Redirect4D-Bench: A Scalable Benchmark for Camera Redirection of Monocular Dynamic Videos with Pseudo-4D Ground Truth

Wei Cao[1], Hao Zhang[1], Jiapeng Tang[2], Yulun Wu[1],
Yingying Li[1], Ning Yu[3], Shenlong Wang[1], Yaoyao Liu[1]

[1] University of Illinois Urbana-Champaign
[2] Technical University of Munich [3] Netflix

Abstract

Camera redirection aims to replay a monocular source video along a requested camera path. Recent methods can generate plausible redirected videos, but current evaluations miss clear failures: an output may ignore the requested path, distort the foreground subject, or place it incorrectly in the new view. Reliable evaluation needs target views along the requested path, yet existing sources are limited: synthetic data has a domain gap, multi-camera capture is scarce and covers only recorded views, and monocular web videos lack observations from arbitrary camera paths. As a result, evaluations on real videos often rely on realism or semantic metrics, which do not directly measure camera following or subject placement. To tackle this, we introduce *Redirect4D-Bench*, a benchmark for camera redirection of real monocular dynamic videos, whose training-free construction pipeline scales the benchmark to more source clips, producing pseudo-4D ground truth for each requested camera path. The pipeline reconstructs a 4D scene for each curated clip, so multiple camera paths can be defined over the same dynamic event. It then renders depth and subject masks along each path as the pseudo-4D ground truth. Based on this pseudo-4D ground truth, we design new metrics for camera accuracy and subject fidelity. Benchmark results show that CLIP, FID/FVD, and VBench can miss visible trajectory and subject failures, while Redirect4D-Bench metrics reveal complementary errors in camera following and subject preservation.

1 Introduction

Camera redirection asks a model to replay a monocular source video along a requested camera path. As video generation moves beyond fixed viewpoints, camera redirection has become a natural interface for controlling how dynamic scenes are viewed [14, 54, 59]. Success requires more than visual realism: the output must follow the requested camera motion while preserving the dynamic foreground subject in the target view. Recent systems make this task increasingly plausible through camera-conditioned generation [17], video re-rendering [4, 70], and geometry-guided synthesis [21, 48, 68].

However, existing evaluations often rely on generic video quality or semantic similarity, which may fail to capture whether the output follows the requested camera path and preserves the foreground subject. Consequently, a redirected video that scores well on these evaluations may still fail the actual task: it can appear realistic while ignoring the target trajectory, damaging the foreground subject, or placing the subject in the wrong target-view location.

Reliable evaluation of camera redirection on real videos is difficult because references for the requested path are usually unavailable. Without such references, it is hard to judge whether the generated video follows the requested camera path and preserves the foreground subject in the new

Preprint.

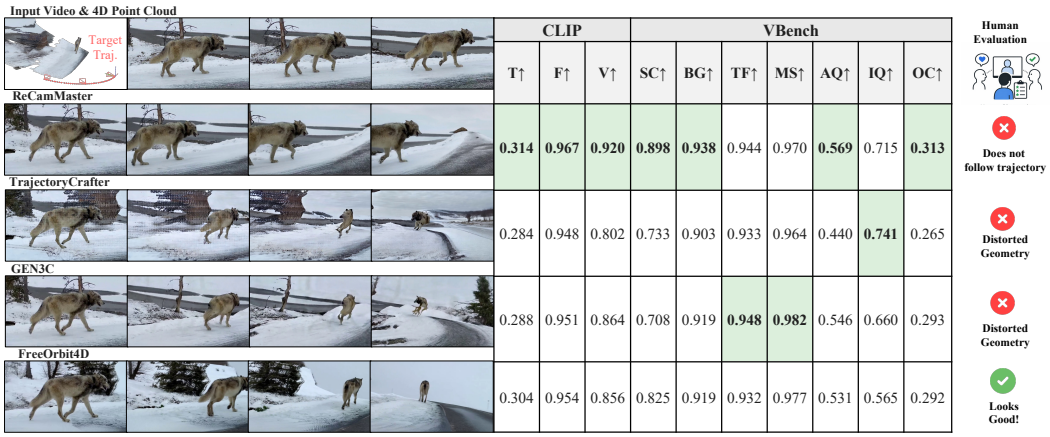


Figure 1: **High video-metric scores do not imply successful camera redirection.** CLIP-T/F/V [45] denote text alignment, adjacent-frame consistency, and source-video consistency; VBench [25] SC/BG/TF/MS/AQ/IQ/OC denote subject consistency, background consistency, temporal flickering, motion smoothness, aesthetic quality, imaging quality, and overall consistency. The shaded cells mark the best metric scores, but the corresponding outputs either do not follow the requested camera path or deform the subject. The human evaluation instead selects the output that follows the target trajectory and preserves the foreground subject.

view, as illustrated in Fig. 1. Existing metrics capture only part of this evaluation goal. Full-reference fidelity and perceptual metrics such as PSNR [26], SSIM [61], and LPIPS [72] require physically captured target-view RGB frames along the same trajectory. Distributional, semantic, and quality metrics such as FID [19], FVD [53], CLIP [45], and VBench [25] measure realism, similarity, or temporal quality, but they are not tied to the requested camera path. Camera-pose errors are closer to the task [4, 17], but they do not verify whether the foreground subject remains recognizable or correctly placed. This missing-reference problem is fundamental for real dynamic videos. No existing data source jointly provides in-the-wild realism and per-path target views: studio multi-camera capture covers only staged subjects, synthetic and hybrid corpora differ from real videos [4, 48, 68], and monocular web videos observe only the source camera path. Thus, a useful in-the-wild benchmark must construct a per-case reference for each requested trajectory.

To address this, we introduce *Redirect4D-Bench*, a benchmark for camera redirection of real monocular dynamic videos, whose training-free construction pipeline scales the benchmark to more source clips, producing pseudo-4D ground truth for each requested camera path. The pipeline reconstructs a 4D scene for each curated clip, so multiple camera paths can be defined over the same dynamic event. For each redirection case, Redirect4D-Bench combines a curated source clip, the reconstructed 4D scene, a parameterized camera path, and rendered depth and subject masks as the pseudo-4D ground truth. Together, these assets specify the camera path and subject placement that each generated video should satisfy. The retained benchmark covers human, robot, animal, and other dynamic subjects, with most cases requiring large-angle camera motion.

Based on this pseudo-4D ground truth, we design two new metrics. *Camera accuracy* measures whether the camera trajectory estimated from the generated video follows the benchmark target trajectory. *Subject fidelity* measures whether the foreground subject is detected, recognizable, and spatially aligned with the target-view subject-mask pseudo ground truth. Across four representative camera-redirection systems, these metrics reveal that camera following and subject preservation are partially decoupled, while CLIP, FID/FVD, and VBench scores can rank visually apparent task failures inconsistently. Our contributions are:

- We introduce Redirect4D-Bench, a scalable benchmark for camera redirection of real monocular dynamic videos. Its training-free pseudo-4D ground truth construction pipeline reconstructs a 4D scene for each curated clip and defines multiple camera paths over a common dynamic event.

- We propose task-specific camera-accuracy and subject-fidelity metrics, measuring whether a generated video follows the requested camera path and whether the foreground subject remains recognizable and correctly localized.
- We benchmark representative camera-redirection methods and show that common video metrics can miss visible task failures, while Redirect4D-Bench metrics expose complementary errors in camera following and subject preservation.

2 Related Work

Camera Redirection and Camera-Aware Generation. Camera redirection aims to resynthesize a source video along a user-specified camera trajectory, enabling post-capture viewpoint control and free-viewpoint replay of dynamic scenes [14, 68, 70]. Recent camera-aware generation methods use explicit camera conditioning and camera-control analysis [2, 3, 62], ray or relative camera encodings [17, 34], multi-video and time-camera control [32, 59], and training-free control [20]. They operate on images or videos through generative novel-view synthesis [54, 67, 69], video re-rendering [4, 24, 70], spatiotemporal caches [48], geometry guidance [21, 68], and 4D point-cloud proxies [9, 36]. Related control and evaluation problems also appear in pseudo-simulation [8], scene understanding [75], policy-safety evaluation [74], and active view selection [63]. Unlike CameraBench [37], which evaluates camera-motion recognition in existing videos, Redirect4D-Bench targets generated redirected videos and tests whether they follow the requested camera trajectory while preserving the foreground subject under that trajectory.

Video Data and Benchmark References. Web videos have long supported large-scale video benchmarks, from action and video understanding [1, 18, 30] to video-language learning and video generation data [6, 12, 29, 42, 44, 64]. However, these datasets provide semantic labels, captions, or unconstrained videos rather than trajectory-specific geometric references. Dynamic camera-pose and spatially annotated web-video datasets such as DynPose100K [49] and SpatialVID [56] scale pose, depth, and motion annotations for spatial learning, but they do not assign target trajectories to source clips or provide target-view subject references for redirection. Camera-control and view-synthesis systems therefore rely on posed static-scene videos [38, 77], driving data [50], synthetic multi-view data [16], hybrid web/static-view corpora [4, 48, 54, 68], or small dynamic benchmarks limited to recorded views [15]. Renderable 3D/4D representations can be built with NeRFs [43, 51], Gaussian or feed-forward splatting [11, 13, 27, 31], dynamic radiance fields [14], motion-aware Gaussians [33, 58], point maps [71], depth/pose estimation [28, 35], and non-rigid surface reconstruction [7, 9, 52]. Redirect4D-Bench curates monocular web clips and reconstructs a 4D scene for each clip to derive target-trajectory depth and subject-mask pseudo ground truth.

Evaluation Metrics. Existing evaluations measure useful but incomplete aspects of camera redirection. Full-reference fidelity and perceptual metrics such as PSNR [26], SSIM [61], and LPIPS [72] require captured target-view RGB frames along the same trajectory. Distributional metrics such as FID [19] and FVD [53], semantic metrics such as CLIP [45], and video-generation benchmarks such as EvalCrafter [41] and VBench [25] measure realism, similarity, and temporal quality, but they are not tied to the requested camera path or target-view subject placement. Motion-oriented benchmarks [39] and camera-pose errors used in camera-control evaluation [4, 17, 60] are closer to the task, but they do not jointly evaluate camera following, subject recognition, and target-view localization for each source clip. Redirect4D-Bench therefore evaluates camera accuracy for trajectory following together with subject fidelity for detection, recognition, and target-view localization.

3 Pseudo-4D Ground Truth Construction

For each source clip and target camera path, evaluating camera redirection requires a reference that specifies how the dynamic scene should be observed. No existing data source provides such references for in-the-wild dynamic scenes. Redirect4D-Bench addresses this by constructing *pseudo-4D ground truth*: for each curated source clip, we reconstruct a temporally aligned, foreground-complete 4D point-cloud representation that fills in subject geometry occluded in the source view, express target camera trajectories in its coordinate space, and render pseudo ground-truth target masks along each trajectory. Figure 2 summarizes the pipeline; the next three subsections describe source-video curation, 4D reconstruction with trajectory rendering, and the resulting dataset contents.

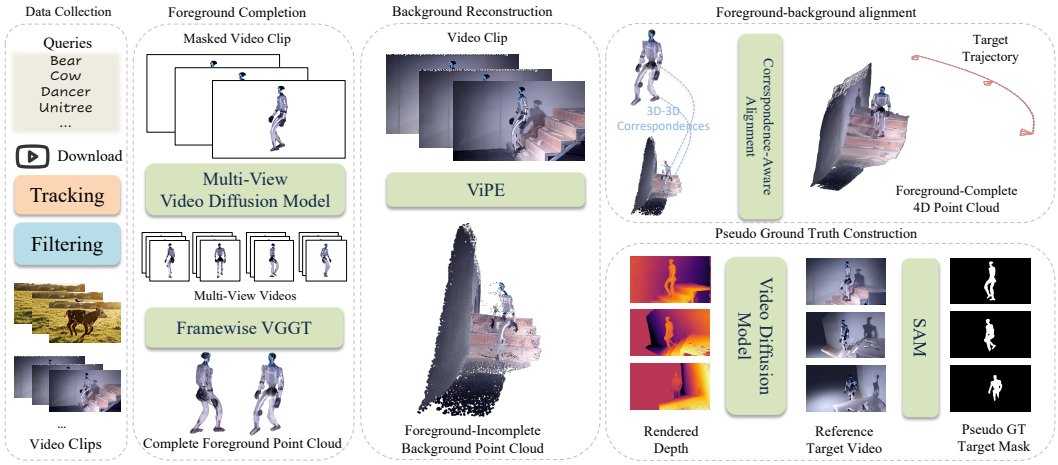


Figure 2: **Redirect4D-Bench construction pipeline.** Curated source clips and foreground masks feed two parallel branches: foreground completion from multi-view observations, and metric background reconstruction with ViPE [23]. We align the completed foreground to the metric scene through 3D–3D correspondences at foreground-mask pixels, yielding a shared 4D scene. This scene is rendered along target trajectories to produce the pseudo ground-truth target mask for evaluation.

3.1 Source Video Collection and Curation

Redirect4D-Bench starts from in-the-wild monocular clips with moving foreground subjects that admit reliable tracking and 4D reconstruction. Adapting the web-video pipeline of Animal-in-Motion [73], we download candidate YouTube videos from category-level queries, split them into shot-level clips, and filter out clips that are off-topic, multi-shot, or lack useful foreground motion. For each retained clip, we obtain object tracks and foreground masks with a Grounded-SAM pipeline [47] that combines Grounding DINO [40] with SAM-based video segmentation [46], discard tracks with severe overlap, truncation, very small subjects, unstable identity association, or infeasible crops, and standardize the remainder into object-centric sequences. After a final manual review, the retained tracks are passed to the 4D reconstruction stage described in Sec. 3.2.

3.2 4D Reconstruction and Pseudo Ground Truth Rendering

For every source clip, we build a renderable 4D point-cloud representation that combines a completed dynamic foreground with a metric static background. Following [9], we reconstruct foreground and background in two coordinate spaces: the foreground in canonical object space for subject completion, and the background in global scene space, in which target trajectories are specified.

Foreground completion. We use the curated masks (Sec. 3.1) to isolate the subject and generate multi-view observations with SV4D 2.0 [66]. Source and synthesized views are fed to VGGT-1B [57], whose point-map head predicts dense per-pixel 3D coordinates; masking to subject pixels gives a per-frame foreground point cloud expressed in canonical object space.

Metric-scale background reconstruction. For background reconstruction, we use ViPE [23], which we found to produce a more stable metric scene on spatially extended web-video clips than PAGE-4D [76]. ViPE recovers source-camera poses and intrinsics; combined with metric depth, these unproject background pixels into the global scene space in which we both specify target trajectories and evaluate the recovered camera motion (Sec. 4).

Foreground-background alignment. The completed foreground and the metric background initially live in different coordinate spaces. Using foreground-mask pixels in the source view, we establish 3D–3D correspondences between the completed foreground point map and the metric source-view point map [9]. From these correspondences, we estimate a per-frame scale-and-translation transform, apply it to the full completed foreground, and smooth the result over time. The resulting 4D scene

combines the metric background and per-frame completed foreground point clouds in the shared global scene space, ready for trajectory rendering.

Pseudo ground truth rendering. Each target trajectory is a parameterized camera arc whose center is obtained by unprojecting the first-frame center pixel at the reconstructed depth; the same depth gives the base orbit radius, with the source camera at the origin. The arc is specified by yaw, pitch, roll, and a radius scale; together with the source clip, the resulting 45 per-frame poses define a camera-redirection case. We rasterize the 4D point cloud along the target poses to obtain depth and foreground-mask scaffolds [9], which condition Wan2.2 VACE [55] together with a reference image and text prompt to synthesize the target-view RGB. The foreground-mask scaffold is then refined via SAM [10] video propagation on the synthesized RGB sequence, yielding the pseudo ground-truth target mask.

3.3 Dataset Contents and Statistics

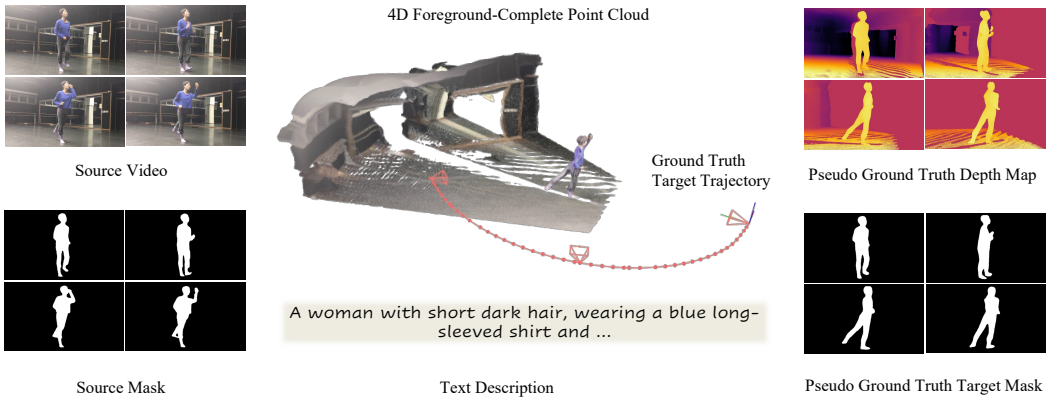


Figure 3: **Per-case data bundle.** Each retained redirection case contains a 45-frame source clip with foreground masks, a foreground-complete 4D point cloud, a parameterized target trajectory, a text prompt, a rendered target-view depth map, and the pseudo ground-truth target mask used as the per-case subject-fidelity target.

Figure 3 shows the per-case bundle. Redirect4D-Bench comprises 83 cases from 62 source clips at 15 fps and 832×480 resolution, spanning human, robot, and animal subjects. Most cases ($78/83$) require camera arcs spanning at least $90°$, making the benchmark a stress test of trajectory following rather than small viewpoint perturbation. Because the construction pipeline is training-free, additional cases can be added at bounded per-case cost (see Sec. A.1). Target-view RGB is not provided as ground truth: camera redirection reveals previously occluded background regions that admit many plausible completions. Evaluation instead relies on two well-defined targets: the requested camera trajectory (for camera accuracy) and the pseudo ground-truth target mask (for subject fidelity). Detailed dataset statistics, safety filtering, and the benchmark release policy appear in Sec. A.1.

4 Benchmark Metrics

Standard video metrics do not verify whether a generated video follows the requested camera path or preserves the foreground subject in the target view. Redirect4D-Bench therefore introduces two task-specific metric families: camera accuracy and subject fidelity.

4.1 Task Definition

A redirection case consists of a source clip $\mathcal{V}^{src} = \{\mathbf{I}_t^{src}\}_{t=1}^{T}$ and a target camera trajectory $\{\boldsymbol{\pi}_t^{tgt}\}_{t=1}^{T}$. A method outputs a redirected video $\widehat{\mathcal{V}}^{tgt}$ of length T, which Redirect4D-Bench evaluates along two axes (Fig. 4): camera accuracy (Sec. 4.3), asking whether the recovered camera path follows the target trajectory; and subject fidelity (Sec. 4.4), asking whether the foreground subject remains recognizable and correctly localized in the target view.

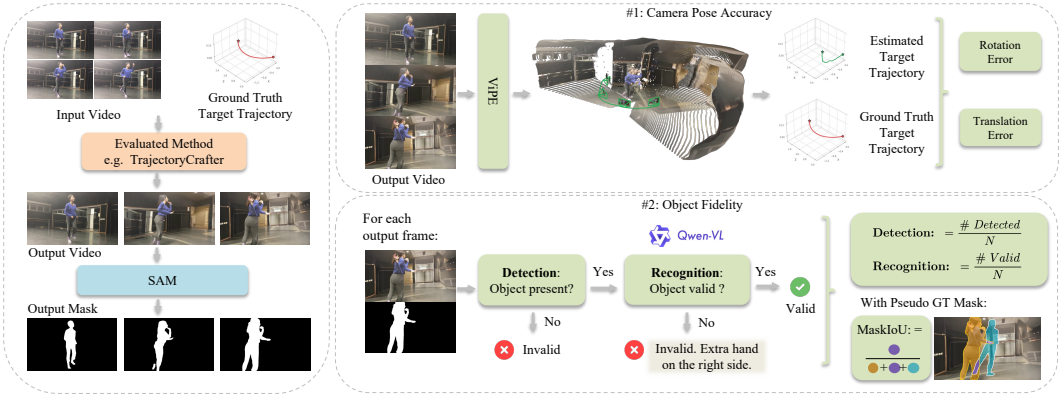


Figure 4: **Redirect4D-Bench metric protocol. (Left)** The evaluated method receives the source video and target camera trajectory and produces an output video, from which SAM [10] extracts a per-frame subject mask. **(#1, top right) Camera accuracy:** ViPE [23] recovers an estimated camera trajectory from the output video and visualizes it alongside the target in the reconstructed scene; we report rotation and translation errors. **(#2, bottom right) Subject fidelity:** for each output frame, Detection checks whether the subject is present and Qwen3-VL [5] Recognition judges whether the rendered subject is structurally plausible (e.g., catching defects like extra limbs); on recognized frames, MaskIoU is computed against the per-case pseudo ground-truth target mask.

4.2 Limitations of Existing Metrics

Full-reference fidelity metrics. PSNR [26], SSIM [61], and LPIPS [72] compare each generated frame with a captured target-view frame along the same camera path. For in-the-wild monocular videos, this reference does not exist: synchronized multi-view rigs cover only controlled studio settings, and synthetic data does not match the appearance of real scenes. Pixel-level fidelity, therefore, cannot rank methods on whether they correctly redirected real dynamic scenes.

Distributional and quality metrics. FID [19], FVD [53], CLIP similarities [45], and VBench quality dimensions [25] measure realism, semantic similarity, and temporal quality at the dataset level. None of them are conditioned on the requested camera path: a generated video can score well by remaining visually plausible while ignoring the path or placing the foreground subject incorrectly, as illustrated in Fig. 1. They are therefore useful as diagnostics for general video quality but cannot rank methods on whether they actually executed the requested redirection.

Camera-pose metrics. Rotation and translation errors directly assess whether a generated video follows the target camera trajectory, and we adopt this style of measurement in Sec. 4.3. On their own, they remain incomplete: a method can match the target trajectory while the foreground subject disappears, distorts, or appears in the wrong target-view location. Pose accuracy alone is therefore not a faithful proxy for successful camera redirection, which is why Redirect4D-Bench complements it with subject fidelity (Sec. 4.4).

4.3 Camera Accuracy

Camera accuracy asks whether the generated video $\widehat{\mathcal{V}}^{tgt}$ follows the target trajectory $\{\boldsymbol{\pi}_t^{tgt}\}_{t=1}^{T}$ (Fig. 4, camera-pose branch). Baselines do not necessarily expose their internal camera poses, so we evaluate every method through a uniform pose-recovery protocol that estimates a recovered trajectory $\{\widehat{\boldsymbol{\pi}}_t\}_{t=1}^{T}$ from the generated video itself; this makes the metric independent of any baseline-specific camera representation or output format.

Pose recovery and alignment. We run ViPE [23] with its `lyra` configuration on each generated video to recover per-frame poses $\widehat{\boldsymbol{\pi}}_t = (\widehat{R}_t, \widehat{\mathbf{t}}_t)$. To remove the arbitrary global offset inherent to monocular pose recovery, both target and recovered trajectories are expressed relative to their first frame, so the metric reports per-frame deviation from the requested arc rather than absolute world

position. Because monocular pose recovery returns translations only up to a per-clip scale, $\mathrm{TransErr}$ is computed in the same metric units as the target trajectory after aligning the two first frames, and it should be read alongside RotErr, which is scale-invariant. Fig. 4 (top right) shows the scene that ViPE reconstructs from the generated video, alongside the estimated and target trajectories plotted in separate 3D coordinate frames.

Rotation and translation error. Following camera-control works [4, 17], per-frame errors against the aligned target $\boldsymbol{\pi}_t^{tgt} = (R_t^{tgt}, \mathbf{t}_t^{tgt})$ are

$$\mathrm{RotErr}_t = \arccos\!\left(\frac{\mathrm{tr}(R_t^{tgt}\widehat{R}_t^{\top})-1}{2}\right), \qquad \mathrm{TransErr}_t = \|\mathbf{t}_t^{tgt} - \widehat{\mathbf{t}}_t\|_2.$$

Benchmark-level RotErr and $\mathrm{TransErr}$ are computed by averaging per-frame errors within each case, then over the 83 retained cases.

4.4 Subject Fidelity

Following the target camera path does not guarantee successful redirection: the generated video must also render a recognizable foreground subject and place it at the correct target-view location. Subject fidelity (Fig. 4, subject-fidelity branch) is therefore a per-frame cascade over target-view subject masks: detection and recognition measure whether the subject is rendered correctly, and MaskIoU measures whether it is correctly localized.

Reference and predicted masks. For each generated video, we apply SAM [10] to obtain a predicted subject mask at every frame and compare against the per-case pseudo ground-truth target mask from Sec. 3. Subject-fidelity metrics use the clip length T defined in Sec. 4.1 ($T = 45$) as their denominator; by design, every retained case has a non-empty pseudo-GT target mask in all T frames (Sec. 3.2). Frames where SAM returns an empty or invalid mask therefore count directly as failures.

Detection and recognition. Detection asks whether the subject is rendered where the reference expects it: $\mathrm{detected}(t)$ holds iff the predicted SAM mask and the reference target mask at frame t are both non-empty. Recognition refines this by asking Qwen3-VL-32B [5] whether the predicted subject is structurally plausible for the case's category; $\mathrm{recognized}(t)$ holds iff the frame is both detected and judged plausible. For instance, a human silhouette with an extra limb attached or a fragmented body fails recognition (Fig. 4, bottom right). The two rates are

$$D = \frac{|\{t : \mathrm{detected}(t)\}|}{T}, \qquad R = \frac{|\{t : \mathrm{recognized}(t)\}|}{T}, \qquad R \le D.$$

Benchmark-level D and R are computed per case, then averaged over the 83 retained cases.

Subject localization. On recognized frames, we compute MaskIoU between predicted and reference masks. The conditional score averages over recognized frames; the recognition-weighted score instead averages over all T frames, counting non-recognized frames as zero:

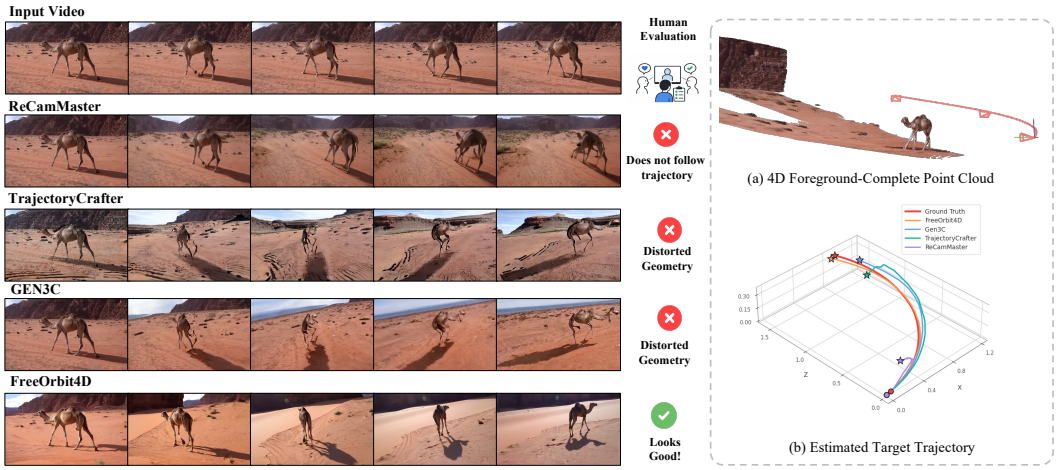
$$\mathrm{cMaskIoU} = \operatorname*{mean}_{t:\mathrm{recognized}(t)} \mathrm{MaskIoU}(t),$$

$$R \cdot \mathrm{cMaskIoU} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{MaskIoU}(t)\,\mathbf{1}[\mathrm{recognized}(t)].$$

$R \cdot \mathrm{cMaskIoU}$ is our primary localization metric: it rewards both rendering the correct subject and placing it at the correct target-view location.

5 Experiments

We apply the evaluation protocol of Sec. 4 to four camera-redirection baselines, presenting both a case-level (Fig. 5) and a benchmark-level (Table 1) comparison.



	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	**0.317**	**0.983**	**0.935**	**0.931**	0.911	**0.962**	**0.984**	0.507	0.709	0.219	27.313	0.727	**1.000**	0.756	0.095	0.071
TrajectoryCrafter	0.311	0.949	0.846	0.796	0.907	0.933	0.968	0.498	**0.748**	**0.223**	3.985	0.152	**1.000**	0.044	0.493	0.022
Gen3C	0.301	0.949	0.849	0.814	**0.939**	0.958	**0.984**	0.498	0.685	0.215	7.099	0.157	**1.000**	0.178	0.507	0.090
FreeOrbit4D	0.304	0.934	0.858	0.874	0.934	0.952	0.982	**0.554**	0.703	**0.223**	**0.793**	**0.027**	**1.000**	**1.000**	**0.919**	**0.919**

Figure 5: **Case-level comparison on a camel case. (Left)** the source clip and the four baselines' generated target-view frames, each with a human-evaluation mark. **(Right) (a)** the foreground-complete 4D point cloud with the target trajectory arc; **(b)** a 3D plot overlaying the four baselines' recovered trajectories on the target trajectory. The accompanying per-case table reports scores on the same sample. FID/FVD are omitted because they are aggregate distribution metrics; cIoU abbreviates cMaskIoU; **bold** marks the best score per column.

5.1 Baselines

We evaluate four camera-redirection systems on the 83 retained Redirect4D-Bench cases. All baselines receive the same source clip and target camera trajectory; we re-format the trajectory to match each method's expected input where needed, but the camera path itself is unchanged, so output differences reflect the methods rather than the inputs.

(i) ReCamMaster [4] is a video re-rendering model conditioned on the source video, a text prompt, and the target camera trajectory. We feed the benchmark trajectory directly into its native camera encoder, after resampling it to ReCamMaster's expected inference length.

(ii) TrajectoryCrafter [68] estimates per-frame depth from the source video, unprojects it into a dynamic point cloud, renders the cloud along the target trajectory, and refines the result with a CogVideoX-based diffusion model [65]. For a fair comparison, we replace its DepthCrafter [22] front-end with ViPE [23], matching Gen3C and our pseudo-GT pipeline.

(iii) Gen3C [48] builds a spatiotemporal 3D cache from the source video and renders it along the target camera path as the diffusion condition. We feed ViPE [23] depth, poses, and intrinsics into its dynamic-video path, with each trajectory mapped to Gen3C's native yaw/pitch/roll/scale parameters.

(iv) FreeOrbit4D [9] performs training-free camera redirection by reconstructing a 4D scene from the source clip and refining the renders with video diffusion; we run its published configuration on each Redirect4D-Bench case without modification.

5.2 Results

We present results in two steps: a single representative case (Fig. 5) and the aggregate over all 83 retained cases (Table 1).

Table 1: **Benchmark-level comparison.** Traditional metrics on the left, Redirect4D-Bench metrics on the right. CLIP [45]: CLIP-F is adjacent-frame consistency, CLIP-V is prediction-vs-source consistency. VBench [25] columns SC/BG/TF/MS/AQ/IQ/OC stand for subject, background, temporal-flickering, motion-smoothness, aesthetic, imaging, and overall consistency. **Bold** = best per column.

	Traditional metrics												Redirect4D-Bench metrics					
	CLIP ↑			Dist. ↓		VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	FID	FVD	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	0.324	**0.970**	**0.881**	**52.429**	**662.048**	**0.884**	0.900	**0.953**	**0.981**	**0.508**	0.605	0.274	20.81	0.596	0.965	0.907	0.246	0.227
TrajectoryCrafter	0.324	0.958	0.832	83.559	1082.468	0.836	0.895	0.932	0.971	0.472	**0.631**	0.271	5.26	0.379	**1.000**	0.530	0.495	0.253
Gen3C	0.321	0.954	0.849	77.513	1047.264	0.839	0.902	0.942	0.980	0.480	0.543	0.272	5.25	0.312	**1.000**	0.596	0.486	0.289
FreeOrbit4D	**0.326**	0.953	0.843	67.305	925.250	0.860	**0.915**	0.937	0.976	0.496	0.610	**0.282**	**5.10**	**0.280**	**1.000**	**0.990**	**0.907**	**0.898**

Case-level comparison. Figure 5 shows a representative large-angle camel case. Visually, ReCamMaster produces plausible frames but does not follow the requested camera path; TrajectoryCrafter and Gen3C follow the trajectory more closely but distort the camel's body; only FreeOrbit4D follows the path while keeping the camel intact. The traditional CLIP and VBench scores contradict this ranking, scattering "best" cells across all four methods including the visibly broken ones. The Redirect4D-Bench columns instead concentrate the best cells on FreeOrbit4D, because they separately measure the three things a human observer attends to: camera following, subject recognition, and target-view localization. A method that fails on any one of these is therefore visible in the corresponding column instead of being averaged out by other strengths.

Benchmark-level comparison. Table 1 aggregates the comparison over all 83 retained cases. The same pattern persists. Traditional metric winners remain scattered: ReCamMaster takes FID/FVD and several CLIP/VBench columns, FreeOrbit4D wins CLIP-T, Background Consistency, and Overall Consistency, and TrajectoryCrafter wins Imaging Quality. These metrics reward realism, similarity, and temporal smoothness, but none reveal whether the requested redirection actually happened.

The Redirect4D-Bench metrics, in contrast, give a clear two-axis diagnosis. *Camera accuracy.* FreeOrbit4D, Gen3C, and TrajectoryCrafter cluster near 5° mean rotation error, while ReCamMaster drifts by 20.81°, roughly four times worse, confirming that ReCamMaster does not follow the benchmark trajectory at scale. *Subject fidelity.* The four baselines split into two failure modes. TrajectoryCrafter and Gen3C detect the subject reliably ($D = 1.000$) but fail recognition: only $R = 0.530$ and $R = 0.596$ of their frames are judged structurally plausible by Qwen3-VL, consistent with the camel-case distortions. ReCamMaster shows the opposite, recognizing the subject in $R = 0.907$ of frames but placing it at the wrong target-view location (R·cMaskIoU $= 0.227$). Only FreeOrbit4D scores well on both axes (RotErr $= 5.10^\circ$, R·cMaskIoU $= 0.898$), agreeing with the human verdict on the camel case (Fig. 5).

Camera following and subject fidelity therefore need to be reported as separate axes: a method can follow the trajectory while breaking the subject, or preserve plausible appearance while missing the path. The two failure modes we observe (broken-subject and misplaced-subject) are distinct defect categories that traditional video metrics conflate into a single "visual quality" score. Ranking by traditional metrics alone therefore risks selecting visibly broken outputs as winners, while the Redirect4D-Bench axes restore the link between the metric and the human-visible failure mode.

6 Conclusion

Evaluating camera redirection on real monocular videos is held back by a missing-reference problem: no captured target view exists for an arbitrary camera path, so video-quality metrics measure realism rather than per-trajectory task success. Redirect4D-Bench addresses this with a scalable, training-free pipeline that produces per-trajectory pseudo ground truth for real dynamic scenes, and defines two task-specific metrics (camera accuracy and subject fidelity) that align with human visual judgment where CLIP, FID/FVD, and VBench disagree. Across our four baselines, the two axes expose distinct failure modes (broken-subject vs. misplaced-subject) that traditional metrics conflate, indicating that future methods should be reported on both axes. The pipeline currently assumes a single dominant, source-visible subject; lifting this constraint and scaling the construction code to more clips are natural extensions.

Acknowledgments

This research is supported by the National Artificial Intelligence Research Resource (NAIRR) Pilot under award NAIRR250199 and the NVIDIA Academic Grant Program. Computational resources are also provided by Delta and DeltaAI at the National Center for Supercomputing Applications (NCSA) through ACCESS allocations CIS250012, CIS250816, and CIS251188. S. W. is also supported by NSF Awards #2525287, #2404385, #2414227, #2340254, #2312102, and #2331878, and research grants from IBM, Meta, NVIDIA, and Intel.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. VD3D: Taming large video diffusion transformers for 3D camera control. *arXiv preprint arXiv:2407.12781*, 2024.

[3] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. AC3D: Analyzing and improving 3D camera control in video diffusion transformers. In *CVPR*, pages 22875–22889, 2025.

[4] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. ReCamMaster: Camera-controlled generative rendering from a single video. In *ICCV*, pages 14834–14844, 2025.

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[6] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021.

[7] Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2vecsets: 4D latent vector set diffusion for non-rigid shape reconstruction and tracking. In *CVPR*, pages 20496–20506, 2024.

[8] Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, et al. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218*, 2025.

[9] Wei Cao, Hao Zhang, Fengrui Tian, Yulun Wu, Yingying Li, Shenlong Wang, Ning Yu, and Yaoyao Liu. FreeOrbit4D: Training-free arbitrary camera redirection for monocular videos via foreground-complete 4D reconstruction. In *ACM SIGGRAPH Conference Papers*, 2026.

[10] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[11] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *CVPR*, pages 19457–19467, 2024.

[12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70M: Captioning 70m videos with mulieee transactions on image processingle cross-modality teachers. In *CVPR*, 2024.

[13] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In *ECCV*, pages 370–386, 2024.

[14] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, pages 5712–5721, 2021.

[15] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022.

[16] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *CVPR*, pages 3749–3761, 2022.

[17] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

[18] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, volume 30, 2017.

[20] Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

[21] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. EX-4D: Extreme viewpoint 4D video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

[22] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025.

[23] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

[24] Zhening Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y. Wang, Joan Lasenby, and Chun-Hao Huang. SpaceTimePilot: Generative rendering of dynamic scenes across space and time. *arXiv preprint arXiv:2512.25075*, 2025.

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024.

[26] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008.

[27] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. AnySplat: Feed-forward 3D gaussian splatting from unconstrained views. *ACM Transactions on Graphics*, 44(6):1–16, 2025.

[28] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In *ICCV*, pages 20658–20671, 2025.

[29] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. MiraData: A large-scale video dataset with long durations and structured captions. *NeurIPS*, 37:48955–48970, 2025.

[30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

[32] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J. Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *NeurIPS*, 37:16240–16271, 2024.

[33] Jiahui Lei, Yijia Weng, Adam W. Harley, Leonidas Guibas, and Kostas Daniilidis. MoSca: Dynamic gaussian fusion from casual videos via 4D motion scaffolds. In *CVPR*, pages 6165–6177, 2025.

[34] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025.

[35] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSAM: Accurate, fast and robust structure and motion from casual dynamic videos. In *CVPR*, pages 10486–10496, 2025.

[36] Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4D point clouds. In *CVPR*, 2026.

[37] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, et al. Towards understanding camera motions in any video. *arXiv preprint arXiv:2504.15376*, 2025.

[38] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *CVPR*, pages 22160–22169, 2024.

[39] Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. VMBench: A benchmark for perception-aligned video motion generation. In *ICCV*, pages 13087–13098, 2025.

[40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[41] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *CVPR*, pages 22139–22149, 2024.

[42] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[43] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

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

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[46] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[47] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[48] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3C: 3D-informed world-consistent video generation with precise camera control. In *CVPR*, pages 6121–6132, 2025.

[49] Chris Rockwell, Joseph Tung, Tsung-Yi Lin, Ming-Yu Liu, David F. Fouhey, and Chen-Hsuan Lin. Dynamic camera poses and where to find them. In *CVPR*, pages 12444–12455, 2025.

[50] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020.

[51] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023.

[52] Jiapeng Tang, Wei Cao, Biao Zhang, Chang Luo, Yaoyao Liu, and Matthias Nießner. Motion2VecSets: Non-rigid shape reconstruction and tracking with 4D latent set diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[53] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *ICLR 2019 Workshops*, 2019.

[54] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *ECCV*, pages 313–331. Springer, 2024.

[55] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[56] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. SpatialVID: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025.

[57] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025.

[58] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4D reconstruction from a single video. In *ICCV*, pages 9660–9672, 2025.

[59] Yiming Wang, Qihang Zhang, Shengqu Cai, Tong Wu, Jan Ackermann, Zhengfei Kuang, Yang Zheng, Frano Rajič, Siyu Tang, and Gordon Wetzstein. BulletTime: Decoupled control of time and camera pose for video generation. *arXiv preprint arXiv:2512.05076*, 2025.

[60] Zhaoqing Wang, Xiaobo Xia, Zhuolin Bie, Jinlin Liu, Dongdong Yu, Jia-Wang Bian, and Changhu Wang. Taming camera-controlled video generation with verifiable geometry reward. *arXiv preprint arXiv:2512.02870*, 2025.

[61] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

[62] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH Conference Papers*, pages 1–11, 2024.

[63] Yulun Wu, Ruyi Zha, Wei Cao, Yingying Li, Yuanhao Cai, and Yaoyao Liu. Active view selection with perturbed gaussian ensemble for tomographic reconstruction. *arXiv preprint arXiv:2603.06852*, 2026.

[64] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, pages 5036–5045, 2022.

[65] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[66] Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4D generation. In *ICCV*, pages 13248–13258, 2025.

[67] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. NVS-Solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024.

[68] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, pages 100–111, 2025.

[69] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[70] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. ReCapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *CVPR*, pages 2050–2062, 2025.

[71] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

[72] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[73] Brian Nlong Zhao, Jiajun Wu, and Shangzhe Wu. Web-scale collection of video data for 4D animal reconstruction. *arXiv preprint arXiv:2511.01169*, 2025.

[74] Hongkuan Zhou, Wei Cao, Aifen Sui, and Zhenshan Bing. What matters to enhance traffic rule compliance of imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:2309.07808*, 2023.

[75] Hongkuan Zhou, Stefan Schimid, Yicong Li, Lavdim Halilaj, Xiangtong Yao, and Wei Cao. Predicting the road ahead: A knowledge graph based foundation model for scene understanding in autonomous driving. In *European Semantic Web Conference*, pages 116–132. Springer, 2025.

[76] Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. PAGE-4D: Disentangled pose and geometry estimation for VGGT-4D perception. *arXiv preprint arXiv:2510.17568*, 2025.

[77] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using mulieee transactions on image processinglane images. *ACM Transactions on Graphics*, 37(4):65, 2018.

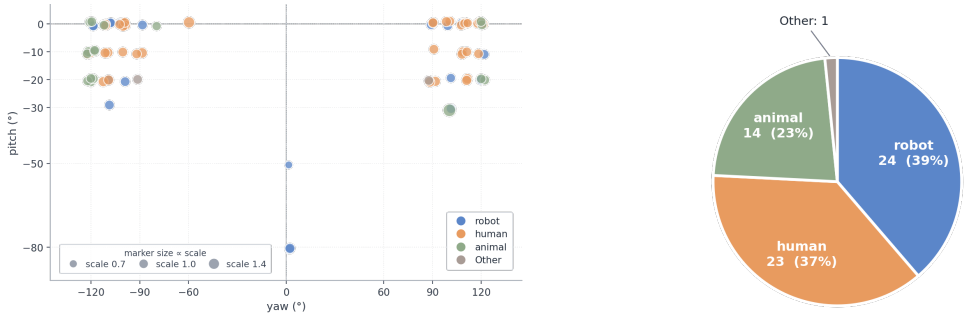
A Appendix

A.1 Dataset Statistics

The current Redirect4D-Bench release contains 62 curated source video clips and 83 retained redirection cases, each source clip standardized to 45 frames at 15 fps and 832×480 resolution. All quantitative results reported in the paper are computed on these 83 retained cases.

Safety filtering. Of the 103 target trajectories produced by the construction pipeline, we exclude 20 cases that involve minors as the foreground subject, leaving the 83 retained cases for evaluation.

Release policy. The benchmark is released on Hugging Face. The public dataset contains derived assets (4D point clouds, target trajectories, rendered geometry, subject masks, prompts, and metadata) plus a small public sample for inspection. Original source RGB is not redistributed; following [73], the released code provides a recovery script using YouTube identifiers in the metadata.



(a) Trajectory Parameters Distribution

(b) Subject Categories Distribution

Figure 6: **Dataset statistics. (a)** Yaw and pitch parameters for the retained target trajectories. Each point is one redirection case; color denotes the subject family (robot, human, animal, other) and marker size denotes the radius scale (0.7, 1.0, or 1.4). Yaw values cluster between $\pm 90°$ and $\pm 120°$, with most pitch values in $[-30°, 0°]$. **(b)** Source-clip counts across subject families in the current release: robot 24 (39%), human 23 (37%), animal 14 (23%), other 1 ($\sim 2\%$), totaling the 62 source clips that yield 83 redirection cases. The wide yaw spread makes the benchmark a stress test for trajectory following rather than small viewpoint perturbation.

A.2 Qwen3-VL Subject-Recognition Prompt

The subject-recognition component in Sec. 4.4 uses Qwen3-VL-32B [5] as a binary judge: given a cropped generated frame and a subject class `{cls}`, the judge decides whether the foreground subject is catastrophically broken (e.g., extra or missing limbs, fragmented body, fused parts). The prompt template is shown in Fig. 7; separate defect lists are used for biological/articulated and mechanical subjects, and the model's textual response is parsed into a binary recognition decision.

```
[Shared Instruction]
You are shown a single frame cropped from an AI-generated video.  The crop contains a {cls}.  The
camera angle may be unusual.

Question:  is the {cls} catastrophically broken, i.e., so badly deformed that it is clearly not a
real {cls}?

[Biological / Articulated Subjects]
Only answer "yes" if ONE of these is undeniably visible:
  a.  DUPLICATION: two heads, two tails, two torsos, or two complete instances of a body part
that should appear only once.
  b.  WRONG LIMB COUNT: a clearly visible extra limb, or a main limb entirely missing where
anatomy demands it.
  c.  MELTED / BLOBBED: the {cls}'s body or limbs have dissolved into a continuous amorphous blob
with no recognizable structure.
  d.  REVERSED JOINT: a knee, elbow, hock, wrist, or ankle bends in the opposite direction to the
species' natural anatomy.
If the {cls} looks like a plausible albeit imperfect rendering of a real {cls}, answer "no".
Imperfect details, blur, odd poses, long or S-curved necks, humps, trunks, stripes, fur / skin
texture, dance poses, unusual camera angles, cropped parts, self-occlusion, and natural
class-specific features are NEVER defects.

[Mechanical Subjects]
Only answer "yes" if ONE of these is undeniably visible:
  a.  DUPLICATION: two of a main component that should appear only once (e.g., two gear hubs, two
robot torsos / heads, or two complete structural pieces).
  b.  MELTED / BLOBBED: structural parts have dissolved into a continuous amorphous blob with no
recognizable mechanical form.
  c.  IMPOSSIBLE STRUCTURE: a rigid component fluidly bent like rubber, a hinge twisted beyond
its physical range, or a snapped structural piece floating apart.
  d.  MIS-ATTACHED PART: a main component sprouts from a clearly impossible location, such as a
robot limb attached at the head.
If the {cls} looks like a plausible albeit imperfect rendering of a real {cls}, answer "no".
Unusual camera angles, cropped parts, stylized aesthetics, visible internal mechanisms, unusual
colors, minor jitter, and class-specific mechanical features are NEVER defects.

[Output Rule]
When in doubt, always answer "no".  Flag "yes" only for defects that would stop someone scrolling
past on social media to say "wait, that's wrong."

Respond with exactly one line.  Start with "no" if the {cls} is plausible; otherwise start with

"yes" followed by a brief description of which part is catastrophically broken.
```

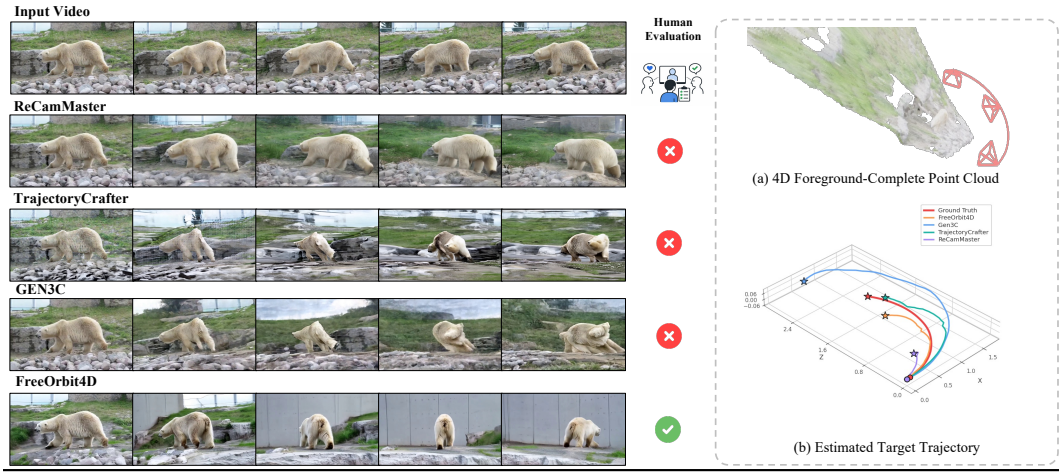
Figure 7: **Qwen3-VL recognition prompt.** The prompt receives a crop from a generated frame and a subject class placeholder `{cls}`. Separate defect lists are used for biological/articulated and mechanical subjects. The response is parsed as a binary recognition decision: catastrophic defects make the frame invalid, while plausible but imperfect renderings remain valid.

A.3 Compute Resources

All experiments run on a single internal node with $8\times$ NVIDIA A100-SXM4-80GB GPUs. We organize the budget by stage.

4D-reconstruction pipeline. The construction pipeline (Sec. 3) is a one-time cost per dataset version. End-to-end on the 83 retained cases it consumes roughly 40 A100-hours, dominated by foreground completion (SV4D 2.0 [66] pseudo multi-view synthesis and VGGT-1B [57] point-map prediction), background reconstruction (ViPE [23] `lyra` configuration), and target-view rendering with Wan2.2 VACE [55]. Per-stage wall times on 8 GPUs: 5 min for source-RGB caching, 5 min for SAM source-mask sharding, and 5 h for the reconstruction-render-refine loop.

Baseline inference. Each baseline is run once per case on a single A100-80GB. Per-case averages over the 83 retained cases are: ReCamMaster [4] $\sim$5.7 min/case ($\sim$8 A100-hours total), TrajectoryCrafter [68] $\sim$7 min/case ($\sim$10 A100-hours total), Gen3C [48] $\sim$17 min/case ($\sim$23 A100-hours total), and FreeOrbit4D [9] $\sim$10 min/case ($\sim$14 A100-hours total). Sharded over 8 GPUs the wall-clock times are approximately 1 h, 1.5 h, 2 h 50 min, and 1.7 h respectively.



	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	0.324	**0.980**	**0.945**	**0.931**	0.943	**0.948**	**0.982**	**0.549**	0.648	0.294	31.930	0.835	**1.000**	**1.000**	0.402	0.402
TrajectoryCrafter	0.324	0.955	0.864	0.773	0.925	0.896	0.956	0.429	**0.652**	**0.310**	4.123	0.268	**1.000**	0.444	0.636	0.283
Gen3C	**0.342**	0.945	0.853	0.779	0.932	0.911	0.964	0.471	0.571	0.305	23.953	0.665	**1.000**	0.800	0.622	0.497
FreeOrbit4D	0.298	0.963	0.816	0.852	**0.953**	0.932	0.973	0.516	0.628	0.279	**3.529**	**0.156**	**1.000**	**1.000**	**0.981**	**0.981**

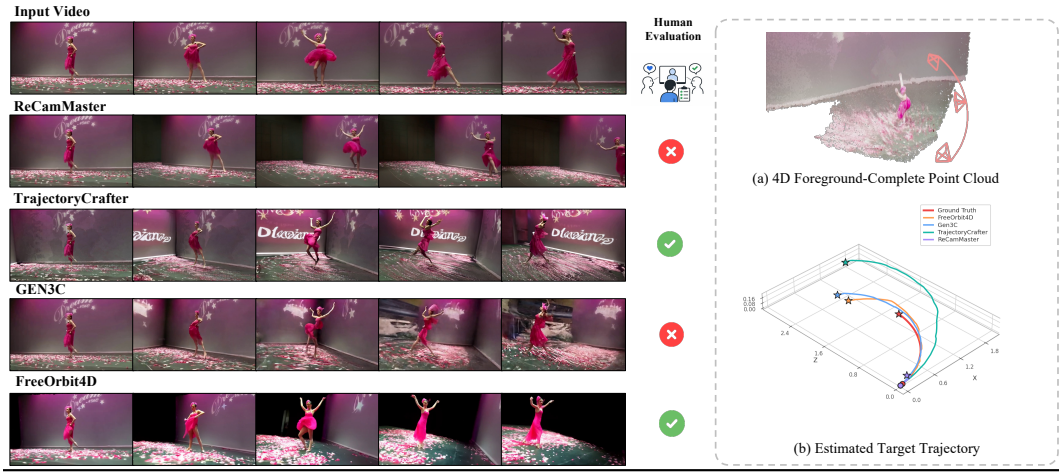
Figure 8: **Bear case.** ReCamMaster [4] earns the best score on six traditional CLIP/VBench columns by emitting frames close to the source view, but its recovered camera barely moves and drifts 31.9° from the requested arc, leaving the bear in a wrong target-view location (R·cMaskIoU $= 0.402$). TrajectoryCrafter [68] and Gen3C [48] follow the trajectory but break the bear's body, dropping recognition to $R = 0.444$ and 0.800. Only FreeOrbit4D [9] both follows the arc (RotErr $= 3.53°$) and preserves the bear (R·cMaskIoU $= 0.981$).

Metric evaluation. Per evaluated method, the camera-accuracy protocol re-runs ViPE [23] on the generated video; this takes ∼3.6 A100-hours total (∼27 min wall on 8 GPUs). Subject-mask extraction with SAM [10] costs under 1 A100-hour per method (∼5 min wall per shard on 8 GPUs). Subject recognition with Qwen3-VL-32B [5] in 4-bit quantization costs ∼10 A100-hours per method (∼75 min wall on 8 GPUs).

Aggregate budget. Producing all numbers reported in the paper (the four-baseline benchmark plus all camera-accuracy and subject-fidelity metrics on 83 cases) takes roughly 200 A100-80GB GPU-hours on top of the one-time ∼40-hour construction cost. This excludes preliminary runs and failed configurations that are not reported in the paper, which together used a comparable amount of compute, primarily for tuning baseline trajectory encoding and validating the metric pipeline.

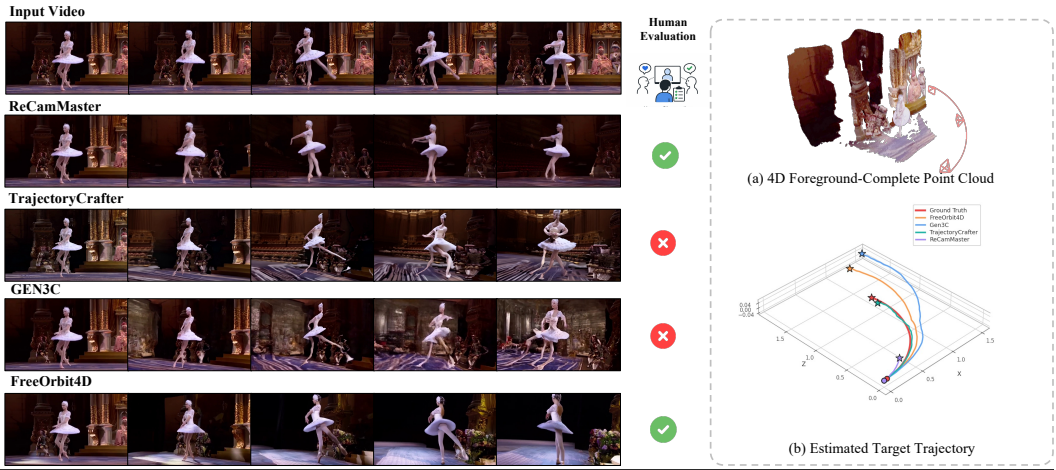
A.4 Additional Per-Case Comparisons

We provide seven additional per-case comparisons (Figures 8 to 14) using the same layout as Figure 5. Each figure displays, on the left, the input source video and the four baselines' generated target-view videos, with a human-evaluation mark per baseline; on the right, panel (a) shows the foreground-complete 4D point cloud with the target trajectory arc, and panel (b) overlays the target trajectory with each baseline's ViPE-recovered trajectory in 3D. The accompanying per-case table reports the same columns as Table 1 except that aggregate distribution metrics (FID/FVD) are omitted; cIoU abbreviates cMaskIoU, and **bold** marks the best score per column. Each case is chosen to expose a specific discrepancy between traditional video metrics (CLIP [45], VBench [25]) and task-level visual assessment: traditional scores frequently reward outputs that are visibly broken in ways the Redirect4D-Bench metrics surface directly.



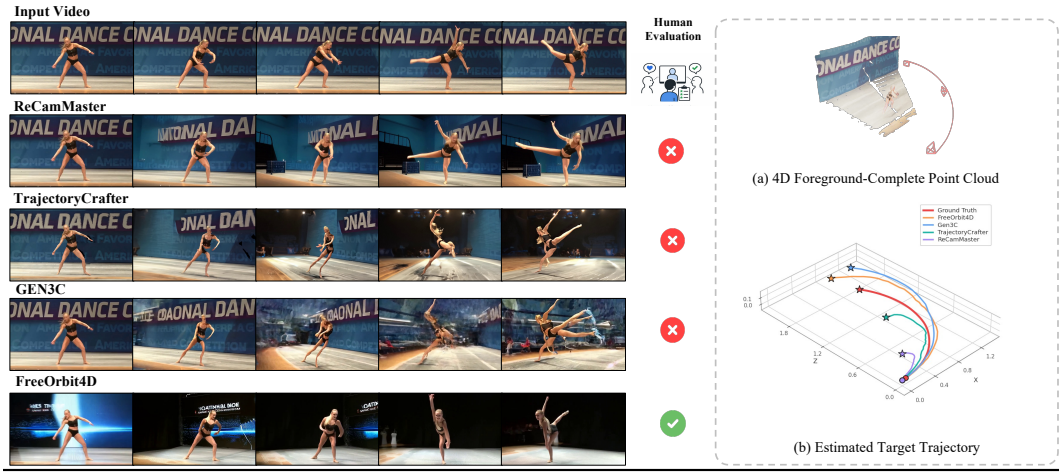
	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	0.318	**0.976**	0.863	**0.898**	0.889	**0.969**	**0.988**	0.547	0.566	0.302	24.855	0.673	**1.000**	**1.000**	0.043	0.043
TrajectoryCrafter	0.330	0.954	0.846	0.855	0.872	0.935	0.975	0.533	0.530	0.277	14.633	1.157	**1.000**	0.756	0.407	0.308
Gen3C	**0.349**	0.963	**0.881**	0.827	0.886	0.940	0.981	0.502	0.549	**0.306**	18.838	0.777	**1.000**	0.800	0.373	0.298
FreeOrbit4D	0.343	0.967	0.875	0.873	**0.917**	0.940	0.974	**0.569**	**0.585**	0.305	**12.104**	**0.639**	**1.000**	**1.000**	**0.886**	**0.886**

Figure 9: **Dancer case 1.** ReCamMaster [4] scores best on three VBench dimensions and produces visually plausible frames, yet its camera drifts 24.9° off the requested arc and the dancer ends up almost entirely outside the target view (R·cMaskIoU $=$ 0.043). TrajectoryCrafter [68] and FreeOrbit4D [9] both follow the trajectory while keeping the dancer recognizable, with FreeOrbit4D placing her best (R·cMaskIoU $=$ 0.886 vs. 0.308).



	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	0.345	**0.988**	**0.945**	0.872	0.890	**0.971**	**0.987**	0.538	0.496	**0.261**	24.466	0.588	**1.000**	**1.000**	0.408	0.408
TrajectoryCrafter	**0.354**	0.975	0.886	0.832	**0.937**	0.939	0.977	0.540	0.567	0.254	1.953	**0.052**	**1.000**	0.422	0.519	0.219
Gen3C	0.348	0.969	0.880	0.834	0.901	0.934	0.977	**0.544**	0.521	0.256	2.781	0.548	**1.000**	0.222	0.567	0.126
FreeOrbit4D	0.332	0.966	0.907	**0.875**	0.909	0.955	0.984	0.499	**0.647**	0.257	**1.648**	0.427	**1.000**	**1.000**	**0.905**	**0.905**

Figure 10: **Dancer case 2.** TrajectoryCrafter [68] (RotErr 1.95°, R·cMaskIoU $=$ 0.219) and Gen3C [48] (RotErr 2.78°, R·cMaskIoU $=$ 0.126) track the trajectory closely yet break the ballerina's limbs. ReCamMaster [4] keeps the dancer intact but barely orbits her (RotErr $=$ 24.5°, R·cMaskIoU $=$ 0.408); only FreeOrbit4D [9] jointly satisfies both axes (RotErr $=$ 1.65°, R·cMaskIoU $=$ 0.905).

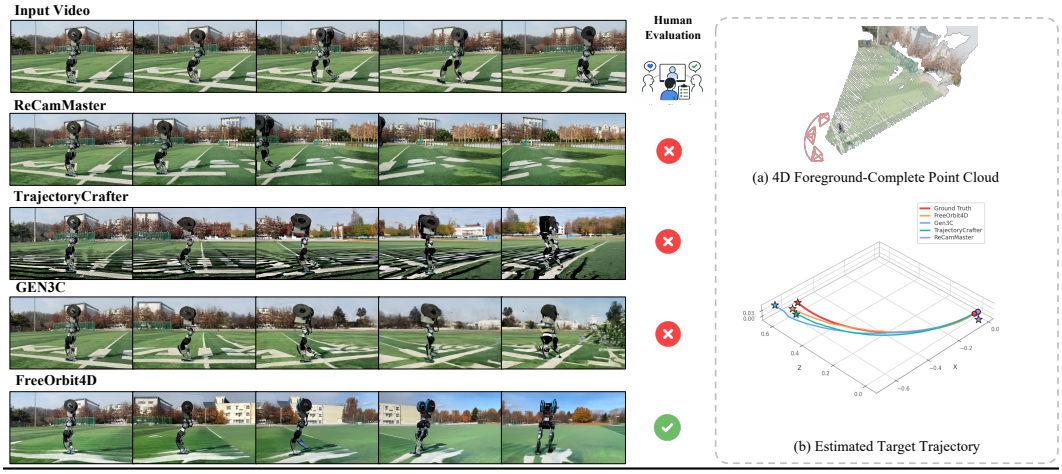


	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	0.294	**0.979**	**0.910**	**0.915**	0.920	**0.959**	**0.985**	0.449	**0.713**	0.274	24.440	0.615	**1.000**	**1.000**	0.235	0.235
TrajectoryCrafter	**0.316**	0.960	0.862	0.861	0.907	0.939	0.977	**0.451**	0.560	**0.295**	6.469	0.491	**1.000**	0.578	0.412	0.238
Gen3C	0.304	0.948	0.832	0.826	0.882	0.927	0.972	0.431	0.576	0.289	**1.977**	**0.237**	**1.000**	0.311	0.238	0.074
FreeOrbit4D	0.269	0.960	0.795	0.846	**0.934**	0.952	0.975	0.416	0.623	0.274	4.823	0.269	**1.000**	**1.000**	**0.899**	**0.899**

Figure 11: **Dancer case 3.** Gen3C [48] wins both camera-pose columns (RotErr $= 1.98°$, TransErr $= 0.24$ m) yet renders only fragments of the dancer (R·cMaskIoU $= 0.074$). TrajectoryCrafter [68] also tracks the path (RotErr $6.47°$) but distorts her body (R·cMaskIoU $= 0.238$); ReCamMaster [4] barely orbits and ends up off-screen (RotErr $24.4°$, R·cMaskIoU $= 0.235$). FreeOrbit4D [9] is the only baseline that keeps the dancer intact along the arc (R·cMaskIoU $= 0.899$).

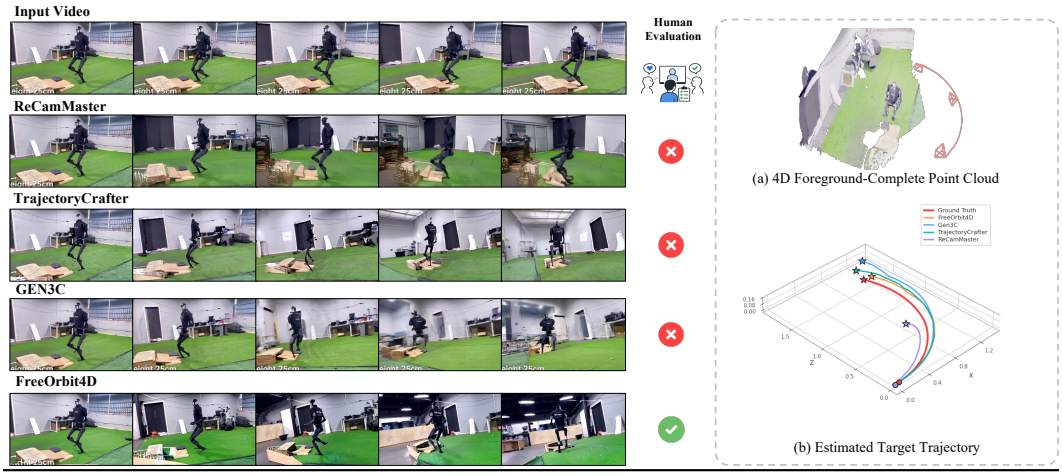
	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	0.300	**0.974**	0.833	**0.839**	0.869	**0.901**	0.957	**0.472**	0.690	**0.294**	32.941	0.897	**1.000**	**1.000**	0.072	0.072
TrajectoryCrafter	**0.311**	0.944	0.789	0.778	0.869	0.875	0.940	0.418	**0.697**	0.253	**3.050**	0.162	**1.000**	0.733	0.480	0.352
Gen3C	0.310	0.939	0.830	0.807	0.899	0.896	**0.959**	0.430	0.552	0.280	10.171	0.108	**1.000**	0.689	0.548	0.378
FreeOrbit4D	0.310	0.933	**0.874**	0.830	**0.913**	0.872	0.931	0.466	0.693	0.285	4.200	**0.089**	**1.000**	**1.000**	**0.886**	**0.886**

Figure 12: **Deer case.** ReCamMaster [4] wins several VBench columns yet diverges $32.9°$ in rotation and the deer disappears almost entirely from the late frames (R·cMaskIoU $= 0.072$). TrajectoryCrafter [68] (RotErr $3.05°$) and Gen3C [48] (RotErr $10.17°$) follow the path more closely but leave the deer fragmented (R·cMaskIoU $= 0.352$ and 0.378). FreeOrbit4D [9] both follows the arc (RotErr $= 4.20°$) and preserves the deer (R·cMaskIoU $= 0.886$).



	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	**0.323**	**0.962**	**0.900**	**0.879**	0.891	**0.949**	**0.981**	**0.533**	0.619	**0.311**	28.421	**0.327**	**1.000**	**0.956**	0.294	0.281
TrajectoryCrafter	0.306	0.943	0.828	0.814	0.875	0.906	0.961	0.438	0.616	0.292	10.922	0.784	**1.000**	0.089	0.483	0.043
Gen3C	0.282	0.928	0.732	0.776	0.867	0.927	0.975	0.453	0.462	**0.311**	**2.412**	0.726	**1.000**	0.067	0.372	0.025
FreeOrbit4D	0.321	0.944	0.841	0.871	**0.926**	0.917	0.976	0.513	**0.689**	0.297	8.779	0.607	**1.000**	**0.956**	**0.881**	**0.842**

Figure 13: **Robot case 1.** ReCamMaster [4] wins TransErr (0.327 m) but its camera barely orbits (RotErr $= 28.4°$) and the robot is placed in the wrong location (R·cMaskIoU $= 0.281$). Gen3C [48] (RotErr $2.41°$) and TrajectoryCrafter [68] track the path but lose the robot in most frames (R·cMaskIoU $= 0.025$ and 0.043). Only FreeOrbit4D [9] keeps the robot recognizable along the arc (R·cMaskIoU $= 0.842$).



	Traditional metrics										Redirect4D-Bench metrics					
	CLIP ↑			VBench ↑							Camera acc. ↓		Object fidelity ↑		Subject loc. ↑	
Method	T	F	V	SC	BG	TF	MS	AQ	IQ	OC	RotErr	TransErr	D	R	cIoU	R·cIoU
ReCamMaster	**0.335**	**0.981**	**0.925**	**0.895**	**0.937**	**0.954**	0.984	**0.561**	**0.726**	0.300	17.599	0.509	**1.000**	0.978	0.257	0.252
TrajectoryCrafter	**0.335**	0.974	0.876	0.851	0.921	0.934	0.979	0.480	0.682	0.325	**1.368**	0.134	**1.000**	0.244	0.302	0.074
Gen3C	0.323	0.971	0.883	0.874	0.918	0.943	**0.985**	0.514	0.586	0.332	1.464	0.209	**1.000**	0.622	0.127	0.079
FreeOrbit4D	0.319	0.952	0.865	0.870	0.910	0.912	0.981	0.549	0.659	**0.338**	1.477	**0.123**	**1.000**	**1.000**	**0.930**	**0.930**

Figure 14: **Robot case 2.** TrajectoryCrafter [68], Gen3C [48], and FreeOrbit4D [9] all reach low rotation error (under $2°$), yet only FreeOrbit4D preserves the target-view robot (R·cMaskIoU $=$ 0.930): ReCamMaster [4] recognizes the robot but mislocates it (R·cMaskIoU $= 0.252$ at RotErr$=$ $17.6°$), while TrajectoryCrafter and Gen3C fail recognition on most frames.