

A Camera Pose Estimation Results based on MAST3R

To further validate the generality of PoseCrafter on different pose estimation backbones, we replaced DUST3R [3] with MAST3R [4], another transformer-based framework augmented with a dense local feature head and efficient sparse matching. We evaluated this configuration on four benchmarks with varying yaw ranges: Cambridge Landmarks (50°–65°), ScanNet (50°–65°), DL3DV-10K (50°–90°), and NAVI (50°–90°). As shown in Table 11 and Table 12, using MAST3R directly for pose estimation results in lower accuracy compared to DUST3R. A possible explanation is that the sparse matching algorithm employed by MAST3R may be less effective for image pairs with small or no overlap. Nonetheless, our proposed framework consistently improves estimation accuracy over the baseline pose estimation models, including MAST3R. This demonstrates that our approach generalizes well across different pose estimation backbones by synthesizing intermediate frames that are more suitable for pose estimation.

Table 11: Camera pose estimation results on outward-facing datasets (Cambridge Landmarks and ScanNet) based on MAST3R. We report rotation recall ($R@ \theta \uparrow$), translation recall ($T@ \theta \uparrow$), mean rotation error (MRE \downarrow), mean translation error (MTE \downarrow), and AUC₃₀ \uparrow .

Method	Input	Cambridge Landmarks					ScanNet								
		R@5°	R@15°	R@30°	MRE	AUC ₃₀	R@5°	R@15°	R@30°	T@5°	T@15°	T@30°	MRE	MTE	AUC ₃₀
MASt3R	Pair	9.03	43.75	57.99	51.61	36.15	25.00	55.17	62.93	6.90	30.17	48.28	36.34	50.13	26.84
Ours(MASt3R)	Hybrid video	25.00	59.03	69.10	41.02	49.83	31.90	65.52	77.59	18.97	45.69	63.79	29.97	38.50	40.66

Table 12: Camera pose estimation results on center-facing datasets (DL3DV-10K and NAVI) based on MAST3R. We report rotation recall ($R@ \theta \uparrow$), translation recall ($T@ \theta \uparrow$), mean rotation error (MRE \downarrow), mean translation error (MTE \downarrow), and AUC₃₀ \uparrow .

Dataset	Method	Input	R@5°	R@15°	R@30°	T@5°	T@15°	T@30°	MRE	MTE	AUC ₃₀
DL3DV-10K	MASt3R	Pair	7.00	63.67	97.00	26.67	72.67	91.67	15.18	13.31	53.49
	Ours(MASt3R)	Hybrid video	7.67	68.67	97.33	30.00	76.33	92.00	14.09	12.91	55.36
NAVI	MASt3R	Pair	43.97	93.39	96.89	50.97	89.49	97.67	8.30	7.43	76.54
	Ours(MASt3R)	Hybrid video	43.97	93.77	97.28	51.75	91.83	97.28	7.71	6.98	77.26

Table 13: Additional comparison with VGGT on Cambridge Landmarks. We report mean rotation error (MRE \downarrow), rotation recall ($R@ \theta \uparrow$), and AUC₃₀ \uparrow .

Method	Input	MRE	R@5°	R@15°	R@30°	AUC ₃₀
Dust3R	Pair	18.14	40.34	71.25	82.99	61.98
Ours	Hybrid video	11.40	55.21	89.93	93.75	77.41
VGGT	Pair	20.17	40.00	70.17	82.29	60.54
VGGT _{Ours}	Hybrid video	17.88	42.43	84.40	85.76	65.15

B Additional Comparison with VGGT

To further analyze the generality of our pipeline, we conducted comparative experiments with VGGT [39] on the Cambridge Landmarks dataset under yaw changes of [50°–65°].

We first evaluated the two models in challenging cases where only image pairs with small or no overlap were provided. As shown in the second and fourth rows of Table 13, DUST3R consistently outperforms VGGT on such data.

We then integrate these two models into our pipeline and evaluate their performance. As shown in the third and fifth rows, both configurations achieve significant improvements over their directly estimated counterparts. The version using DUST3R achieves a higher accuracy than the version using VGGT. These results demonstrate that our method is compatible with different pose estimators and consistently enhances their performance.

Overall, DUS3R appears better suited than VGGT for small- or non-overlapping image pairs, and our framework provides more noticeable improvements when combined with DUS3R. While fine-tuning VGGT on small-overlap data or adjusting the hyperparameters of our pipeline may further enhance its performance, we leave such extensions for future work.

C Comparison of DUS3R Confidence-based Selection and FMS

We compare DUS3R confidence-based frame selection with our proposed Feature Matching Selector (FMS) to evaluate whether confidence scores can serve as a viable alternative. In this setting, top-ranked frames were selected from hybrid videos using four different confidence thresholds (20%, 40%, 60%, and 80%). The results are summarized in Table 14.

When the threshold was set to 20% or 40%, pose estimation accuracy decreased compared to using the full hybrid video sequence. Increasing the threshold to 60% or 80% improved performance, but the accuracy still remained lower than that achieved by our proposed FMS.

In terms of efficiency, confidence-based selection introduces substantial overhead, as all video frames must be processed by DUS3R to compute confidence maps prior to selection. This step incurs significant time and memory costs, and higher thresholds further increase runtime as more frames are selected. By contrast, our FMS requires only a single feature extraction step, immediately identifying the most informative frames with superior efficiency and accuracy. These findings highlight the practical advantages of our proposed FMS over confidence-based selection in both computational efficiency and pose estimation performance.

Table 14: Comparison of DUS3R confidence-based selection and FMS on Cambridge Landmarks. We report pose estimation results under yaw changes of $[50^\circ-65^\circ]$. Frames were selected from hybrid videos using four DUS3R confidence thresholds (20%, 40%, 60%, and 80%). We report mean rotation error (MRE \downarrow), rotation recall ($R@ \theta \uparrow$), $AUC_{30} \uparrow$ and pose estimation time.

Method	MRE \downarrow	$R@5^\circ$	$R@15^\circ$	$R@30^\circ$	$AUC_{30} \uparrow$	Pose Estimation Time
Conf(20%)	14.66	54.17	85.07	89.24	72.45	2.79min
Conf(40%)	14.36	57.30	87.15	90.97	74.24	2.91min
Conf(60%)	12.35	54.17	90.28	93.06	76.60	3.47min
Conf(80%)	12.41	53.47	90.63	93.06	76.46	4.13min
Ours_{w/o} FMS	13.24	54.51	89.24	92.71	76.13	2.56min
Ours	11.40	55.21	89.93	93.75	77.41	0.18min

D Visualization Results of Limitation

Although PoseCrafter achieves robust results across various benchmarks, certain challenging scenes still degrade its performance. In the hybrid video generation (HVG) stage, severe illumination differences between the start and end frames will introduce obvious artifacts in the synthesized intermediate views. The **red-boxed** regions in Figure 4 highlight these artifacts, which would be marked as low confidence and discarded in the downstream pose estimation backbone, e.g., DUS3R in this work. Such an obvious removal of information may significantly degrade the final performance. Moreover, in the feature match selector (FMS) module, scenes dominated by uniform or repetitive textures hinder the reliable extraction and matching of key points, resulting in fewer RANSAC [10] inliers and lower pose accuracy. The **red** lines in Figure 5 show examples of incorrect correspondences in such scenes.

E More Visualizations of Our Generated Videos

To further illustrate the superiority of our proposed hybrid video generation(HVG), we present additional side-by-side comparisons of intermediate frames produced by DynamiCrafter, ViewCrafter, and PoseCrafter. In Figure 6, DynamiCrafter [6] delivers temporally smooth transitions but exhibits progressive blur and geometric drift in the middle of generated sequences. Although ViewCrafter [9]



Figure 4: Hybrid Video Generation artifacts under severe illumination differences. Blue and yellow outlines indicate the start and end frames, respectively. Confidence images denote images filtered with the predicted confidence map in the subsequent DUS3R model. The red boxes highlight regions affected by artifacts in these cases. We can observe that these regions have quite low confidence.

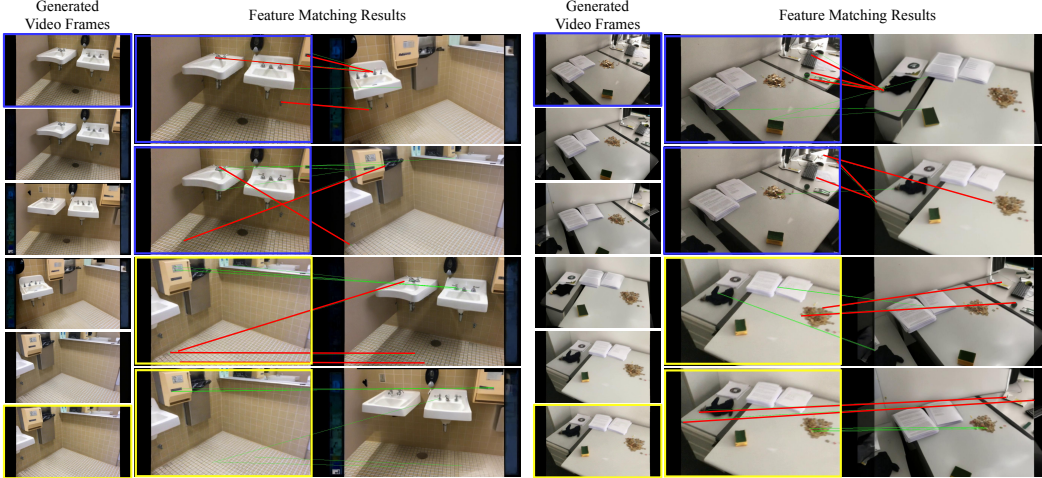


Figure 5: Feature Matching Selection failures in low-texture regions. Blue and yellow outlines indicate the start and end frames, respectively. The red and green lines indicate incorrect and correct correspondences, respectively. Incorrect correspondences lead to errors in inlier counting, affecting the accuracy of subsequent frame selection.

can produce sharp results with minimal blur, using only input image pairs with small overlap often leads to structural artifacts and misalignments. To address these limitations, we combine DynamiCrafter and ViewCrafter to complement each other’s strengths. Specifically, we first use DynamiCrafter to synthesize intermediate “relay” frames, effectively augmenting the input image pair with frames that have larger overlaps. These relay frames are then passed to ViewCrafter to generate clearer and more geometrically consistent results. As shown in Figure 6, our proposed approach successfully produces frames that are both visually sharp and structurally reliable.

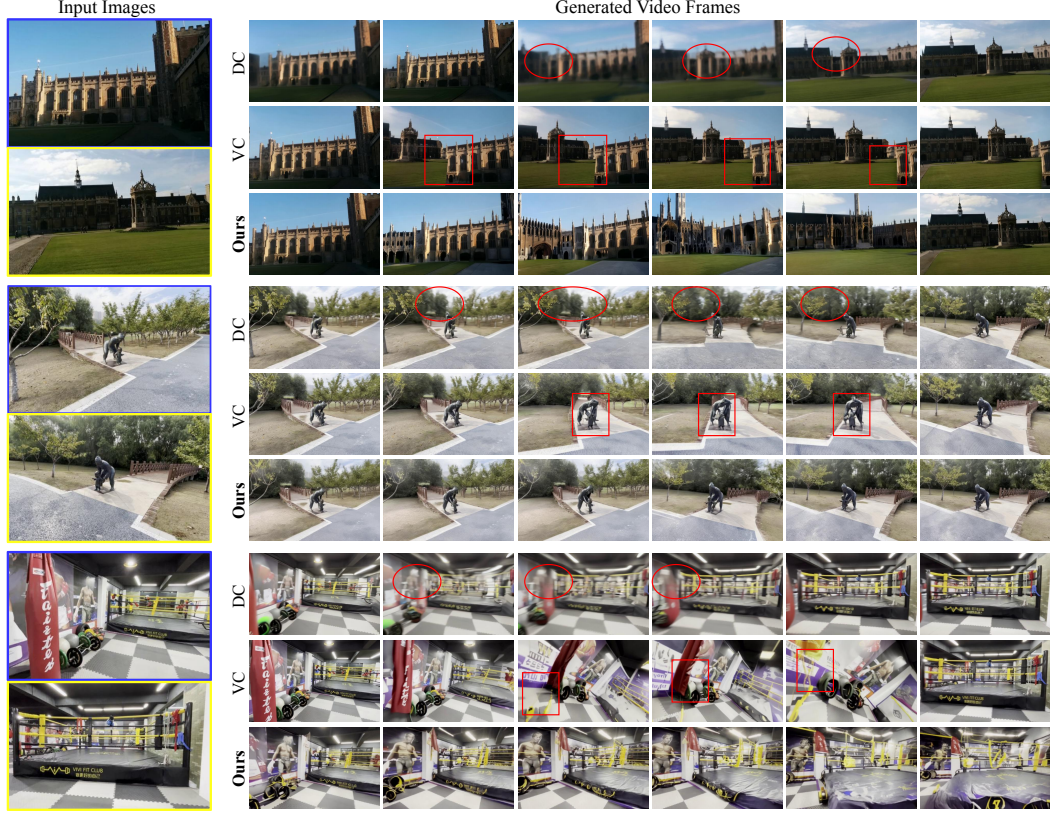


Figure 6: Comparative video synthesis results. Each row shows intermediate frames synthesized between the same start frame (blue box) and end frame (yellow box), generated by different methods: DynamiCrafter (DC), ViewCrafter (VC), and our Hybrid Video Generation (Ours). DC produces smooth motion but exhibits progressive blur and geometric drift in the middle of sequences (highlighted in red circle). Since VC is sensitive to the pose of the input image pair, it tends to produce structural misalignments in our small-overlap setting (highlighted in red box). By coupling DC and VC together, our method delivers sharp, geometrically consistent video frames throughout, correcting both the blur of DC and the misalignments of VC.