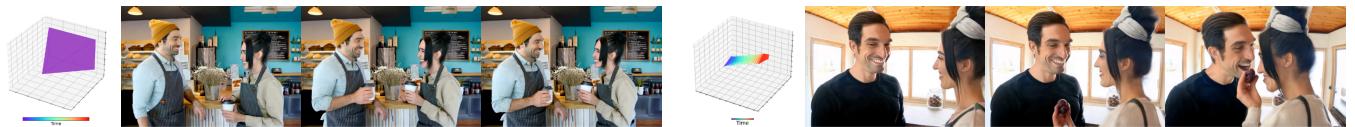


Fig. 11. Generation results using noise blending with independently customized models for multi-subject video synthesis.

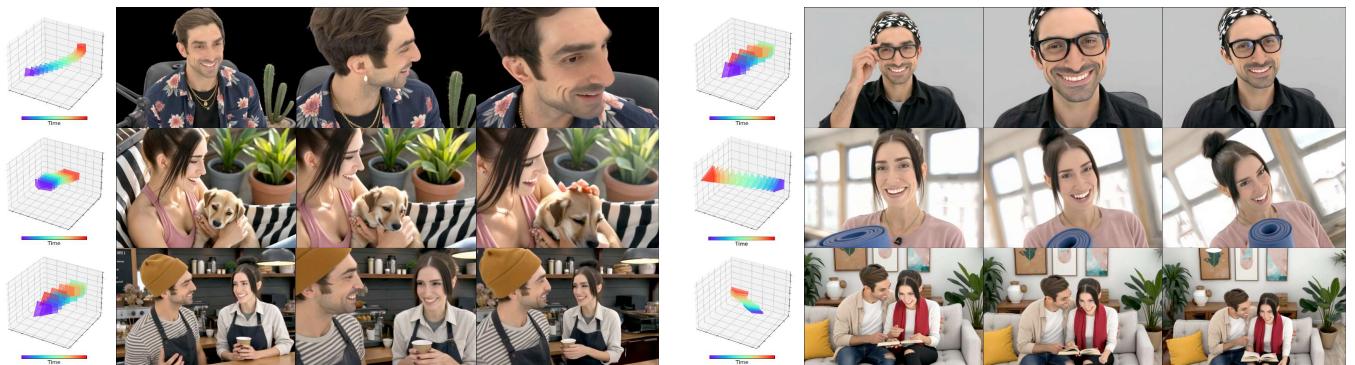


Fig. 12. Image-to-video (I2V) generation results with camera control for single and multiple subjects.

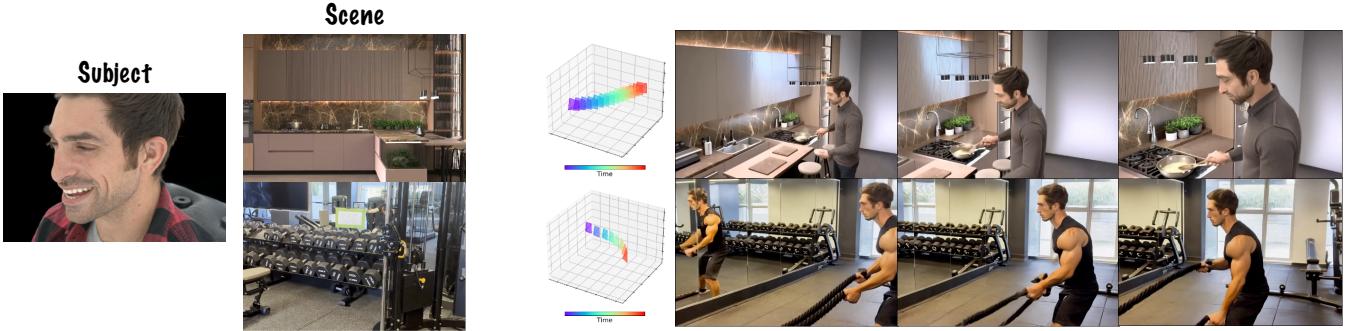


Fig. 13. Customization of a single subject across two scenes with camera control, demonstrating subject–scene interaction during generation.

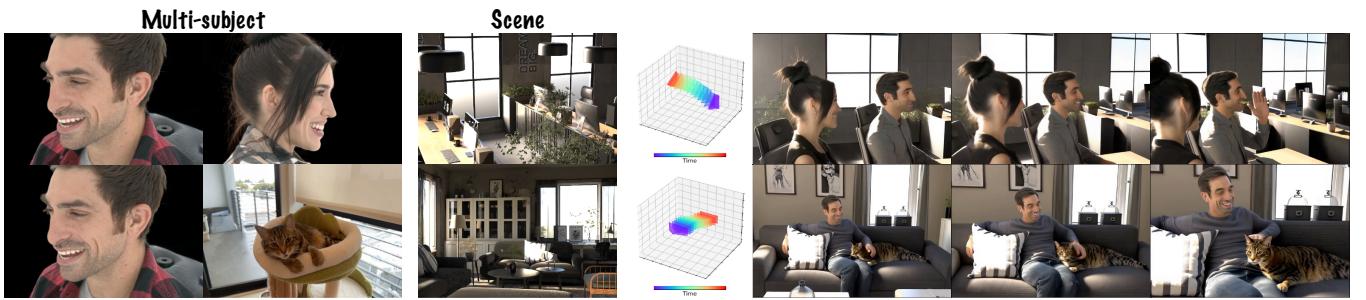


Fig. 14. Customization of multiple subjects across two scenes with camera control, demonstrating both subject–scene and inter-subject interactions during generation.



Fig. 15. Customization with motion and layout control. The first and third rows show generated videos; the second and fourth rows show the corresponding source videos. The results demonstrate that the generated subjects preserve both spatial layout and motion from the source videos.

## References

- Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. 2025a. AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers. *CVPR* (2025).
- Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. 2025b. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control. *ICLR* (2025).
- Jianhong Bai, Menghan Xia, Xiao Fu, Xiantao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. 2025. ReCamMaster: Camera-Controlled Generative Rendering from A Single Video. *arXiv preprint arXiv:2503.11647* (2025).
- Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul A. Beardsley, Craig Gotsman, Robert W. Sumner, and Markus H. Gross. 2011. High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph.* 30, 4 (2011), 75. doi:10.1145/2010324.1964970
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv* (2023).
- Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*.
- Ryan Burgerl, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. 2025. Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise. *CVPR* (2025).
- Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. 2010. Probabilistic Deformable Surface Tracking from Multiple Videos. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 6314)*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer, 326–339. doi:10.1007/978-3-642-15561-1\_24
- Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. 2025. Multi-subject Open-set Personalization in Video Generation. *arXiv* (2025).
- Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Ivan Skorokhodov, Jun-Yan Zhu, Kfir Aberman, Ming-Hsuan Yang, and Sergey Tulyakov. [n.d.]. VideoAlchemy: Open-set Personalization in Video Generation. ([n.d.]).
- Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baqoan Chen. 2024. 4D-Rotor Gaussian Splatting: Towards Efficient Novel View Synthesis for Dynamic Scenes. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) (*SIGGRAPH '24*). Association for Computing Machinery, New York, NY, USA, Article 87, 11 pages. doi:10.1145/3641519.3657463
- Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul E. Debevec. 2011. Comprehensive Facial Performance Capture. *Comput. Graph. Forum* 30, 2 (2011), 425–434. doi:10.1111/j.1467-8659.2011.01888.x
- Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2021. Sample and Computation Redistribution for Efficient Face Detection. *arXiv* (2021).
- Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*.
- Poly Haven. 2025. Collection of HDRIs from Poly Haven. <https://polyhaven.com/hdris> Accessed: 2025-03-07.
- Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024b. Cameractrl: Enabling camera control for text-to-video generation. *arXiv* (2024).
- Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. 2025. CameraCtrl II: Dynamic Scene Exploration via Camera-controlled Video Diffusion Models. *arXiv preprint arXiv:2503.10592* (2025).
- Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgerl, Ning Yu, et al. 2024a. DiffRelight: Diffusion-Based Facial Performance Relighting. In *SIGGRAPH Asia*.
- Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. 2024. Training-free Camera Control for Video Generation. *arXiv* (2024).
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.
- Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, et al. 2024a. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv* (2024).
- Yuzhou Huang, Ziyang Yuan, Quande Liu, Qulin Wang, Xiantao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. 2025. ConceptMaster: Multi-Concept Video Customization on Diffusion Transformer Models Without Test-Time Tuning. *arXiv preprint arXiv:2501.04698* (2025).
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024b. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*.
- Mustafa Isik, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *TOG* (2023).
- Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. 2024. Identity Decoupling for Multi-Subject Personalization of Text-to-Image Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=tEEpVPDafR>
- Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024a. HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19734–19745.
- Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. 2024b. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *TOG* (2023).
- Minchul Kim, Anil K Jain, and Xiaoming Liu. 2022. Adaface: Quality adaptive margin for face recognition. In *CVPR*.
- Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizehu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. 2024. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*.
- Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. 2024. Collaborative Video Diffusion: Consistent Multi-video Generation with Camera Control. *arXiv* (2024).
- Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. 2024. Wonderland: Navigating 3D Scenes from a Single Image. *arXiv* (2024).
- Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaiyan Chen, Jiaqi Wang, and Yi Jin. 2024. MotionClone: Training-Free Motion Cloning for Controllable Video Generation. *arXiv* (2024).
- Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. 2024a. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv* (2024).
- Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Qian He, and Xinglong Wu. 2025. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079* (2025).
- Shilong Liu, ZhaoYang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*.
- Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. 2019. Neural volumes: learning dynamic renderable volumes from images. *ACM Trans. Graph.* 38, 4 (2019), 65:1–65:14. doi:10.1145/3306346.3323020
- Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. 2024. VideoStudio: Generating Consistent-Content and Multi-Scene Videos. In *ECCV*.
- Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In *3DV*.
- Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2024a. 3D Gaussian Blend-shapes for Head Avatar Animation. In *ACM SIGGRAPH Conference Proceedings, Denver, CO, United States, July 28 - August 1, 2024*.
- Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyi Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. 2024b. Magic-me: Identity-specific video customized diffusion. *arXiv* (2024).
- Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, Vishal M Patel, and Paul Debevec. 2025. Lux Post Facto: Learning Portrait Performance Relighting with Conditional Video Diffusion and a Hybrid Dataset. *CVPR* (2025).
- Sida Peng, Chen Geng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2023. Implicit Neural Representations with Structured Latent Codes for Human Body Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.
- Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2024. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20299–20309.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khader, Roman Rädele, Chloe Rolland, Laura Gustafson, et al. 2024. Sam

- 2: Segment anything in images and videos. *arXiv* (2024).
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*.
- Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. 2021. Light field networks: Neural scene representations with single-evaluation rendering. *NeurIPS* (2021).
- Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. 2024. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv* (2024).
- Wan Team. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. (2025).
- Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. 2025. Continuous 3D Perception Model with Persistent State. *arXiv* (2025).
- Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. 2024a. Humanvid: Demystifying training data for camera-controllable human image animation. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshu Chen, Menghan Xia, Ping Luo, and Ying Shan. 2024b. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*.
- Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J. Fleet. 2024. Controlling Space and Time with Diffusion Models. *arXiv* (2024).
- Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. 2024. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*.
- Zijian Zhang Rox Min Zuozhuo Dai Jin Zhou Jiangfeng Xiong Xin Li Bo Wu Jianwei Zhang Katrina Wu Qin Lin Aladdin Wang Andong Wang Changlin Li Duojuin Huang Fang Yang Hao Tan Hongmei Wang Jacob Song Jiawang Bai Jianbing Wu Jinbao Xue Joey Wang Junkun Yuan Kai Wang Mengyang Liu Pengyu Li Shuai Li Weiyang Wang Wenqing Yu Xinchi Deng Yang Li Yanxin Long Yi Chen Yutao Cui Yuanbo Peng Zhentao Yu Zhiyu He Zhiyong Xu Zixiang Zhou Zunnan Xu Yangyu Tao Qinglin Lu Songtao Liu Dax Zhou Hongfa Wang Yong Yang Di Wang Yuhong Liu Weijie Kong, Qi Tian and along with Caesar Zhong Jie Jiang. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv* (2024).
- Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. 2024b. Motionbooth: Motion-aware customized text-to-video generation. *arXiv* (2024).
- Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 2024a. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv* (2024).
- Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. 2024. Video Diffusion Models are Training-free Motion Interpreter and Controller. *NIPS* (2024).
- Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. 2024. CamCo: Camera-Controllable 3D-Consistent Image-to-Video Generation. *arXiv* (2024).
- Zhuoyi Yang, Jianyan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024a. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv* (2024).
- Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. 2024b. Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting. In *International Conference on Learning Representations (ICLR)*.
- Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025. TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models. *arXiv preprint arXiv:2503.05638* (2025).
- Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv* (2024).
- Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shuping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. 2025b. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv preprint arXiv:2501.04001* (2025).
- Shanghai Yuan, Jinfu Huang, Xianyi He, Yunyan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. 2025a. Identity-Preserving Text-to-Video Generation by Frequency Decomposition. In *CVPR*.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2025. Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport. *ICLR* (2025).
- Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817* (2018).

## A Appedix

We conducted four user studies, detailed below.

### A.1 User study on identity preservation compared to baselines

We conducted a user study with 19 participants to evaluate six different video generation methods in terms of multi-view identity preservation, facial realism, and text-video alignment. Participants assessed videos generated by both baseline methods and our approach across 60 prompts, each associated with one of two reference identities (Emily or Alex).

For each prompt, participants were provided with multi-view reference images of the target identity and asked to select the best-performing video for each evaluation criterion. The evaluation questions were as follows:

- **Multi-view identity preservation:** “Which video best preserves the identity of the subject in the reference images?”
- **Facial realism:** “Which video shows the most natural and realistic integration of the human face (i.e., the face does not appear artificially pasted onto the scene)?”
- **Text-video alignment:** “Which video aligns most accurately with the given text prompt in terms of content, actions, and overall depiction?”

Our method was selected as the best in 81.3% of responses for multi-view identity preservation, 70.5% for facial realism, and 74.1% for text-video alignment. A screenshot and detailed statistics of the user study are shown in fig. 16.

### A.2 User study on the effect of joint-subject data

We conducted a user study across 60 prompts with 18 participants to evaluate the impact of joint-subject data on multi-subject video generation. Participants were asked to select the video that best preserved subject identity and depicted the most natural interaction between the subjects. The evaluation questions were:

- **Multi-view identity preservation:** “Which video best preserves the identity of the subject in the reference images?”
- **Natural multi-subject interaction:** “In the videos, two people interact with each other. Which video depicts the most natural and realistic interaction between them?”

The results show that the model trained with joint-subject data was preferred in 63.4% of cases for identity preservation and 72.9% of cases for natural interaction, highlighting the importance of joint-subject examples for improving multi-subject interaction realism. A screenshot and detailed statistics of the user study are shown in fig. 17.

### A.3 User study on the effect of relit data

To evaluate the impact of incorporating relit 4DGS videos into the training data—aimed at enhancing lighting realism and variability—we conducted an ablation study comparing models trained with and without relit data. In a user study, 18 participants viewed pairs of videos generated by these two models across 60 prompts and selected the video with more realistic lighting and better identity preservation. The evaluation questions were:

- **Multi-view identity preservation:** “Which video best preserves the identity of the subject in the reference images?”
- **Natural lighting:** “In some videos, the lighting appears uniformly flat, while in others, there is more variation in brightness and shadows, making the scene look more natural. Which video has the most realistic lighting with noticeable variations instead of flat illumination?”

The results show that the model trained with relit data was preferred in 83.9% of cases for lighting realism and 63.0% of cases for identity preservation, indicating that incorporating relit data not only improves lighting quality but also benefits identity consistency. A screenshot and detailed statistics of the user study are shown in fig. 18.

### A.4 User study on customizing image-to-video model

We conducted a user study with 18 participants across 60 prompts to evaluate the effect of customizing the image-to-video (I2V) model on identity preservation. For each prompt, participants were shown a pair of videos—one generated by a non-customized I2V model and the other by a customized I2V model—and asked the following question:

- **Multi-view identity preservation:** “Which video best preserves the identity of the subject in the reference images?”

The results show that 65.4% of participants preferred the videos generated by the customized I2V model, highlighting the importance of customizing the I2V model for achieving high-quality, identity-consistent video generation. A screenshot and detailed statistics of the user study are shown in fig. 19.

**Input**

**A**

**B**

**C**

**D**

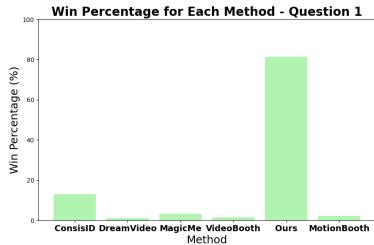
**E**

**F**

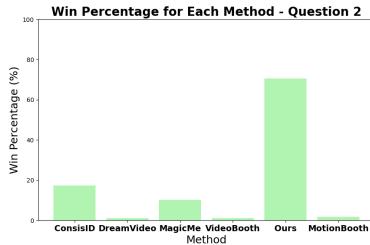
**Prompt:** A video of AS man in a red headscarf and black attire shows him warmly embracing a colorful patterned cushion. AS man smiles gently, occasionally glancing around the room, which is adorned with woven wall decor and a shelf of baskets and plants. A softly glowing candle and vibrant yellow flowers add warmth to the cozy setting.

1. Which video best preserves the identity of the subject in the input reference?  
 A  B  C  D  E
2. Which video shows the most natural and realistic integration of the human face (i.e., the face does not appear artificially pasted onto the scene)?  
 A  B  C  D  E
3. Which video aligns most accurately with the given text prompt in terms of content, actions, and overall depiction?  
 A  B  C  D  E

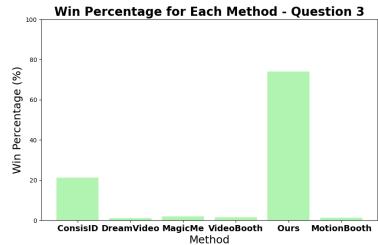
(a) User-study interface and questions for comparison with baselines.



(b) Statistics for the first question on identity preservation.



(c) Statistics for the second question on natural integration of human face.



(d) Statistics for the third question on text video alignment.

Fig. 16. Screenshots and statistics of the user-study questionnaire on comparison with baselines in text-to-video customization. Our method secures the highest user-preference share in all questions.



A video of AS man and EW woman captured from the front, in a bright, minimalist setting, showcases a joyous moment shared between AS man and EW woman. EW woman, dressed in a black dress, is seen with a hand on hip, while AS man, also wearing a stylish black dress, links arms with EW woman with a black prosthetic arm. Both AS man and EW woman are adorned with subtle jewelry, radiating confidence and camaraderie, their expressions filled with laughter and warmth.

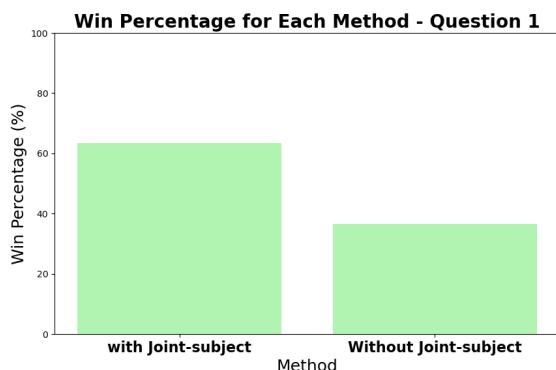
1. Which video best preserves the identity of the subjects in the input reference?

A     B

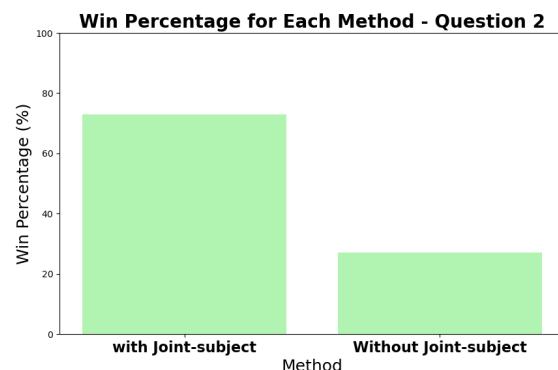
2. In the videos, two people interact with each other. Which video depicts the most natural and realistic interaction between them?

A     B

(a) User-study interface and questions for the effects of joint-subject data.



(b) Statistics for the first question on identity preservation.



(c) Statistics for the second question on natural human interaction.

Fig. 17. Screenshots and statistics of the user-study questionnaire on the effects of the joint-subject data in multi-subject generation. Using jointly-subject data secures the higher user-preference share in both questions.



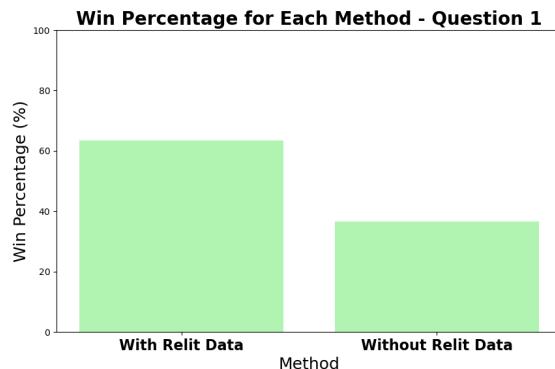
1. Which video best preserves the identity of the subject in the input reference?

A     B

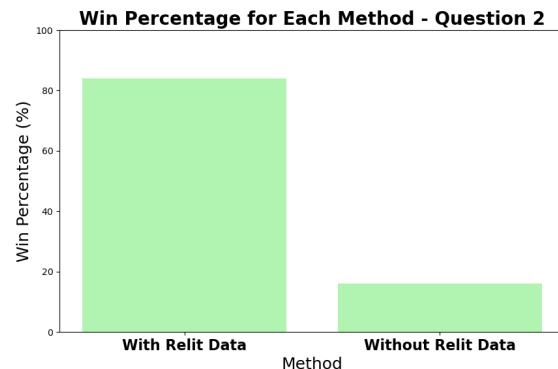
2. In some videos, the lighting appears uniformly flat, while in others, there is more variation in brightness and shadows, making the scene look more natural. Which video has the most realistic lighting with noticeable variations instead of flat illumination?

A     B

(a) User-study interface and questions for the effects of relit data.



(b) Statistics for the first question on identity preservation.



(c) Statistics for the second question on realistic lighting.

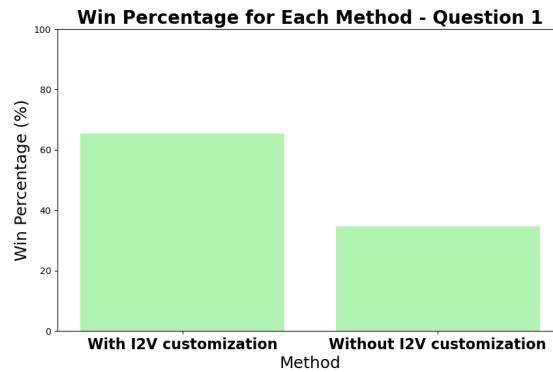
Fig. 18. Screenshots and statistics of the user-study questionnaire on the effects of the relit data data in generating videos with realistic and diverse lighting. Adding relit data secures the higher user-preference share in both questions.



1. Which video best preserves the identity of the subject in the input reference?

A     B

(a) User-study interface and questions for the effects of customizing image-to-video model.



(b) Statistics for the question on identity preservation.

Fig. 19. Screenshots and statistics of the user-study questionnaire on the effects of customizing image-to-video model in preserving the identity of the subjects. Customizing image-to-video model secures the higher user-preference share in identity preservation.