arXiv:2504.09472v3 [cs.CV] 27 Mar 2026

Zero-Shot Personalized Camera Motion Control for Image-to-Video Synthesis

Pooja Guhan[1], Divya Kothandaraman[1], Geonsun Lee[1], Tsung-Wei Huang[2], Guan-Ming Su[2], Dinesh Manocha[1]

[1]University of Maryland College Park, Maryland, USA, 20742,.
[2]Dolby Laboratories, Sunnyvale, California, USA, 94085.

Contributing authors: pguhan@umd.edu; dkr@umd.edu; gsunlee@umd.edu; tsung-wei.huang@dolby.com; guanming.su@dolby.com; dmanocha@umd.edu;

Abstract

Specifying nuanced and compelling camera motion remains a significant hurdle for non-expert creators using generative tools, creating an "expressive gap" where generic text prompts fail to capture cinematic vision. This barrier limits individual creativity and restricts the accessibility of cinematic production for small-scale industries and educational content creators. To address this, we present a zero-shot diffusion-based framework for personalized camera motion control, enabling the transfer of cinematic movements from a single reference video onto a user-provided static image without requiring 3D data, predefined trajectories, or complex graphical interfaces. Our technical contribution involves an inference-time optimization strategy using dual Low-Rank Adaptation (LoRA) networks, with an orthogonality regularizer that encourages separation between spatial appearance and temporal motion updates, alongside a homography-based refinement strategy that provides weak geometric guidance. We evaluate our approach using a new metric, CameraScore, and two distinct user studies. A 72-participant perceptual study demonstrates that our method significantly outperforms existing baselines in motion accuracy (90.45% preference) and scene preservation (70.31% preference). Furthermore, a 12-participant task-based interaction study confirms that our workflow significantly improves usability and creative control ($p < 0.001$) compared to standard text- or preset-based prompts. We hope this work lays a foundation for future advancements in camera motion transfer across diverse scenes.

Keywords: camera motion, creativity support, machine learning, video editing

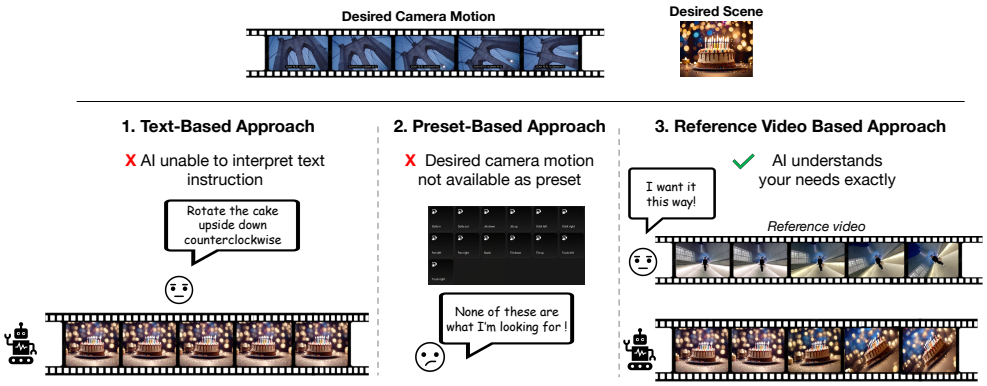


Fig. 1 A comparison of three types of instruction interface available to replicate a desired camera motion on a desired scene. (1) Text Based Approach: The instruction may not be interpreted by AI correctly (2) Preset Based Approach: The preset with the exact camera motion may not be available for AI to use (3) Reference Video Based Approach: Intuitive input and AI understands what you need exactly.

1 Introduction

Digital video has become a vital medium for communication and creative expression. The rise of generative artificial intelligence (AI) [1] offers unprecedented opportunities for both professional and novice creators, enabling the synthesis of personalized content with unique characters [2], customized backgrounds [3], and tailored motions [4, 5]. Recent advances in personalized image [6–8] and video generation [9–11] empower creators to infuse their work with distinctive elements, pushing the boundaries of creativity well beyond training data limitations and finding significant application in animation and filmmaking [12].

Yet, a critical component of cinematic narrative remains elusive: *camera motion* [13]. This *subtle but impactful visual shifting and movement guides perspective*, evokes emotion [14–16], and often determines the difference between a static scene and a captivating sequence. Current generative systems offer only limited control. Abstract text prompts (e.g., "zoom in") lack the granularity for nuanced cinematic intent, while graphical motion panels require technical skill and patience that reinforces the expressive gap for casual creators. This barrier not only limits individual creativity but also restricts the accessibility of cinematic production for small-scale industries and educational content creators, aligning with broader concerns about inclusive innovation highlighted in UN Sustainable Development Goal 9 [17].

To bridge this gap, we introduce *exemplar-based camera motion personalization*: the ability to transfer the specific cinematic movements observed in a reference video onto a new scene of interest. In this paradigm, users can *show and tell* the desired motion by supplying a short clip whose camera behavior they like, and applying that motion to their own still image. This capability democratizes video production [18], allowing creators to borrow and adapt professional techniques effortlessly. However, existing methods fall short. Video foundation models like MovieGen [19] and Veo [20]

produce generic motion effects tied to text prompts, not personalized trajectories. Related motion transfer works, such as COMD [21] and AnimateDiff [9], rely on multiple reference videos, which is poorly suited for capturing the distinctive pacing and angles of a single exemplar. While recent foundation models and motion transfer methods enable impressive video synthesis, they either expose only coarse text-level control or require multiple references, explicit 3D trajectories, or specialized motion modules. We instead focus on the practical scenario of *single-exemplar, cross-scene camera motion transfer* onto *arbitrary* static images using off-the-shelf diffusion backbones.

Controllable camera motion is central not only to professional cinematography but also to emerging AI-mediated media workflows, from short-form online video to virtual production and immersive experiences. By lowering the technical barrier to reusing cinematic camera styles, our method illustrates how AI for video processing can support creative industries, democratize access to advanced visual storytelling techniques, and contribute to innovation in digital content production.

Main Contributions

We present a method for *personalized camera motion transfer* from a *single* reference video to a static image, empowering creators to realize their own scenes with cinematic movements directly inspired by professional footage. Our contributions are summarized as follows:

1. We introduce a zero-shot approach that *transfers camera motion from a single reference video to a static image*, without requiring additional camera trajectories or 3D reconstruction. Our key technical component is an inference-time optimization on a pretrained text-to-video diffusion model using dual LoRA networks with an orthogonality regularizer that encourages separation between spatial appearance updates and temporal motion updates. We further leverage homography representations from classical computer vision to guide the generated video along the desired camera motion, producing coherent and realistic sequences.
2. For quantitative evaluation of this task, we introduce CameraScore, a homography-based metric that assesses temporal consistency between the dominant camera-induced motion in the generated video and in the reference, and we report it alongside established representation and text–video alignment metrics. Extensive quantitative and qualitative analyses, including a comparison study with 72 participants, demonstrate that our method significantly outperforms relevant prior methods, achieving 90.45% preference for motion accuracy and 70.31% preference for scene preservation.
3. We design an intuitive workflow that allows creators to animate still images simply by providing a reference video, avoiding reliance on abstract prompts or generic presets. A user study with 12 participants confirms that our interface improves usability and creative control significantly compared to standard text- or preset-based prompts ($p<0.001$).

2 Related Works

AI techniques have been widely applied to video and motion processing tasks, including reconstructing sharp video bursts from a single motion-blurred image [22] and stereoscopic video deblurring with transformer architectures [23]. These works illustrate the broader role of AI in understanding and manipulating motion in video. In this paper, we focus instead on controllable camera motion transfer and creator-facing interaction, rather than recovering latent motion from degraded footage.

2.1 User Interfaces for Video Creation and Controllability

Traditional video editing tools pose significant barriers due to complexity, driving research toward accessible interfaces leveraging natural language and sketching [24]. Modern generative models offer powerful text-to-video synthesis [20, 25], but *lack geometric control*. Integrating cinematic camera motion *requires specialized expertise* for manual calibration, homography estimation, and 3D trajectory reconstruction—a major hurdle for non-specialists [26]. While recent works such as VGGT [27] provide strong automatic estimators of camera motion, depth and point clouds from monocular videos, they still require non-trivial integration and may not cover the full diversity of the in-the-wild, stylized or low-light footage creators wish to repurpose. Our approach takes a *complementary* stance. Instead of predicting full 3D trajectories, we use homography-based 2D cues for weak guidance that can be obtained directly from the reference video and seamlessly plugged into existing diffusion pipeline.

2.2 Video Personalization, Controllability, and Motion Transfer

Advancements in pretrained models [28–30] have fueled interest in fine-grained control [9, 31–36].

Fine-tuning and Local Motion: Zero-shot methods like Tune-A-Video [32, 37] use spatial/temporal LoRA [38] to personalize based on a few images/videos. While effective for local subject motion, they struggle with camera motion transfer due to the difficulty in separating scene characteristics from camera dynamics during fine-tuning.

Explicit Control & Trajectories: Other methods use specialized motion modules [9, 39] or require explicit camera inputs such as 3D trajectories and SfM poses [40–42] or detailed natural language descriptions of camera behavior [43]. Approaches based on 3D/4D reconstruction, including ReCapture [41], also incur substantial computational and data costs. In a related direction, CamCloneMaster [44], ReCamMaster [45], and Akira [46] explore camera control or cinematic transfer using multiple reference videos or NeRF/ray-based scene representations, and therefore assume access to accurate camera poses or structured scene geometry. By contrast, we focus on single-exemplar, cross-scene camera motion transfer onto arbitrary static images using only 2D cues derived from the reference video and off-the-shelf text-to-video diffusion backbones, without requiring any additional 3D reconstruction or motion annotations.

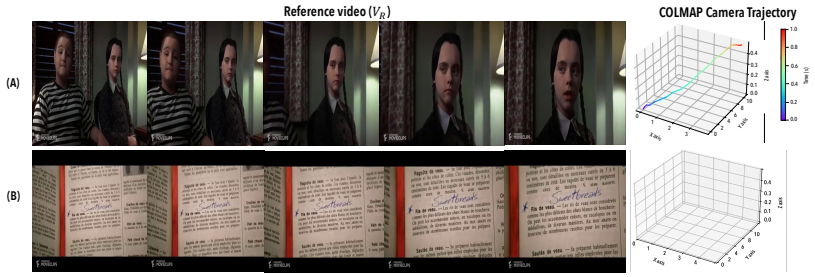


Fig. 2 This showcases examples to depict scenarios where COLMAP can and cannot work reliably. In video (A), the camera effect observed is due to the explicit movement of the camera while in (B), the effect is obtained due to changes in camera focal length. COLMAP is not able to converge for videos like (B).

2.3 Image Personalization and Novel View Synthesis (NVS)

Image personalization excels at customizing objects, poses, and styles [3, 47–54]. However, applying these to video motion is challenging due to the missing temporal dimension. Similarly, methods like ControlNet [55] require substantial supervised fine-tuning.

Our task, transferring camera motion to a static image, is akin to Novel View Synthesis (NVS) [41, 56, 57], which generates frames from new perspectives. A key limitation of NVS, however, is the typical requirement for detailed camera trajectory information to define the desired view angle. Robust 4D NVS also often struggles with generalizability and requires extensive training data. Our objective is unique: to develop a *generalizable, data-efficient, zero-shot* approach that implicitly transfers camera motions from a single reference video, thereby circumventing the need for explicit trajectory input or costly domain-specific data collection.

3 Method

To enable personalized and intuitive camera motion transfer (CMT), our system utilizes a pretrained text-to-video diffusion model conditioned on visual content and structured text prompts. This foundational method powers our zero-shot transfer pipeline, integrating complex camera motion from a single reference video onto a user-provided image without requiring explicit 3D reconstruction.

Problem Description. Given a *single* reference video V_R (with prompt t_R), a user image I_u (with prompt t_u), our goal is to generate a new video V_u by transferring the camera motion from V_R onto I_u in a zero-shot manner. We use a pretrained text-to-video diffusion backbone, e.g., zeroscope [58]. t_R is structured as a two-part, comma-separated prompt: visual content of V_R, followed by 'camera motion of $< v >$', where $< v >$ is a random token to handle arbitrary camera motions. t_u describes the content of I_u.

Key Challenges. Approaching CMT by first extracting camera trajectories using methods like COLMAP and then integrating them with diffusion models [40] has two key issues: *(1)* Camera motions arise from explicit movement or focal length

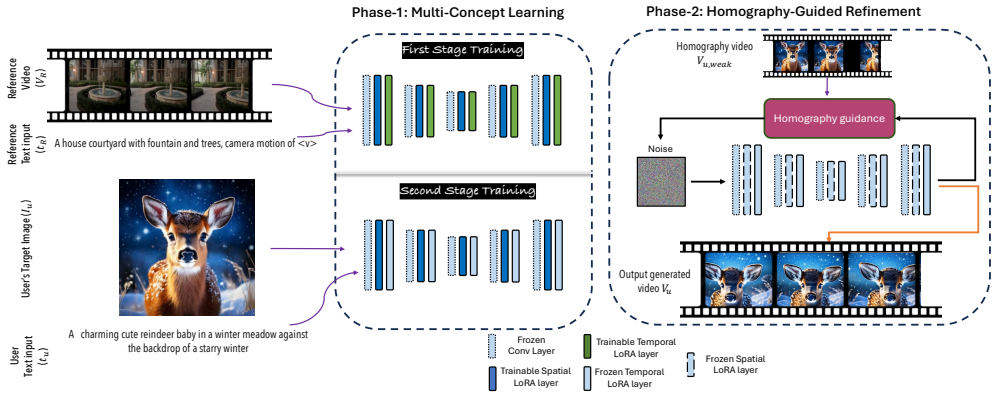


Fig. 3 We present a zero-shot method to transfer camera motion visible in a reference video V_R onto a user provided image I_u to generate video V_u. It's a two-phase algorithm. The first phase involves learning multiple concepts associated with the spatial and temporal features of the reference video as well as the spatial characteristics of the user-provided image. We propose the use of a spatial-temporal orthogonality loss to better learn these concepts. The second phase consists of using a homography-based guidance to refine the generated video to preserve the scene in I_u as well as the camera motion obtained from the reference video V_R.

changes (e.g., dolly zoom). COLMAP *struggles* with the latter (Fig. 2). *(2)* These methods *require training specialized modules*, which is *impractical for personalization* or limited-example scenarios. To overcome these, we propose a novel *zero-shot* strategy.

Method Overview.

Our two-phase algorithm is shown in Fig. 3: (i) *Multi-concept finetuning* that learns to separate temporal and spatial features from V_R and blend the spatial characteristics of I_u using a novel orthogonality based regularizer, while allowing them to be integrated into a single controllable model; (ii) *Homography-guided inference* to ensure V_u preserves both the I_u scene and the V_R camera motion. We now discuss each of them in detail.

3.1 Phase-1: Multi-Concept Finetuning

Network Structure: Our zero-shot algorithm requires discerning three CMT-relevant concepts: V_R's *spatial characteristics*, V_R's *camera motion (temporal) features*, and I_u's *spatial characteristics*. We use a *multi-concept* LoRA fine-tuning approach on the text-to-video backbone. LoRA [38] enables efficient fine-tuning by constraining the weight update $W_0 + \Delta W = W_0 + BA$, where $B \in \mathcal{R}^{d \times r}$, $A \in \mathcal{R}^{r \times k}$, and rank r is small:

$$W_0 + \Delta W = W_0 + BA \tag{1}$$

We use dual LoRA matrices: *spatial LoRAs* in spatial self-attention layers to capture spatial characteristics, and *temporal LoRAs* in temporal self-attention layers to model camera motion features. Cross-attention layers remain fixed.

Training Strategy: We employ a *two-stage* training paradigm.
First stage training: This differential training captures V_R's spatial and temporal concepts. Temporal LoRAs are trained on all V_R frames for comprehensive motion dynamics. Spatial LoRAs are trained on a single, random V_R frame per step, promoting generalization. We use random masking for spatial LoRAs [59] for regularization. The training loss is:

$$\mathcal{L}_{first_stage} = \mathcal{L}_{temporal} + \delta\mathcal{L}_{spatial} \tag{2}$$

where $\mathcal{L}_{temporal}$ and $\mathcal{L}_{spatial}$ are the noise prediction losses [60] for their respective LoRAs, and $\delta \in (0, 1)$.
Second stage training: This stage adapts the model to I_u while maintaining V_R's temporal characteristics. Spatial LoRAs are fine-tuned using I_u [61], leveraging the spatial-temporal relationship learned from V_R in the first stage. To prevent spatial fine-tuning from suppressing temporal features, we introduce an orthogonality loss ($\mathcal{L}_{ortho}$). This constraint ensures spatial features remain orthogonal to temporal features, preserving the temporal dynamics while seamlessly integrating I_u's spatial characteristics. The second stage training loss is:

$$\mathcal{L}_{second_stage} = \mathcal{L}_{spatial} + \lambda\mathcal{L}_{ortho} \tag{3}$$

where λ is a hyperparameter and:

$$\mathcal{L}_{ortho} = \phi(W_{spatial_LoRA}, W_{temporal_LoRA})_k \tag{4}$$

$\phi(x, y)_k$ is the inner product between the k most significant values of $W_{spatial_LoRA}$ and $W_{temporal_LoRA}$ (the respective LoRA weight parameters).

3.2 Phase-2: Homography Guided Inference

We introduce a weak supervision strategy during the *inference denoising* stage to synchronize motion and appearance in V_u. We leverage *homography* [62], proven effective for novel view synthesis [63], as a weak signal for how I_u's scene elements *might* change under the desired camera motion. Homography computes the transformation mapping one image onto another (rotation/translation) using key point matching (e.g., SIFT [64]) and RANSAC [65].

Inference begins with the fine-tuned backbone using prompt $t_I = t_u$+'camera motion of $< v >$'. We compute frame-to-frame homography matrices H_i (from $F_{R,i}$ to $F_{R,i+1}$) using the reference video V_R. We then generate a *pseudo-weak* estimate, $V_{u,weak}$, starting with I_u as the first frame ($F_{P,1}$). Subsequent frames are $F_{P,i+1} = H_i(F_P, i)$, where $i \in [1, N-1]$. At each diffusion denoising step (except the last), the predicted latents z_t are adjusted [63, 66] based on the latents z_P corresponding to F_P:

$$\hat{z}_t = z_t - \lambda_G \nabla(z_t - z_P)^2 \tag{5}$$

3.3 CameraScore - A New Metric

Popular motion transfer metrics like optical flow [67] and COLMAP [68] are unsuitable for cross-scene CMT due to *structural incompatibility, scaling differences, video quality dependency, and computation cost* (details in Appendix).

Therefore, we propose CameraScore, a homography-based metric for evaluating camera motion transfer across structurally different scenes:

$$\frac{1}{N}\sum_{i=1}^{N-1}||\mathcal{H}_{R_i} - \mathcal{H}_{G_i}||^2 \tag{6}$$

where $\mathcal{H}_{R_i}$ is the homography matrix between the i^{th} and $(i+1)^{th}$ frame of the reference video, and $\mathcal{H}_{G_i}$ is the same for the generated video.

Why Homography? Homography estimation effectively captures relative transformations (panning, tilting, zooming) between two planes, aligning perfectly with our goal to transfer holistic camera motion rather than pixel-level details (unlike optical flow). By considering homography between consecutive frame pairs, we capture local motions that approximate planar or rotational transformations. The cumulative effect represents the complex camera path, providing a valuable and computationally efficient metric compared to COLMAP or optical flow.

4 Experiments and Results

4.1 Dataset

Since *no* dataset exists specifically for zero-shot single-exemplar CMT, we curated a diverse dataset of 680 reference video and user scene combinations, comparable in size to works like DreamBooth [69] and MotionDirector [70]. Reference videos, 2–4 seconds long, feature a wide range of cinematic motions: panning, zooming, tilting, dolly shots, and 3D rotations, capturing both *subtle* and *dynamic* styles from movies, documentaries, and animation. User scenes were generated using SOTA text-to-image models, spanning landscapes and urban settings. We used BLIP to caption images; three experienced annotators provided the content descriptions for t_R. This diversity ensures rigorous testing of our method's robustness and adaptability. We will release this dataset post-acceptance.

4.2 Implementation Details

Model Details. We use *zeroscope* [58] as the text-to-video backbone, generating 16 frames per sample.

Training Details. Models are fine-tuned on a *single* A5000 GPU. Both training stages run for 150 epochs each, taking approximately 10 minutes total per sample to generate V_u. This two-stage adaptation is performed per reference video-user image pair and is intended as a personalized model creation step, similar to DreamBooth-style workflows, rather than a single universal model that instantaneously supports

all motion and scenes. We use the Adam optimizer with a learning rate of $5e-4$. All code is implemented using Pytorch [71].

Metrics. We evaluate using three metrics: DINOv2 [72] (scene preservation/similarity to I_u), VideoCLIP [73] (text consistency/alignment), and our proposed CameraScore (motion consistency with V_R).

4.3 Comparison with Prior Work

Ours is the first paper to address *zero-shot* CMT from a *single* video onto a target scene. SOTA foundation models do not natively support this task. We compare our approach against three related *zero-shot* baseline categories: traditional homography, DreamBooth-style fine-tuning (using Tune-A-Video [37] to introduce scene customization), and video diffusion-based object motion transfer (MotionDirector [70]). Other recent camera-control methods based on explicit trajectories or NeRF-style inputs are not directly comparable in our single-exemplar, arbitrary-image setting. Therefore, we restrict our comparisons to only these zero-shot diffusion-based baselines that can be instantiated under the same assumptions.

Fig. 4 presents the *quantitative comparisons*. The homography baseline accurately captures motion but suffers from outpainting artifacts. Tune-A-Video achieves high scene consistency but minimal camera movement. MotionDirector introduces slight motion but struggles with integrating I_u spatial characteristics with V_R temporal features, often suppressing temporal dynamics.

In contrast, Our method achieves the best trade-off. We achieve a higher VideoCLIP score than MotionDirector and Tune-A-Video, indicating superior textual consistency and better adherence to the motion prompt. Our lower CameraScore compared to MotionDirector signifies better motion transfer fidelity. We also attain a higher DINO score than MotionDirector, showcasing improved scene preservation with respect to I_u. Tune-A-Video, which lacks motion, has an inflated DINO score as it simply replicates I_u. Qualitative results (Fig. 6) further validate our approach's effectiveness in transferring camera motion while preserving scene integrity.

4.4 Ablation Experiments

We ablate the impact of three key components: USC (*U*ser *S*cene *C*haracteristics learning: I_u spatial feature learning in Phase-1, excluding orthogonality loss), ORTHO (orthogonality loss function to prevent I_u spatial features from suppressing V_R temporal features), and HG (*H*omography *G*uidance during Phase-2 inference).

Fig. 5 shows the results. Our full approach achieves the *best* trade-off. Whe USC is removed (removing I_u information), the network fails to transfer motion effectively, resulting in a significantly higher CameraScore; HG becomes ineffective without the user scene as a base. When only USC is used, CameraScore deteriorates compared to when ORTHO and HG are included, confirming that without ORTHO, temporal features are disrupted despite better spatial feature capture. Without ORTHO, the camera motion transfer is disrupted as expected. Overall, our findings demonstrate that USC, ORTHO, and HG are complementary and essential for effective motion transfer aligned with the reference video while preserving spatial scene integrity.

4.5 User Study

To evaluate the effectiveness and usability of our method, we conducted two complementary user studies. The first focuses on video quality, assessing how accurately our system reproduces camera motions and preserves user-provided image characteristics. The second focuses on interface usability, comparing different interaction paradigms for directing camera motion. Together these studies provide a holistic understanding of both algorithmic performance and human-centered creative support.

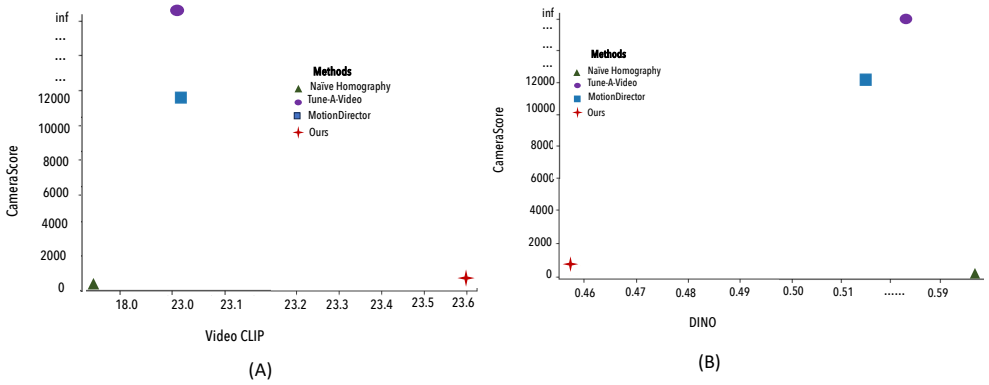


Fig. 4 Graph (a) shows the results of the plot of Video CLIP scores with CameraScore for the different diffusion methods being compared. Graph (b) shows the plot of DINO scores with CameraScore for the different diffusion methods being compared. Our approach achieves the *best trade-off*.

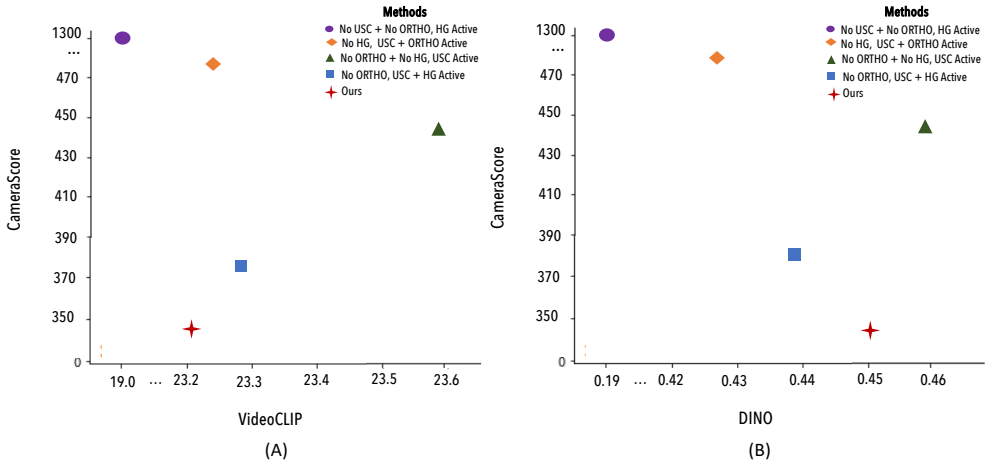


Fig. 5 Graph (A) illustrates the relationship between CameraScore and DINO, while Graph (B) shows the relationship between CameraScore and VideoCLIP. These plots are derived from the ablation experiments conducted to evaluate the significance of various key components in our proposed approach.

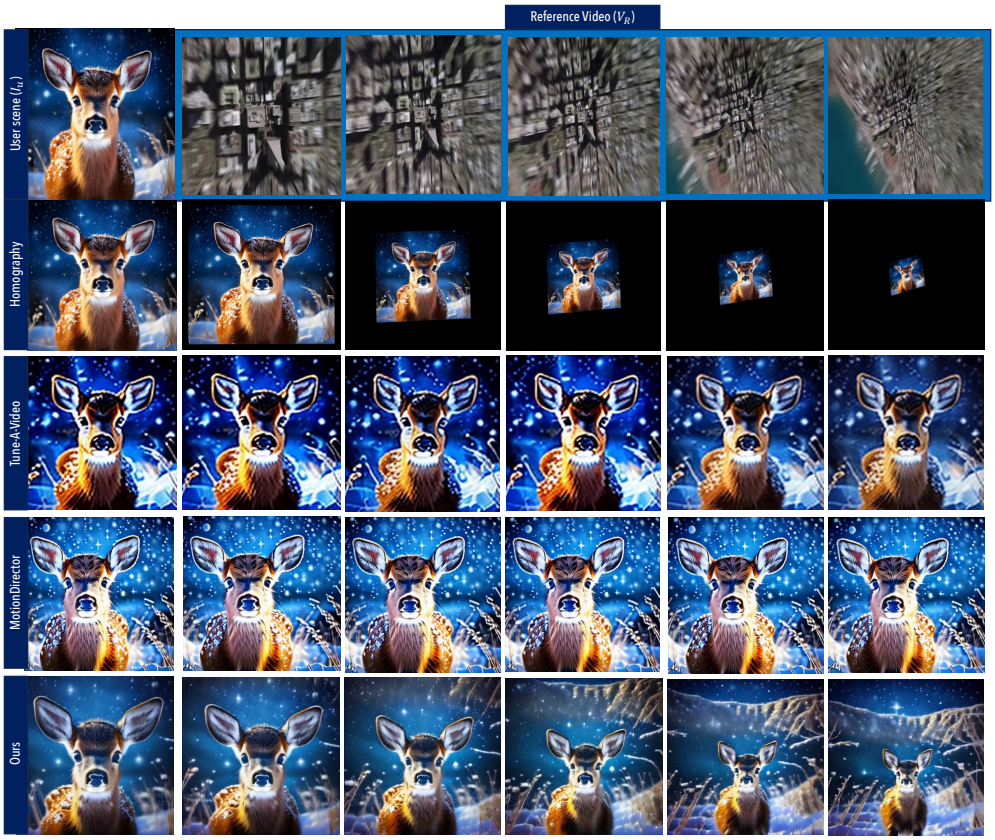


Fig. 6 Qualitative results of our method, demonstrating clear improvements over prior work in transferring camera motion, while preserving scene content (Reference video). *More results in the appendix.*

4.5.1 Video Quality Study

The goal of this study was to validate the fidelity of camera motion transfer while maintaining the visual integrity of the user's scene.

Participants

We recruited 72 participants via an online crowdsourcing platform to evaluate the algorithmic quality of our system's output. The participant pool was diverse in terms of gender and age to reflect a broad, general user base representative of typical video consumers. Specifically, 50% identified as male, 47% identified as female, and 3% preferred not to disclose or identified outside the gender binary. The age group spanned 18 to 65+, with most participants between 24 and 44 years old.

None of the participants were required to have prior expertise in video editing or cinematography, supporting the system's accessibility goals. While not professionally trained, the participants reported frequent exposure to diverse video content in daily

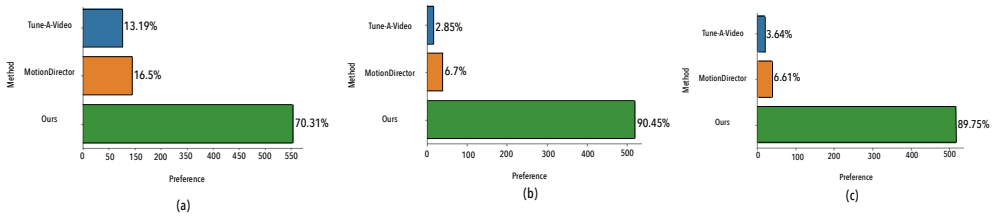


Fig. 7 Graph (a) shows participants evaluating the similarity between the generated video scene and the user-provided scene. Graph (b) depicts participants' preferences for the output video with the desired camera motion among the three methods. Graph (c) indicates participants' choice of the overall best output among the different methods.

life, equipping them with the informal experience necessary to provide meaningful perceptual judgments relevant to non-expert users.

Procedure

To assess the algorithmic quality of our algorithm's camera motion transfer, participants completed a forced-choice comparative evaluation. Each participant was made to evaluate 8 sets of test cases covering a diverse range of camera motions. Each set included:

- A reference video demonstrating the desired camera motion
- The target user image, and
- Outputs from three methods: MotionDirector [70], Tune-A-Video (DreamBooth-modified) [37] and our method.

Participants were asked to compare and choose their preferred output corresponding to each of the following three criteria:

1. **Scene similarity** - how well the generated video preserves the user's image characteristics
2. **Camera motion fidelity** - how closely the camera movement matches the reference video, and
3. **Overall performance** - which output they considered best overall.

A brief tutorial with examples familiarized participants with camera motion transfer concepts and evaluation criteria. To ensure unbalanced judgments, we randomized the order of output options within each test set to prevent positional bias. Additionally, the presentation order of the test sets themselves was randomized for each participant.

Statistical Analysis

For each participant and criterion, we determined which method they selected the most frequently across the 8 trials. These participant-level preferred methods yielded the counts reported in Table 1 and served as the inputs to our inferential tests. Participant preferences were analyzed using a chi-square goodness-of-fit test to evaluate whether

the distribution of choices differed significantly from chance among the three competing methods: MotionDirector, Tune-A-Video(DreamBooth modified) and our method. Under the null hypothesis, each method was equally likely to be preferred at the participant level, with a probability of 1/3 for each method. When the chi-square test indicated a significant deviation from equal preference, we conducted pairwise post-hoc comparisons via binomial tests between our method and each baseline method to identify which pairs contributed to the overall preference differences. Bonferroni correction was applied to adjust for multiple comparisons, setting a significance threshold of $\alpha = 0.025$ for the two post-hoc tests per criterion. Effect sizes were calculated using Cramér's V to quantify the strength of the association between method and preference, with established ndicating small (0.1), medium (0.3) and large (0.5) effects. All statistical analysis were conducted using Python 3.11 and SciPy 1.10.

Results

Table 1 summarizes participant preferences across three criteria: scene similarity, camera motion fidelity and overall performance. Our method was the predominant choice for all criteria, selected by $70.31\%, 90.45\%$ and 89.75% of participants respectively (Figure 7). Post-hoc binomial tests with Bonferroni correction confirmed that our

Table 1 Statistical Analysis of Video Quality User Study

Criterion	Tune-A-Video	MotionDirector	Our Method	Chi-square (χ^2)	p-value	Cramér's	Interpretation
Scene Similarity	9(13.19%)	12(16.5%)	51(70.31%)	45.75	< 0.001	0.56	Large effect
Camera Motion Fidelity	2(2.85%)	5(6.7%)	65(90.45%)	105.25	< 0.001	0.85	Large effect
Overall Performance	3(3.64%)	5(6.61%)	64(89.75%)	100.08	< 0.001	0.83	Large effect

method was significantly preferred over both Tune-A-Video and MotionDirector across all criteria (all adjusted $p < 0.01$). We also observe that the method with the lowest CameraScore (ours) is the one most preferred for motion fidelity in the user study (90.45%), providing qualitative evidence that CameraScore is directionally consistent with human judgments for our scenarios.

4.5.2 User Interaction Study

To understand how effectively our system supports creative workflows, we conducted a comparative study of three interface paradigms for replicating camera motions: (1) text prompt-based controls, (2) preset-based motion panels, and (3) reference video-driven interface (ours). The study aimed to evaluate usability, efficiency, and participants' ability to achieve their intended cinematic outcomes. By observing how participants interact with each interface and collecting subjective feedback, we assessed which design best enables creators to explore, iterate, and express their creative vision.

Participants

We recruited 12 participants (7 female, 5 male) with ages ranging from 19 to 52 years (median = 29.5, SD = 7.42). Participants reported their experience with video editing tools on a 7-point scale, with a median experience level of 4 (SD = 1.26). Their experience with video generation tools was lower, with a median of 2 (SD =

1.35). Participants were recruited through university mailing lists and social media, and received compensation for their time.

Apparatus and Materials

Participants used three different video generation interfaces:

- **Text**: Frame-to-video generation using text prompts with Flow (Veo3 [74])
- **Preset+Text**: Frame-to-video with text prompts plus one of 13 predefined camera motion presets using Veo2 [20]
- **Reference Video (Ours)**: Our developed interface using reference video-driven motion transfer

We selected Google's Flow [75] as our baseline text-based interface due to its state-of-the-art performance in video generation and its widespread adoption in the creative community. Flow represents the current industry standard for text-to-video generation, making it an appropriate comparison point for evaluating alternative interface paradigms. The preset-based condition using Veo2 was included to assess whether structured motion controls could bridge the gap between pure text descriptions and reference-driven approaches.

Each session used standardized input images and inspiration videos to simulate scenarios where users have a specific video outcome in mind.

Procedure

The study followed a within-subjects design with counterbalanced presentation order. Each session began with introductions and consent procedures, followed by an explanation of the study goals and tasks. Participants were instructed to focus on recreating camera movements and effects rather than altering video content, and to disregard video resolution or quality in favor of achieving desired motion effects.

For each interface condition, participants were presented with a creative scenario simulating a common workflow where content creators seek to replicate cinematic effects they have observed. Specifically, participants were shown an input image as their starting frame and an inspiration video demonstrating a desired camera motion or cinematic effect. This setup reflects real-world situations where creators encounter compelling visual content and wish to incorporate similar stylistic elements into their own work.

Participants completed three distinct camera motion tasks in counterbalanced order: (1) an orbit motion involving 3D rotation around a central subject, (2) a zooming motion with varying speed dynamics, and (3) a 2D rotation motion. These tasks were selected to represent common cinematic techniques that vary in complexity and require different levels of spatial understanding and motion control. Participants were then asked to achieve the demonstrated camera motion using the available interface, with unlimited iterations for the Reference Video condition and a maximum of 5 iterations for Text and Preset+Text conditions due to platform token limitations. Participants could iterate until satisfied with their results.

Throughout each trial, we recorded participant's screen to measured task completion time and iteration count from initial idea formulation to final satisfactory

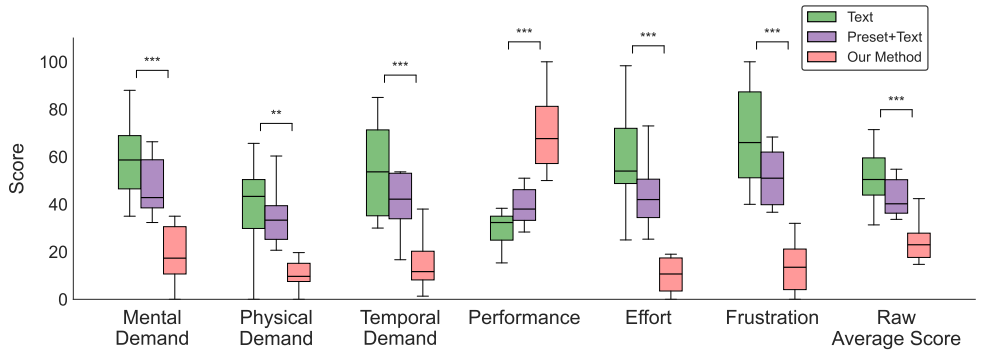


Fig. 8 NASA TLX Results

outcome. After completing each condition, participants filled out questionnaires including the NASA Task Load Index (TLX) [76], System Usability Scale (SUS) [77], and 7-point Likert scales assessing satisfaction with their generated video and preference for the given method.

Measures

We collected both objective and subjective measures to comprehensively evaluate the effectiveness of each interface paradigm. Our objective measures included the number of iterations required to achieve satisfactory results, which served as an indicator of interface efficiency and ease of achieving desired outcomes. We also measured task completion time from initial interaction to final satisfactory outcome, providing insights into the cognitive load and learning curve associated with each interface. Screen recordings of user interactions were captured to enable detailed analysis of user behavior patterns, interaction strategies, and potential usability issues. For subjective assessment, we employed the NASA Task Load Index (TLX) to evaluate perceived workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. This multidimensional assessment helps understand the cognitive burden imposed by each interface design. We administered the System Usability Scale (SUS) to obtain standardized usability scores, enabling comparison with established usability benchmarks and between our three interface conditions. Additionally, we used 7-point Likert scales to assess participants' satisfaction with their generated video outcomes and their overall preference for each method, capturing the subjective quality of the creative experience. Post-session interviews provided qualitative insights into participants' reasoning, preferences, and suggestions for improvement, offering deeper understanding of the quantitative findings.

Results

NASA Task Load Index (TLX)

The NASA-TLX assessment revealed significant differences across all six dimensions of cognitive workload between the three methods. Our Method consistently demonstrated

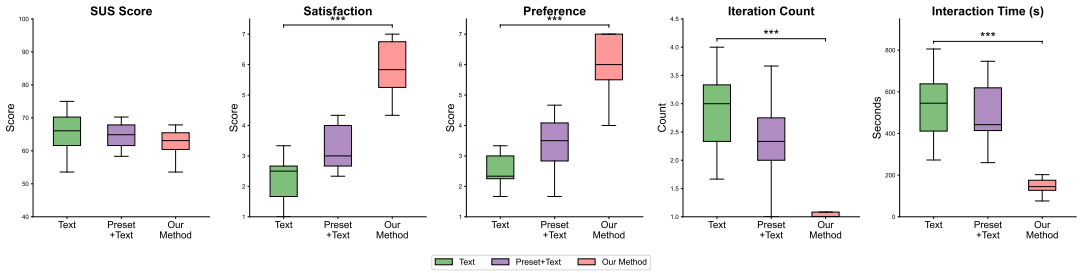


Fig. 9 Interaction Time

the lowest workload scores, indicating superior usability compared to both Text and Preset+Text approaches. (See Figure 8)

Mental demand showed significant differences between methods (ANOVA: $F = 26.62$, $p < 0.001$), with Our Method requiring substantially less mental effort than the baseline approaches. Physical demand also differed significantly (Friedman test: $\chi^2 = 11.64$, $p = 0.003$), though all methods scored relatively low on this dimension as expected for text-based interfaces.

Temporal demand exhibited strong significant differences (Friedman test: $\chi^2 = 18.57$, $p < 0.001$), with Our Method allowing users to complete tasks with less time pressure. Performance ratings differed significantly between methods (Friedman test: $\chi^2 = 21.57$, $p < 0.001$), with Our Method receiving the highest performance scores, indicating users felt they accomplished their goals more effectively.

Effort requirements showed significant variation (Friedman test: $\chi^2 = 20.67$, $p < 0.001$), with Our Method requiring notably less user effort. Frustration levels differed markedly between approaches (ANOVA: $F = 42.13$, $p < 0.001$), with Our Method producing the lowest frustration scores. The overall NASA-TLX scores (Raw Average Score) demonstrated significant differences (ANOVA: $F = 23.60$, $p < 0.001$), confirming Our Method's superior workload profile across all dimensions.

System Usability Scale (SUS)

SUS scores did not reveal significant differences between the three methods (ANOVA: F not reported, $p = 0.230$). See Figure 9). All methods achieved moderate usability scores (Text: $M = 66.77$, $SD = 8.97$; Preset+Text: $M = 64.58$, $SD = 3.90$; Our Method: $M = 62.40$, $SD = 3.98$), suggesting that while users found all approaches reasonably usable, the SUS instrument may not have been sensitive enough to detect the performance differences observed in other measures.

User Satisfaction and Preference

User satisfaction ratings differed significantly between methods (ANOVA: $p < 0.001$), with Our Method receiving substantially higher satisfaction scores ($M = 5.92$, $SD = 0.95$) compared to Preset+Text ($M = 3.25$, $SD = 0.77$) and Text ($M = 2.31$, $SD = 0.72$). This pattern indicates that users found Our Method considerably more satisfying to use.

User preference showed a similar pattern (ANOVA: $p < 0.001$), with Our Method strongly preferred ($M = 5.97$, $SD = 1.06$) over both Preset+Text ($M = 3.36$, $SD = 0.93$) and Text ($M = 2.42$, $SD = 0.65$). These results suggest that users not only performed better with Our Method but also preferred the interaction experience it provided.

Task Efficiency Metrics

Iteration count analysis revealed significant differences between methods (Friedman test: $p < 0.001$). Our Method required significantly fewer iterations to complete tasks ($M = 1.11$, $SD = 0.22$) compared to both Preset+Text ($M = 2.36$, $SD = 0.73$) and Text ($M = 2.94$, $SD = 0.71$). This reduction in required iterations suggests Our Method enabled users to achieve their goals more directly with less trial-and-error.

Interaction time showed dramatic differences between approaches (ANOVA: $p < 0.001$). Our Method required substantially less time to complete tasks ($M = 159.17$ seconds, $SD = 56.43$) compared to Preset+Text ($M = 490.47$ seconds, $SD = 142.51$) and Text ($M = 531.00$ seconds, $SD = 169.60$).

The results demonstrate that Our Method consistently outperformed both baseline approaches across multiple dimensions. Users experienced lower cognitive workload, higher satisfaction and preference, and achieved tasks more efficiently with fewer iterations and significantly less time. While SUS scores showed no significant differences, the convergent evidence from NASA-TLX, satisfaction measures, and objective performance metrics strongly supports the effectiveness of Our Method for this task domain.

These user studies serve as a *complementary evaluation channel* to CameraScore especially for scenes where planar homography assumptions are violated or where subtle cinematic qualities are hard to capture with geometric metrics alone.

5 Discussion

In this section, we reflect on our findings from the design and development of our system to seamlessly transfer camera motions from a reference video to any scene of the user's choice.

Automating Complex Technical Tasks to Foster Accessibility.

Our findings demonstrate that embedding deep automation for technically challenging tasks such as accurate camera motion estimation substantially lowers cognitive barriers for users. By abstracting away complicated mathematical and calibration procedures behind intelligent AI models, systems like ours can empower novices and creators of varying expertise to produce high-quality, professional-grade content effortlessly. This highlights a critical design principle:*effective AI-enhanced tools must prioritize invisibility and simplicity in the user interface while delivering sophisticated capabilities under the hood.* Future designs should ensure that powerful automation complements rather than complicates the creative process, thereby reducing friction and inviting broader adoption.

Harnessing Reference-Driven and Incremental Interaction Paradigm.

In our interaction study, user preferences strongly favored exemplar-based workflows. This emphasizes the need for providing *concrete visual or textual anchor points* facilitates more intuitive and controllable content generation. Users benefit from workflows that allow them to iteratively adjust AI-generated content through small, understandable steps, rather than relying on opaque, fully automated outputs. This incremental, transparent process makes the creative journey more accessible, and predictable. This enables users to maintain control, build confidence and continuously guide AI towards their desired outcome. Designers should move beyond fully automated *black-box* solutions toward collaborative interaction models where users play an active role in guiding the generative behaviors. Such hybrid approaches balance user agency with AI efficiency, enhancing trust and satisfaction.

Building Modular, Extensible, and Future Proof Architecture.

Our system's successful integration of reference video based camera motion inference with generative AI pipelines highlights the importance of *modular and extensible system architectures*. Such designs empower developers to fluidly incorporate state-of-the-art algorithms, respond to emerging user requirements, and maintain system evolvability without disrupting user experience continuity. Prioriziting architecture strategies that enable incremental innovation and cross component synergy will be crucial for sustaining the long-term relevance and usability of AI-powered creative tools.

6 Limitation

Our system represents a significant step forward in zero-shot reference video based camera motion transfer. However, several limitations remain. Technically, our reliance on homography-based guidance can be susceptible to inaccuracies in the presence of substantial foreground object motions within reference videos. This limitation restricts motion transfer fidelity when dynamic objects disrupt stable camera motion estimation. Additionally, our system currently supports camera motion transfer only onto static images rather than full video sequences, limiting its applicability to scenes with moving subjects or complex temporal dynamics. Furthermore, the quality and diversity of generated videos output are inherently bounded by the capabilities of the underlying pretrained generative model, constraining the complexity of camera motions and scene content that can be realistically synthesized. Although we implemented our method on Zeroscope due to availability, the dual-LoRA and homography-guidance strategy is model-agnostic and could be applied to other text-to-video backbones; we leave systematic evaluation across multiple models for future work.

From an interface perspective, although our design lowers barriers by enabling intuitive camera motion specification via single reference videos, it lacks flexibility for more nuanced creative control. Users cannot blend multiple reference camera motions, adjust the point of application of camera motion (for example, allowing users to decide about which object in the scene, should a rotational camera motion occur) or fine-tune camera trajectories interactively. Additionally our user study focused on a small

group of mostly non-expert creators. A more diverse participant base including professional filmmakers and active content creators may reveal additional usability needs and interaction challenges and we plan to evaluate the interface with expert users and more diverse content types in future work.

7 Future Work

Building on these limitations, future research can pursue several promising directions. Enhancing camera motion transfer robustness by integrating object-aware scene understanding, potentially through large language models (LLM) enhanced video analysis could improve handling dynamic objects in reference videos. Extending our system to support full user videos rather than static images would enable richer, temporally coherent personalization workflows. Advances in generative video foundation models open opportunities for our camera motion transfer algorithm to evolve as a modular, plug and play component that leverages improved motion complexity and video fidelity.

From an interaction perspective, users indicated that supporting fusion of multiple reference videos could enrich creative control further. While this capability to the best of our knowledge is not yet algorithmically implemented, designing interfaces that allow users to blend, select or weigh multiple references videos to be incorporated in the same video represents a valuable direction for future interface and algorithm co-design. Additional enhancements could include interactive editing controls to adjust the point of camera motion application and trajectory fine-tuning, as well as recommendation systems for reference video selection. Conducting broader user studies with diverse creator population will further inform the design of expressive accessible tools that align well with varied workflows and expertise levels.

8 Conclusion

We designed and developed a system for reference based camera motion transfer that enables creators to animate static images with cinematic camera movements drawn from a *single* reference video. The system leverages pretrained text-to-video diffusion models and incorporates an inference-time optimization strategy using dual LoRA networks, and homography guidance to achieve coherent and personalized motion effects. Our evaluations included both algorithmic studies, introducing a new homography-based metric and large scale user comparisons and interaction studies comparing interface paradigms for directing camera motion. Results show that our system not only reliably transfers the desired camera motion and preserves user's desired scene compared to prior methods but also enhances usability and creative control for non-expert creators. We hope this work catalyzes future research on human-centered generative video systems that expand the expressive possibility of storytelling.

Table A1 Comparison of our work with related prior works

Method	Zero-Shot strategy	Single Reference Video (input)	User-specified Image (input)	3D data (input)	Output Features
Video Foundation Models [19, 28, 29]	✗	✗	✗	✗	Lacks the ability to transfer camera motion from a video onto another
AnimateDiff-based works [9, 39]	✗	✗	✗	✗	Text to video models that require training a control net and therefore, extensive data corresponding to it.
Uses 3D camera trajectories/3D models [40, 41]	✗	✗	✓	✓	Requires extensive data for training a separate motion and camera control modules or 3D structure-aware modules. Additionally requires extraction of camera trajectories or generation of 3D models
Homography	✓	✓	✓	✗	Can't mimic non-planar camera motions effectively. Black spaces in regions that are outside the boundaries of the user-specified image.
Tune-A-Video based works [32, 37]	✓	✓	✗	✗	Text to image model extended for video generation. Targeted towards object motions. Output video content based on the text prompt provided.
DiTFlow based works†	✓	✓	✗	✗	Can't transfer to images. Video generated using only text prompts.
MotionDirector [70]	✓	✓	✗*	✗	Focuses on object motions. Primarily, output video content based on the text prompt provided
Ours	✓	✓	✓	✗	Require only the reference video and user-specified image with the target being transfer of camera motions

Appendix A Prior Works

We provide a feature-by-feature comparison of prior works in this domain in Table A1. We use ✓ to mark the presence of a feature and ✗ to mark the absence.

For experiments, as discussed in Sec. 4.3, we compare our method only with prior zero-shot-based works like Homography, Tune-A-Video and MotionDirector. We detail these methods and highlight any modifications made to ensure a fair and consistent evaluation below.

Naive Homography. A fundamental approach to understanding scene changes due to camera motion is based on homography. This involves computing homography matrices from transitions observed in the reference video V_R and applying the same sequence to the target image I_u. Specifically, we define:

$$V_u(t) = \mathcal{H}(V_R, t) I_u(t)$$

where $\mathcal{H}(V_R, t)$ represents the homography matrix computed at time step t between consecutive frames of V_R, which is then applied to $I_u(t)$, the corresponding state of I_u. Given the central role of homography in our method, we also investigate the impact of integrating diffusion models for this task. This experiment evaluates the effect of excluding diffusion models on the generated outputs.

Tune-A-Video (Dreambooth-modified). In our method, we have used a video diffusion model. However, these are part of very recent developments within the diffusion literature. Text-to-image models have improved multi-fold since their inception. Tune-A-Video showcased extending these models to generate videos based on the text prompt provided. This method cannot be directly applied to our task, as we need the motion to be applied to a specific scene. We, therefore, modify the method to incorporate a Dreambooth-style [69] training mechanism. This involves finetuning the Tune-A-Video 3D UNet with the video and its corresponding text, followed by finetuning it for the user image I_u. We have used the publicly available implementations of this work and DreamBooth to implement and test this case.

MotionDirector [70]. In contrast to Tune-A-Video, this work does the same task of transferring object motions observed in a reference video but using a video diffusion model. For scene customization, they suggest finetuning two separate UNets - one on the reference video and one on the desired scene. Following this, they replace the spatial LoRA layers in the UNet finetuned on the reference video with the spatial LoRAs of

the UNet finetuned on the image. We use the publicly available implementation for this work.

Appendix B CameraScore

In Section 3.3, we introduced *CameraScore*, a novel metric designed to evaluate the quality of camera motion generated by a method in comparison to the motion in the reference video. A *lower score* indicates better performance. We deliberately avoid assessing camera motion using the estimated 3D trajectory typically obtained via COLMAP, as it proves unreliable or infeasible under specific conditions. For example, when the camera achieves a zoom effect solely by adjusting the focal length without physical movement, or when the scene structure is predominantly flat (i.e., coplanar), COLMAP struggles to generate accurate trajectories.

Figure B1 illustrates this limitation. In Case 1, the zoom-in effect is achieved through the physical movement of the camera toward the subjects, enabling COLMAP to generate a 3D trajectory for the reference video. In contrast, Case 2 demonstrates a zoom-in effect achieved by altering the focal length, with no actual camera movement. Here, COLMAP fails to compute a trajectory. However, in both cases, homography matrices can still be computed. This highlights the advantage of using homography over COLMAP, making it a more robust and versatile metric for evaluating camera motion.

Consider the optical flow maps generated for the reference and generated videos in both Case-1 and Case-2 (Figure B1). While the observed motion in both the reference and generated videos is similar, directly comparing their optical flow maps is challenging due to structural differences between the reference scene and the desired scene. Additionally, optical flow often exhibits significant errors near boundaries, making it unreliable for accurately evaluating the observed camera motions.

Considering the different reference video-image pairs, we have a total of 680 test samples in the dataset. This is comparable to other works in image and video synthesis, such as [69]

Appendix C Qualitative Results

We showcase some qualitative results obtained using our method in Fig.C2-C8. In each example shown, we show the video generated using our method as well as baselines. The same video backbone (zeroscope [58]) has been used in our method and MotionDirector [70] to generate the video output observed. Stable Diffusion v2 [78] has been used as the image generation backbone for generating Tune-A-Video [37] outputs. *The supplementary video shows additional results with different camera motions.*

References

[1] van der Zant, T., Kouw, M. & Schomaker, L. in *Generative artificial intelligence* (ed.Müller, V. C.) *Philosophy and theory of artificial intelligence* 107–120 (Springer, Berlin, Heidelberg, 2013).

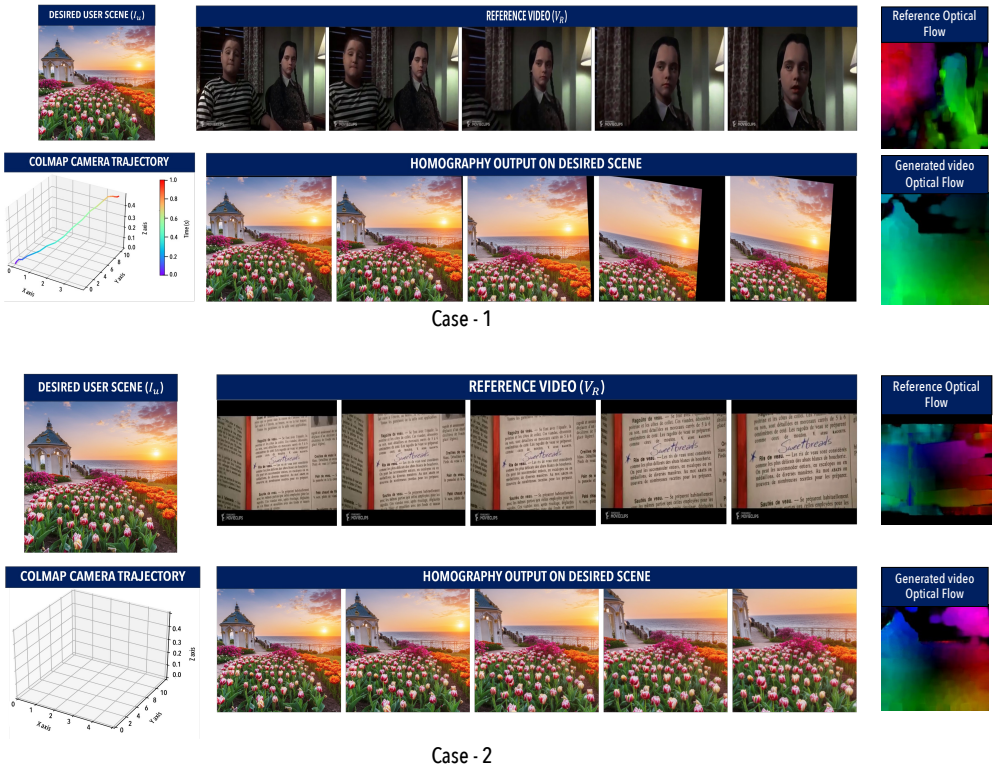


Fig. B1 COLMAP and optical flow are not reliable metrics for comparing the camera motions observed in the reference and the generated videos. COLMAP fails to generate the camera trajectory in Case-2. Comparing the optical flows would be challenging because of structural differences. Therefore, we define *CamerScore*. (Reference video clips: top row,bottom row).

[2] Zhang, Y., Xing, Z., Zeng, Y., Fang, Y. & Chen, K. Society, I. C. (ed.) *Pia: Your personalized image animator via plug-and-play modules in text-to-image models.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7747–7756 (IEEE, 2024).

[3] Li, P. *et al.* Committee, E. . O. (ed.) *Tuning-free image customization with image and text guidance.* (ed.Committee, E. . O.) *European Conference on Computer Vision*, 233–250 (Springer, 2025).

[4] Wu, J. *et al.* Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758* (2024).

[5] Ma, Y. *et al.* Committee, A.-. O. (ed.) *Follow your pose: Pose-guided text-to-video generation using pose-free videos.* (ed.Committee, A.-. O.) *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 4117–4125 (AAAI Press, 2024).

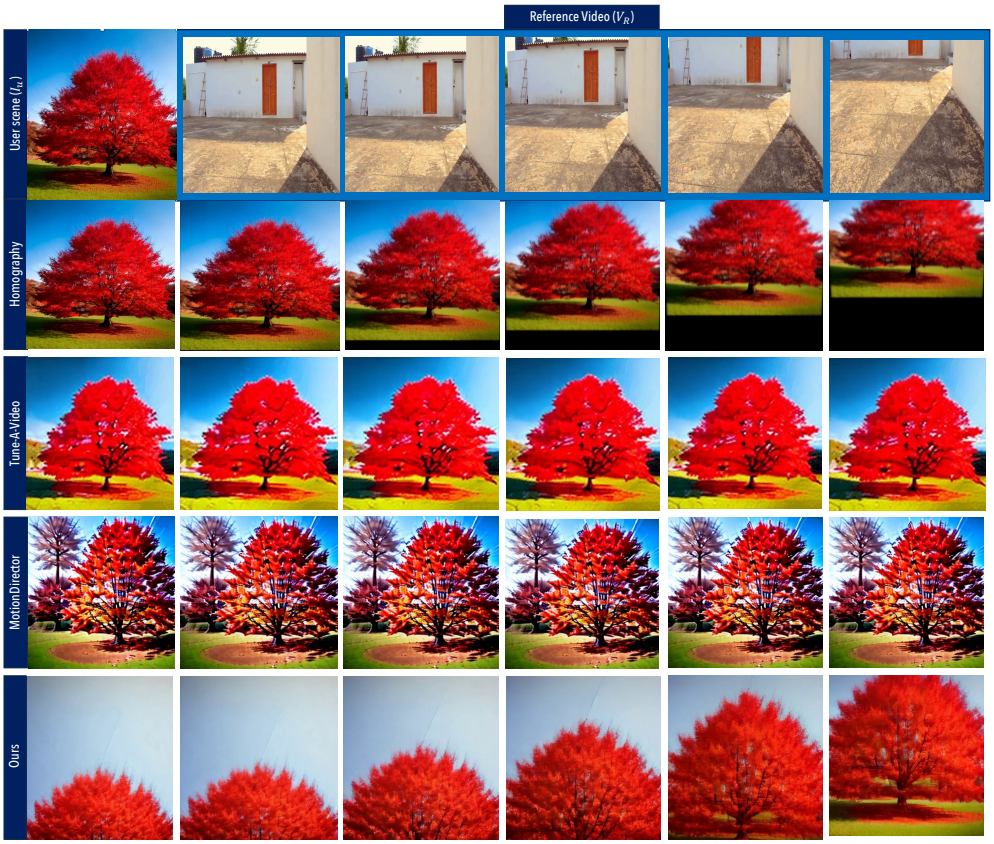


Fig. C2 The first row shows the user-provided scene and the reference video (source) containing camera motion. In contrast to the baselines (Tune-A-Video and MotionDirector), our method more effectively captures the underlying camera motion from the reference video while also better preserving the details of the user-provided scene.

[6] Shi, J., Xiong, W., Lin, Z. & Jung, H. J. Society, I. C. (ed.) *Instantbooth: Personalized text-to-image generation without test-time finetuning.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8543–8552 (IEEE, 2024).

[7] Li, Z. *et al.* Society, I. C. (ed.) *Photomaker: Customizing realistic human photos via stacked id embedding.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8640–8650 (IEEE, 2024).

[8] He, Z. *et al.* Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346* (2024).



Fig. C3 The first row shows the user-provided scene and the reference video (source) containing camera motion. In contrast to the baselines (Tune-A-Video and MotionDirector), our method more effectively captures the underlying camera motion from the reference video while also better preserving the details of the user-provided scene.

[9] Guo, Y. *et al.* Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).

[10] Xing, J. *et al.* Committee, E. . O. (ed.) *Dynamicrafter: Animating open-domain images with video diffusion priors.* (ed.Committee, E. . O.) *European Conference on Computer Vision*, 399–417 (Springer, 2024).

[11] Ma, W. *et al.* Video diffusion generation: comprehensive review and open problems. *Artificial Intelligence Review* **58**, 338 (2025).

[12] Totlani, K. The evolution of generative ai: Implications for the media and film industry. *International Journal for Research in Applied Science and Engineering Technology* **11**, 973–980 (2023).

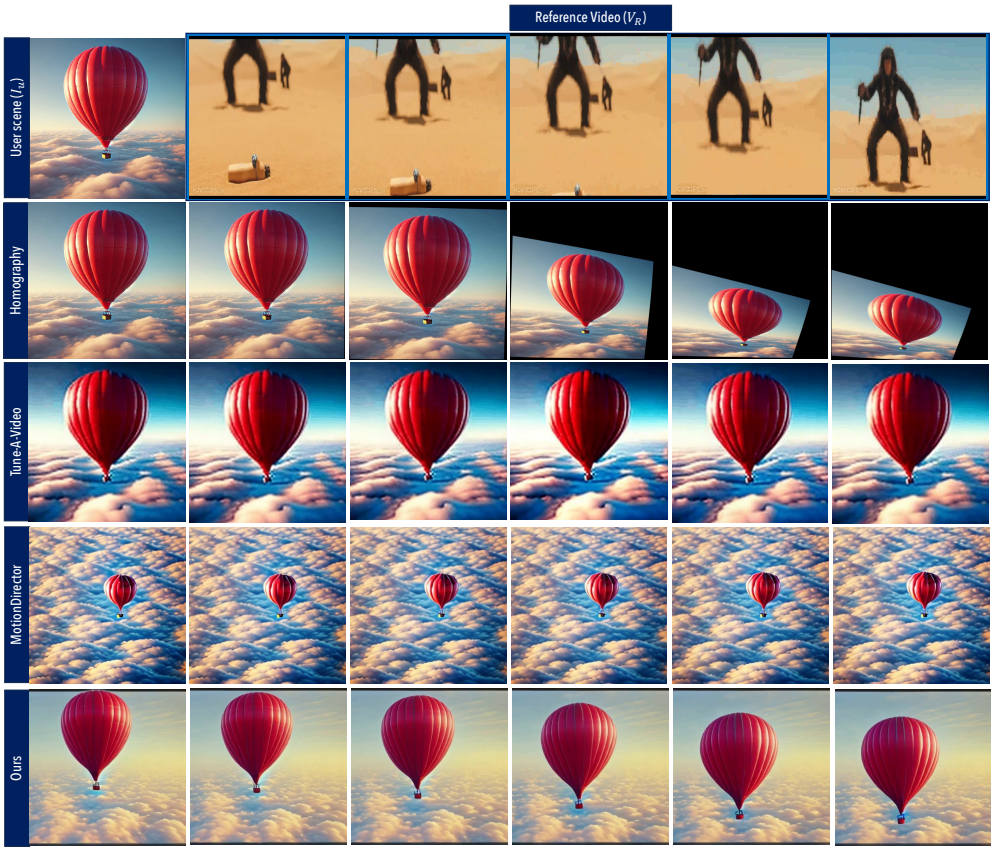


Fig. C4 The first row shows the inputs, i.e., the user-provided scene and the reference video (source) containing camera motion. In contrast to the baselines, the video output generated by our method more effectively captures the underlying camera motion from the reference video while also better preserving the details of the user-provided scene.

[13] Furht, B. (ed.). *Camera Motions*, 57–58 (Springer US, Boston, MA, 2008). URL https://doi.org/10.1007/978-0-387-78414-4_10.

[14] Camera movement terms: Everything you need to know (2024). URL https://www.nfi.edu/camera-movement-terms/.

[15] Videographer tips: How to direct your audience with camera movement (2024). URL https://hayotfilms.com/blog/videographer-tips-camera-movement/.

[16] How to use movement to tell a story in your videography (2024). URL https://rubberduckers.co.uk/how-to-use-movement-to-tell-a-story-in-your-videography/.

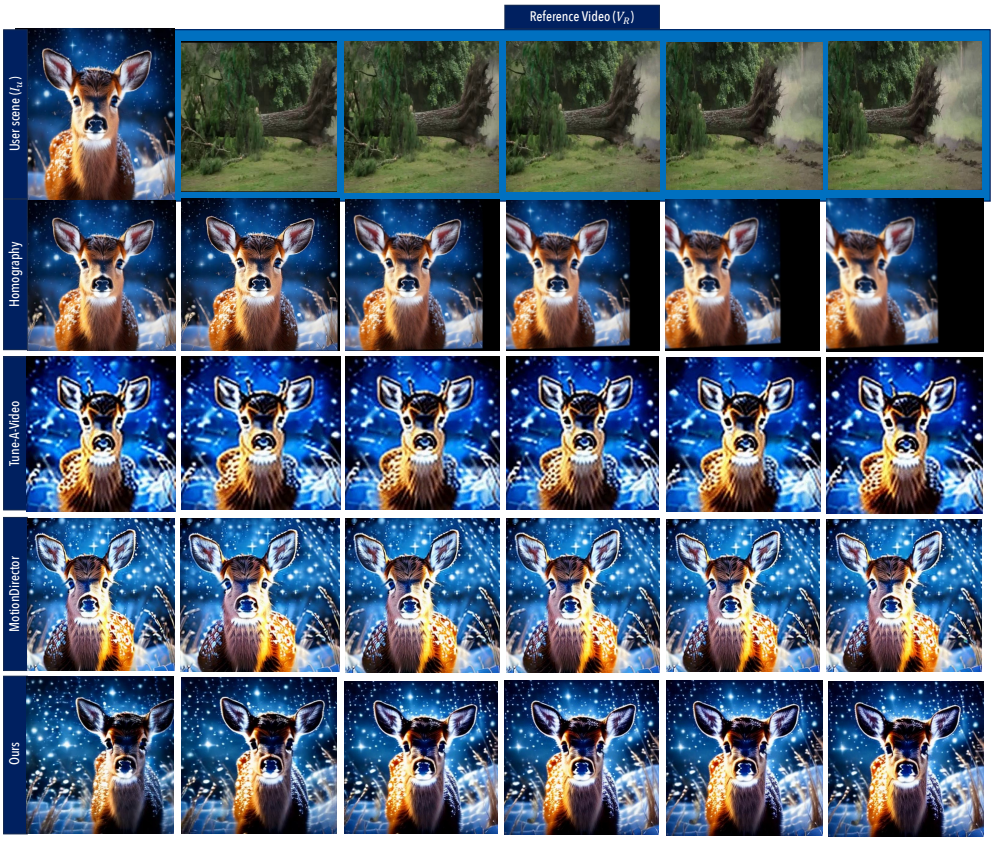


Fig. C5 The first row shows the inputs, i.e. the user-provided scene and the reference video (source) containing camera motion. In contrast to the baselines, the video output generated by our method more effectively captures the underlying camera motion from the reference video while also better preserving the details of the user-provided scene.

[17] Assembly, G. *et al.* Sustainable development goals. *SDGs transform our world* **2030** (2015).

[18] The challenges of shooting photography and videography (2025). URL https://www.adorama.com/alc/shooting-photography-and-videography/amp/.

[19] Movie gen: A cast of media foundation models (2024). URL https://ai.meta.com/research/publications/movie-gen-a-cast-of-media-foundation-models/.

[20] Veo-Team *et al.* Veo 2. *DeepMind Technical Report* (2024). URL https://deepmind.google/technologies/veo/veo-2/.

[21] Hu, T. *et al.* Committee, A. M. . O. (ed.) *Comd: Training-free video motion transfer with camera-object motion disentanglement.* (ed.Committee, A. M. . O.)

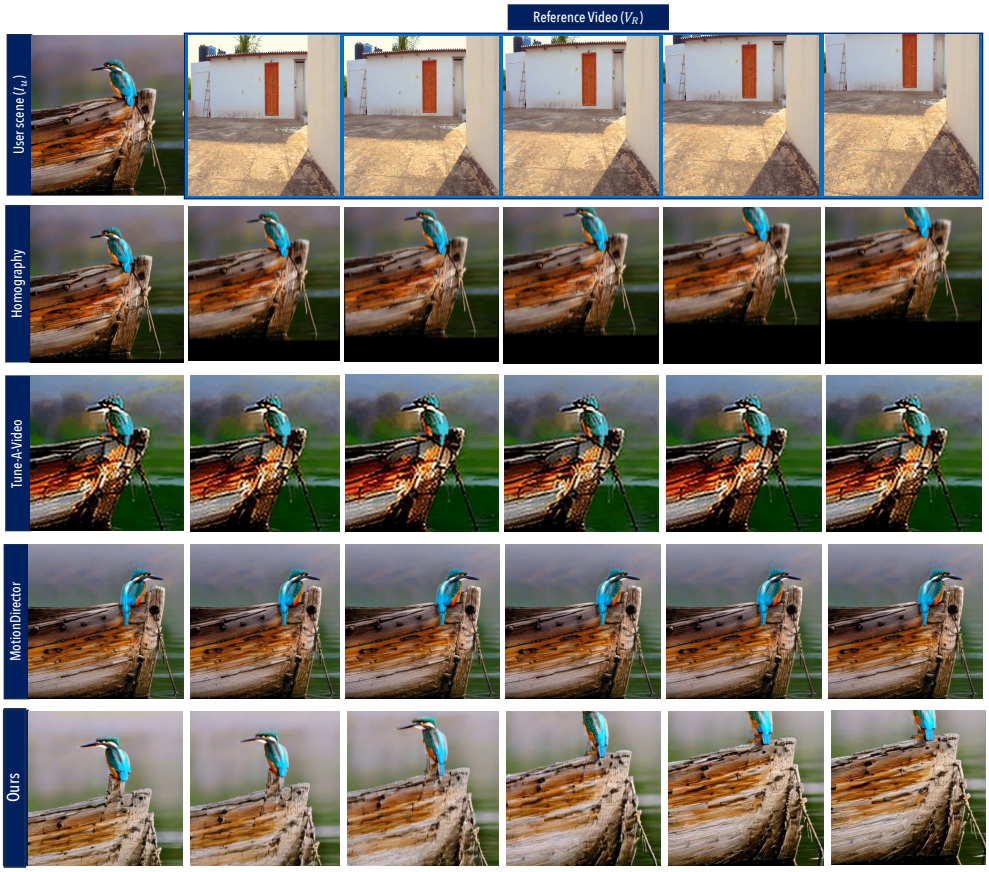


Fig. C6 The first row shows the inputs, i.e. the user-provided scene and the reference video (source) containing camera motion. In contrast to the baselines, the video output generated by our method more effectively captures the underlying camera motion from the reference video while also better preserving the details of the user-provided scene.

Proceedings of the 32nd ACM International Conference on Multimedia, 3459–3468 (Association for Computing Machinery, 2024).

[22] Yosef, E., Elmalem, S. & Giryes, R. Video reconstruction from a single motion blurred image using learned dynamic phase coding. *Scientific Reports* **13**, 13625 (2023).

[23] Imani, H., Islam, M. B., Junayed, M. S. & Ahad, M. A. R. Stereoscopic video deblurring transformer. *Scientific reports* **14**, 14342 (2024).

[24] Tilekbay, B., Yang, S., Lewkowicz, M. A., Suryapranata, A. & Kim, J. Committee, I. . O. (ed.) *Expressedit: Video editing with natural language and sketching.* (ed.Committee, I. . O.) *Proceedings of the 29th International Conference on*

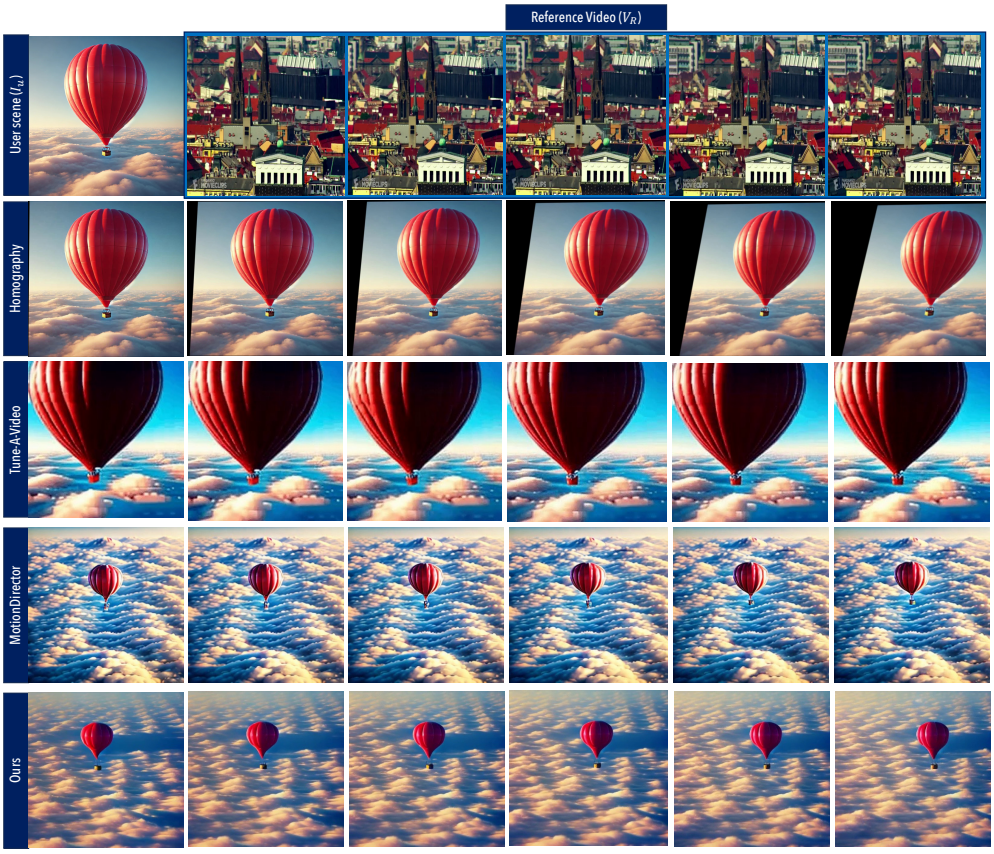


Fig. C7 The first row shows the inputs, i.e., the user-provided scene and the reference video (source) containing camera motion. In contrast to the baselines, the video output generated by our method more effectively captures the underlying camera motion from the reference video while also better preserving the details of the user-provided scene.

Intelligent User Interfaces, 515–536 (Association for Computing Machinery, 2024).

[25] Photoshop, A., Premiere, A., Pro, C. & CS, A. B. Adobe®. *Praha: Adobe, c2019 [cit. 2019-05-21]. Dostupné z: https://www. adobe. com/cz/products/photoshop. html* (2021).

[26] Abbaspour Tehrani, M., Beeler, T. & Grundhofer, A. Society, I. C. (ed.) *A practical method for fully automatic intrinsic camera calibration using directionally encoded light.* (ed.Society, I. C.) *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1106–1114 (IEEE, 2017).

[27] Wang, J. *et al.* Society, I. C. (ed.) *Vggt: Visual geometry grounded transformer.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision*

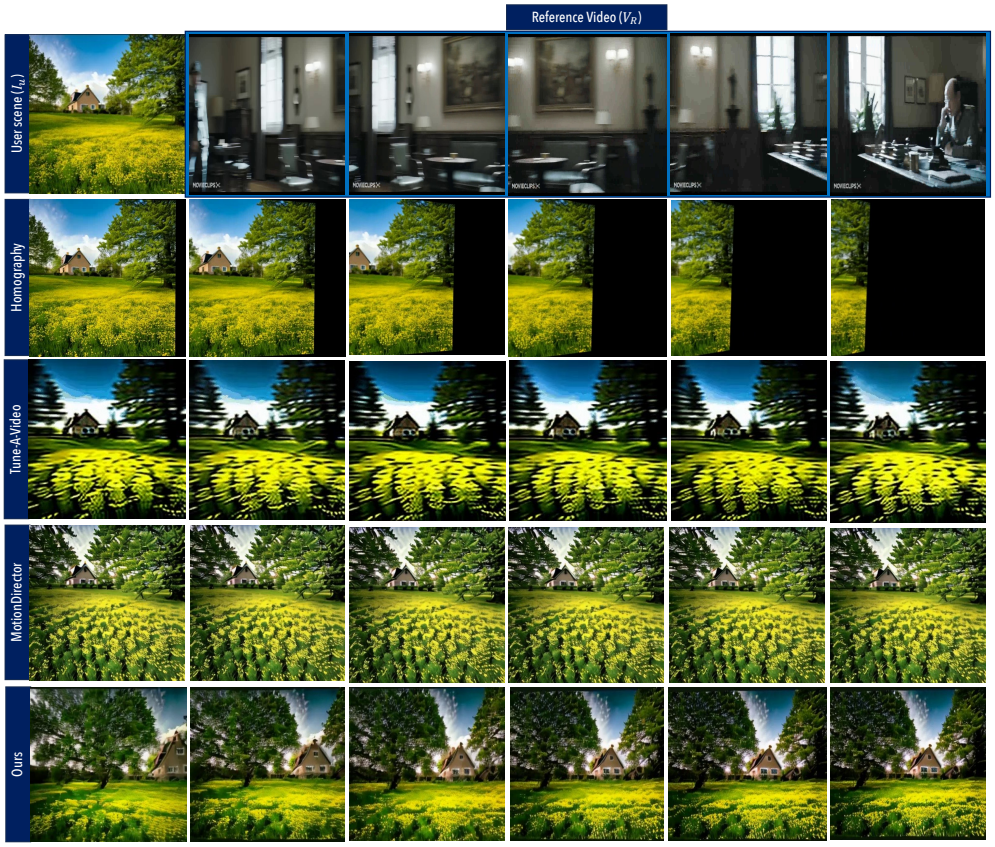


Fig. C8 The first row shows the inputs, i.e., the user-provided scene and the reference video (source) containing camera motion. In contrast to the baselines, the video output generated by our method more effectively captures the underlying camera motion from the reference video while also better preserving the details of the user-provided scene.

and Pattern Recognition, 5294–5306 (IEEE, 2025).

[28] Blattmann, A. *et al.* Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[29] Wang, J. *et al.* Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).

[30] Polyak, A. *et al.* Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720* (2024).

[31] Jeong, H., Park, G. Y. & Ye, J. C. Society, I. C. (ed.) *Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, 9212–9221 (IEEE, 2024).

[32] Ren, Y. *et al.* Customize-a-video: One-shot motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2402.14780* (2024).

[33] Zhang, Y. *et al.* Motioncrafter: One-shot motion customization of diffusion models. *arXiv preprint arXiv:2312.05288* (2023).

[34] Materzynska, J. *et al.* Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966* (2023).

[35] Wang, Z. *et al.* Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962* (2024).

[36] Kothandaraman, D. *et al.* Text prompting for multi-concept video customization by autoregressive generation. *arXiv preprint arXiv:2405.13951* (2024).

[37] Wu, J. Z. *et al.* Society, I. C. (ed.) *Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633 (IEEE, 2023).

[38] Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[39] Hu, T. *et al.* Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789* (2024).

[40] He, H. *et al.* Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).

[41] Zhang, D. J. *et al.* Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003* (2024).

[42] YU, M., Hu, W., Xing, J. & Shan, Y. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638* (2025).

[43] Jiang, H., Wang, X., Christie, M., Liu, L. & Chen, B. Cinematographic camera diffusion model. *Computer Graphics Forum* **43**, e15055 (2024).

[44] Luo, Y. *et al.* Chair, S. A. . P. (ed.) *Camclonemaster: Enabling reference-based camera control for video generation.* (ed.Chair, S. A. . P.) *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 1–10 (Association for Computing Machinery, 2025).

[45] Bai, J. *et al.* Society, I. C. (ed.) *Recammaster: Camera-controlled generative rendering from a single video.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV), 14834–14844 (IEEE, 2025).

[46] Wang, X., Courant, R., Christie, M. & Kalogeiton, V. Society, I. C. (ed.) *Akira: Augmentation kit on rays for optical video generation.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2609–2619 (IEEE, 2025).

[47] Chen, X. *et al.* Society, I. C. (ed.) *Anydoor: Zero-shot object-level image customization.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602 (IEEE, 2024).

[48] Yuan, Z. *et al.* Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784* (2023).

[49] Ma, L. *et al.* Pose guided person image generation. *Advances in neural information processing systems* **30** (2017).

[50] Nguyen, V. N. *et al.* Society, I. C. (ed.) *Nope: Novel object pose estimation from a single image.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17923–17932 (IEEE, 2024).

[51] Xu, Y., Xu, X., Gao, H. & Xiao, F. Sgdm: An adaptive style-guided diffusion model for personalized text to image generation. *IEEE Transactions on Multimedia* (2024).

[52] Ruiz, N. *et al.* Society, I. C. (ed.) *Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6527–6536 (IEEE, 2024).

[53] Tewel, Y., Gal, R., Chechik, G. & Atzmon, Y. Spencer, S. N. (ed.) *Key-locked rank one editing for text-to-image personalization.* (ed.Spencer, S. N.) *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11 (Association for Computing Machinery, 2023).

[54] Alaluf, Y., Richardson, E., Metzer, G. & Cohen-Or, D. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)* **42**, 1–10 (2023).

[55] Zhang, L., Rao, A. & Agrawala, M. Society, I. C. (ed.) *Adding conditional control to text-to-image diffusion models.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847 (IEEE, 2023).

[56] Shi, R. *et al.* Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110* (2023).

[57] Liu, M. *et al.* One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems* **36** (2024).

[58] Zeroscope (2023). URL https://huggingface.co/cerspense/zeroscope_v2_576w.

[59] Zhuang, Z. *et al.* Copra: A progressive lora training strategy. *arXiv preprint arXiv:2410.22911* (2024).

[60] Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021).

[61] Asadi, N. *et al.* Does combining parameter-efficient modules improve few-shot transfer accuracy? *arXiv preprint arXiv:2402.15414* (2024).

[62] Szeliski, R. *Computer vision: algorithms and applications* (Springer Nature, 2022).

[63] Kothandaraman, D., Zhou, T., Lin, M. & Manocha, D. Hawki: Homography & mutual information guidance for 3d-free single image to aerial view. *arXiv preprint arXiv:2311.15478* (2023).

[64] Lowe, G. Sift-the scale invariant feature transform. *Int. J* **2**, 2 (2004).

[65] Derpanis, K. G. Overview of the ransac algorithm. *Image Rochester NY* **4**, 2–3 (2010).

[66] Ho, J. & Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[67] Beauchemin, S. S. & Barron, J. L. The computation of optical flow. *ACM computing surveys (CSUR)* **27**, 433–466 (1995).

[68] Schönberger, J. L. & Frahm, J.-M. Society, I. C. (ed.) *Structure-from-motion revisited.* (ed.Society, I. C.) *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113 (IEEE, 2016).

[69] Ruiz, N. *et al.* Society, I. C. (ed.) *Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510 (IEEE, 2023).

[70] Zhao, R. *et al.* Committee, E. . O. (ed.) *Motiondirector: Motion customization of text-to-video diffusion models.* (ed.Committee, E. . O.) *European Conference on Computer Vision*, 273–290 (Springer, 2025).

[71] Imambi, S., Prakash, K. B. & Kanagachidambaresan, G. Pytorch. *Programming with TensorFlow: Solution for Edge Computing Applications* 87–104 (2021).

[72] Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[73] Xu, H. *et al.* Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084* (2021).

[74] Google DeepMind. Veo 3: State-of-the-art video generation model (2025). URL https://deepmind.google/models/veo/. Accessed: 2025-09-12.

[75] Google. Flow: Ai filmmaking tool designed for veo (2025). URL https://blog.google/technology/ai/google-flow-veo-ai-filmmaking-tool/. Accessed: 2025-09-12.

[76] Hart, S. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload/Elsevier* (1988).

[77] Brooke, J. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* (1996).

[78] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. Society, I. C. (ed.) *High-resolution image synthesis with latent diffusion models.* (ed.Society, I. C.) *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695 (IEEE, 2022).