Can Generative Video Models Help Pose Estimation?

Supplementary Material

1. Qualitative Results

We provide additional qualitative results, including more examples with videos generated from four different prompts and three video generation models across four datasets. For more visualizations and interactive DUSt3R point clouds, please visit our project page: Inter-Pose.github.io.

2. Effectiveness of our method across different yaw changes

In addition to the small overlapping pairs with yaw changes in the ranges of $[50°, 65°]$ for outward-facing datasets and $[50°, 90°]$ for center-facing datasets, as described in the main paper, we conducted further experiments to evaluate the effectiveness of our proposed method on image pairs with either significant overlap or no overlap. These experiments specifically examine the impact of varying yaw angle changes between image pairs.

ScanNet [1]: For this outward-facing, indoor dataset, we sampled 200 pairs with yaw changes in the range of $[0°, 50°]$ to represent pairs with large overlap, and 200 pairs with yaw changes in the range of $[65°, 180°]$ to represent non-overlapping pairs.

DL3DV-10K [3]: This is a dataset consisting of outdoor scenes with center-facing camera viewpoints. We sampled 200 large-overlap pairs (with yaw changes in the range $[0°, 50°]$ and 200 pairs with larger yaw changes in the range $[90°, 180°]$.

For each pair, we the settings described in the main paper by generating four videos using Dream Machine. For each video, We randomly selected 11 subsets of 3 frames, along with the original image pair, and used these subsets as input to the DUSt3R pose estimator. We then computed the total medoid distance of the predicted relative transformations and selected the prediction with the lowest distance as the final relative pose estimate.

In Fig. 1, we present camera pose estimation performance vs. yaw angle change using the metrics of mean rotation error (MRE), mean translation error (MTE), and $AUC_{30°}$. As the yaw angle between input image pairs increases, the overlap between images decreases, resulting in higher MRE and MTE for both DUSt3R and our method. our method consistently achieves lower errors than DUSt3R for yaw changes below $110°$ on both the ScanNet and DL3DV-10K datasets.

We provide quantitative results with more metrics on ScanNet in Tables 5 and 6. For large-overlap pairs, our method, which incorporates generated frames from the video model, outperforms DUSt3R (when DUSt3R only uses the input image pair). Specifically, the mean rotation and translation errors decreased from (11.33°, 22.50°) to (9.12°, 15.75°) when using Dream Machine. For non-overlapping pairs, adding the generated video as input to the pose estimator yields comparable performance to using only the original image pair. This may be due to the ambiguity and multiple possibilities inherent in pairs with no overlap.

Quantitative results for DL3DV-10K are shown in Tables 7 and 8. For large-overlap pairs, our method (using the generated frames from generative videos obtains better results than DUSt3R, reducing mean rotation and translation errors from (4.28°, 11.04°) to (3.23°, 8.16°). For pairs with yaw changes in $[90°, 180°]$, the center-facing nature of the DL3DV-10K dataset still results in some overlapping regions. Incorporating the generated video as input improves performance by increasing $R_{acc}@30°$ and $T_{acc}@30°$ from (85.50%, 87.00%) to (89.50%, 91.50%). These results also indicate that center-facing datasets like DL3DV-10K are significantly easier for pose prediction than ScanNet and Cambridge Landmarks, which have many outward-facing camera viewpoints.

3. Additional results

We provide additional results for our method variants, as well as for the additional open-source video model CogVideoX-Interpolation [2, 4] in Table 3 and 4.

3.1. Variants of our method

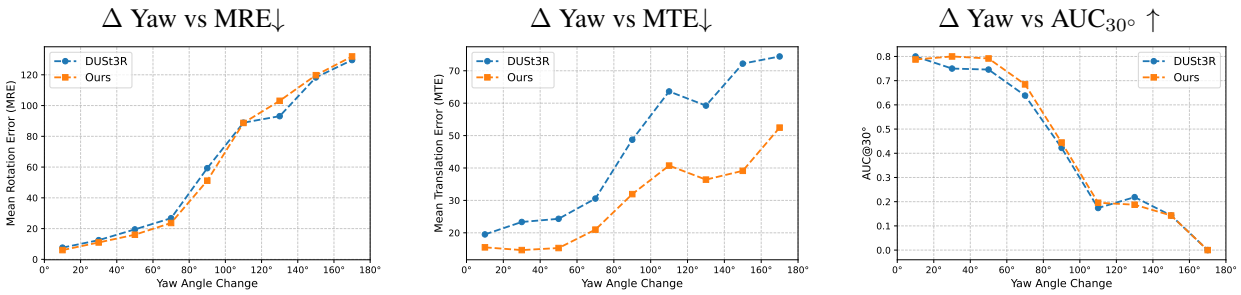
Best Medoid: We use the medoid relative transformation predicted from the generated video with the lowest total medoid distance (see Section 3.2 of the main paper).

Average: To evaluate the contribution of our self-consistency score using the medoid distance, we also evaluate an approach that takes the average of all $n \cdot m$ predictions from the video model. This tells us whether frames from a video model without any heuristic selection can still help with pose estimation.

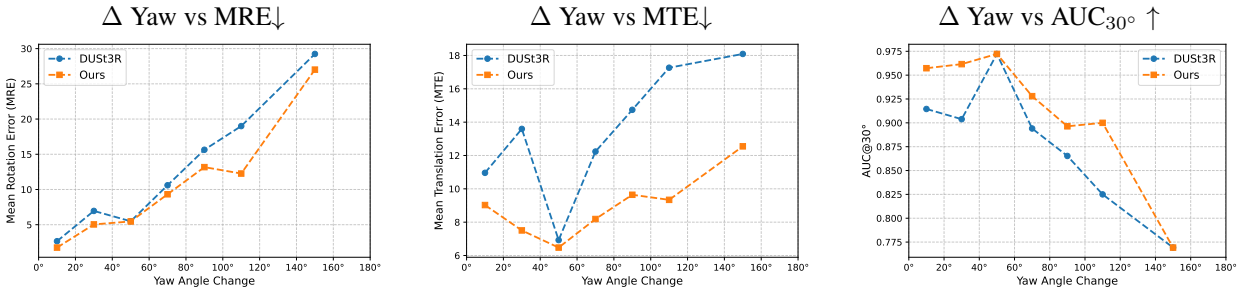
Oracle: This picks the best possible set of poses with the minimal rotation and translation error among all $n \cdot m$ generated predictions from all three video models. This serves as an upper-bound for a ground-truth heuristic selection.

3.2. Effectiveness of self-consistency score

We observe that simply averaging pose predictions from generated frames leads to worse performance than just tak-



(a) Camera Pose Estimation Performance vs. Yaw Angle Change on the ScanNet Dataset.



(b) Camera Pose Estimation Performance vs. Yaw Angle Change on the DL3DV-10K Dataset.

Figure 1. **Camera Pose Estimation Performance vs. Yaw Angle Change on the ScanNet and DL3DV-10K Datasets.** Comparison of Mean Rotation Error (MRE), Mean Translation Error (MTE), and Area Under Curve at 30° ($AUC_{30°}$) across different yaw angle change intervals (0°, 20°, 40°, 60°, etc.) Each data point represents the average value of the respective metric within a specific yaw angle range. Our method consistently achieves lower errors than DUSt3R for yaw angle changes below 110° on both datasets. Due to the limited number of sample pairs with yaw angle changes larger than 120° in the DL3DV-10K dataset, we report the results averaged over the [120°, 180°] range.

ing original image pair as input. For instance, in Table 3 on the Cambridge Landmarks dataset, using our method with the DUSt3R pose estimator, averaging among the predictions using Dream Machine's frames is even worse than not using a video model at all, with the mean rotation error increasing from 13.28° to 21.85°. By using our self-consistency metric, the mean rotation error of predictions with Dream Machine reduces to 11.96°. This validates the necessity and effectiveness of our medoid-based selection strategy in filtering out low-quality videos and unreliable predictions, thereby preventing degeneration in pose accuracy.

The Oracle outperforms all methods by a wide margin. This implies that with sufficient samples, it is possible for a video generation model to produce frames that are highly informative for pose estimation. It also suggests that there is still significant room for improving the selection method for reliably identifying the best generated frames or videos for pose estimation.

3.3. Additional video model

CogVideoX-Interpolation [2, 4]: CogVideoX is a large-scale diffusion transformer model for text-to-video generation, capable of producing continuous videos aligned with text prompts. CogVideoX-Interpolation extends this framework with a modified pipeline that enhances flexibility in keyframe interpolation. It generates 49 frames at a resolution of 720×480.

In Table 3 and 4, using generated frames from the CogVideoX-Interpolation model consistently outperforms the baseline pose estimator, which only uses the original image pair as input, on outward-facing datasets (Cambridge Landmarks and ScanNet). For example, on the ScanNet dataset, using DUSt3R as the pose estimator, the mean rotation and translation errors decrease from (21.31°, 24.72°) to (19.41°, 17.01°). On center-facing datasets (DL3DV-10K and NAVI), results with CogVideoX do not surpass those of commercial video models like Runway in terms of rotation and translation errors. However, they still achieve comparable performance.

3.4. Additional analysis of MASt3R

On the Cambridge Landmarks and ScanNet datasets, many image pairs feature outward-facing camera viewpoints and have no overlap. This lack of overlap and correspondence results in MASt3R exhibiting performance that is significantly worse than that of DUSt3R, especially on the Cambridge Landmarks dataset. As shown in Figure 2, MASt3R completely fails in scenarios with no overlap. Our method, with MASt3R as the pose estimator, still achieves improvements on both outward-facing datasets.

Figure 2. **Failure examples of MASt3R.** We show instances where MASt3R fails to accurately predict poses on non-overlapping pairs from the Cambridge Landmarks (top row) and ScanNet (bottom row) datasets. MASt3R relies on feature matching for pose refinement, which is insufficient and less reliable when pairs lack overlapping regions. In contrast, DUSt3R demonstrates greater robustness in these scenarios.

3.5. Runtime comparisons

The baseline DUSt3R runs in ∼1.8s/pair. Our pipeline adds ∼3s/pair for captioning and additional time for video generation—about 45s/video with DynamiCrafter and 20s/API call with Dream Machine and Runway Gen 3 Turbo, with 4 generated videos per pair. Our method also runs DUSt3R 11 times on multiple subsampled frame sets on each video. Our total runtime is thus about 2-4 min per pair. Although our method is less efficient, it effectively integrates video models in pose estimation. Direct comparisons with commercial models are challenging due to undisclosed hardware configurations and limited access.

4. Ablation Study

4.1. Ablation study on distance metrics

In the main paper, we quantify video inconsistency using the medoid distance D_{med}. We also define the total distance as

$$D_{\text{total}} = D_{\text{med}} + \text{dist}\left(\hat{T}_{\text{med}}, f_{\text{pose}}(\{I_A, I_B\})\right), \quad (1)$$

where $\hat{T}_{\text{med}}$ is the medoid relative pose, and $f_{\text{pose}}(\{I_A, I_B\})$ is the pose estimated from the original image pair. We select the video with the lowest D_{total} and output the predicted medoid relative pose $\hat{T}_{\text{med}}$ as the consensus pose.

In Table 1, we present an ablation study on the distance metrics by comparing predictions based on D_{total}, D_{med}, and D_{bias}, where

$$D_{\text{bias}} = \text{dist}\left(\hat{T}_{\text{med}}, f_{\text{pose}}(\{I_A, I_B\})\right). \quad (2)$$

Intuitively, ensembling predictions from a mixture of experts improves robustness, especially since pose estimators like DUSt3R are specifically trained for pose estimation tasks on 3D datasets. Our results show that when using DUSt3R as the pose estimator,both D_{total} and D_{med} obtain comparable results across most datasets and video models, consistently outperforming the DUSt3R baseline, which only takes original image pairs. However, on the Cambridge Landmarks dataset with Dream Machine as the generative video model, using D_{med} alone increases the rotation error from $11.96°$ to $19.37°$ compared to D_{total}. Adding the bias term improves the worst-case performance, particularly on Cambridge Landmarks with Dream Machine. Ablation studies with MASt3R further confirm that the bias term is beneficial across most datasets and video models. These results demonstrate that incorporating D_{bias} into the distance metric enhances robustness and generalization ability across different datasets and video models.

4.2. Ablation study on the number of input images

The oracle showing the tendency as worse performance when using more video frames, which is likely due to less randomless in sampling, and also video might contain inconsistent content, which might degenerate the performance if the original input pair is less considered in pose estimation and post-optimization process.

We present an ablation study on the number of input images to the pose estimator in Table 2. The baseline DUSt3R takes only the original image pair as input, utilizing two images. To explore the impact of varying the number of input frames, we conducted experiments with 3, 5, 10, 40, and 116 images. These configurations correspond to sampling 1, 3, 8, 38, and 114 frames from the video generated by Dream Machine, respectively. Since the Dream Machine video consists of 114 frames in total, the configuration with 116 images involves sampling all frames once,

while the other configurations involve multiple sampling iterations (11 times for all except the 116-image setup).

The results indicate that using five images, as adopted in the main paper, yields the best performance across most metrics, including Mean Rotation Error (MRE), Mean Translation Error (MTE), and AUC_{30°. In addition, the oracle results reveal a trend of degenerating performance as the number of video frames increases. This decline is likely due to reduced randomness in sampling and the less-emphasisis on the original input pair during the pose estimation and post-optimization processes. Overall, these results indicate that using five frames provides a robust and generalizable approach, avoiding the pitfalls associated with both insufficient and excessive frame counts.

Pose estimator	Input data	Distance metric		Cambridge		ScanNet			DL3DV-10K			NAVI		
		D_{med}	D_{bias}	MRE↓	AUC_{30} ↑	MRE↓	MTE↓	AUC_{30} ↑	MRE↓	MTE↓	AUC_{30} ↑	MRE↓	MTE↓	AUC_{30} ↑
DUSt3R	Pair	–	–	13.28	77.23	21.31	24.72	60.34	10.72	13.08	66.99	8.65	7.88	78.66
Ours	DynamiCrafter	✓	✓	12.70	79.00	18.96	16.42	62.14	10.02	9.13	67.97	8.26	6.57	78.78
	DynamiCrafter	✓		12.58	80.31	18.26	16.58	62.94	9.42	8.93	68.89	7.12	6.31	78.23
	DynamiCrafter		✓	12.88	77.32	20.25	21.91	60.43	10.48	12.34	67.24	8.78	8.05	78.00
DUSt3R	Pair	–	–	13.28	77.23	21.31	24.72	60.34	10.72	13.08	66.99	8.65	7.88	78.66
Ours	Runway	✓	✓	10.78	80.59	19.93	16.31	61.83	9.49	8.81	69.44	8.08	6.24	79.02
	Runway	✓		10.77	80.91	21.27	16.66	61.33	9.11	9.26	69.87	6.70	6.15	78.36
	Runway		✓	12.13	78.52	20.68	18.78	61.29	10.08	12.13	68.04	8.18	7.37	78.66
DUSt3R	Pair	–	–	13.28	77.23	21.31	24.72	60.34	10.72	13.08	66.99	8.65	7.88	78.66
Ours	Dream Machine	✓	✓	11.96	78.67	17.65	15.88	63.06	9.13	8.72	69.11	7.85	6.51	79.06
	Dream Machine	✓		19.37	71.63	18.28	15.49	62.89	8.60	8.48	70.10	7.53	6.66	78.34
	Dream Machine		✓	11.25	79.08	20.24	19.45	60.80	10.15	11.98	67.70	8.36	7.59	78.70

(a) Ablation study of distance metrics on DUSt3R.

Pose estimator	Input data	Distance metric		Cambridge		ScanNet			DL3DV-10K			NAVI		
		D_{med}	D_{bias}	MRE↓	AUC_{30} ↑	MRE↓	MTE↓	AUC_{30} ↑	MRE↓	MTE↓	AUC_{30} ↑	MRE↓	MTE↓	AUC_{30} ↑
MASt3R	Pair	–	–	36.55	55.69	24.35	17.93	55.10	4.13	3.88	87.22	5.59	5.23	80.84
Ours	DynamiCrafter	✓	✓	31.43	60.03	21.97	16.48	57.90	4.49	4.04	85.86	5.29	5.61	80.21
	DynamiCrafter	✓		34.07	55.80	22.00	17.44	55.52	5.16	5.02	82.57	5.60	7.70	73.13
	DynamiCrafter		✓	33.14	57.97	21.83	17.22	57.14	4.38	4.09	85.97	5.34	5.29	80.86
MASt3R	Pair	–	–	36.55	55.69	24.35	17.93	55.10	4.13	3.88	87.22	5.59	5.23	80.84
Ours	Runway	✓	✓	29.04	63.57	21.68	15.28	57.19	4.17	4.01	86.79	5.28	5.2	81.63
	Runway	✓		32.10	59.69	22.50	15.41	53.56	4.87	5.02	83.56	6.14	6.78	77.11
	Runway		✓	29.78	62.54	22.14	16.46	57.00	4.22	4.08	86.91	5.27	5.31	81.02
MASt3R	Pair	–	–	36.55	55.69	24.35	17.93	55.10	4.13	3.88	87.22	5.59	5.23	80.84
Ours	Dream Machine	✓	✓	27.47	63.14	19.91	15.05	58.28	4.30	4.21	85.88	5.66	5.45	81.42
	Dream Machine	✓		28.83	62.56	21.27	15.94	57.17	4.45	4.75	82.56	6.39	7.14	76.19
	Dream Machine		✓	29.29	62.32	21.31	16.81	57.22	4.24	4.16	86.20	5.19	5.39	80.83

(b) Ablation study of distance metrics on MASt3R.

Table 1. Ablation study of distance metrics. Our proposed distance metric incorporates both the medoid distance D_{med} and the bias distance D_{bias}, where D_{bias} is defined as $D_{\text{bias}} = \text{dist}\left(\hat{T}_{\text{med}}, f_{\text{pose}}(\{I_A, I_B\})\right)$. We perform an ablation study to evaluate the contribution of each distance term. While D_{med} and the total distance yield comparable results across most datasets and video models, solely considering D_{med} leads to significantly worse performance on the Cambridge dataset when using the Dream Machine video model. Incorporating the total distance enhances generalization ability and robustness across various datasets and video models.

Pose estimator	Input data	# Images	# Samples	MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			AUC_{30} ↑
						5°	15°	30°	5°	15°	30°	
DUSt3R	Pair	2+0	1	21.31	24.72	65.33	76.33	79.00	48.33	68.33	73.67	60.34
Ours	Pair+Dream Machine	2+1	11	20.41	16.93	67.00	79.00	81.67	50.00	**71.33**	81.33	62.08
		2+3	11	**17.65**	**15.88**	**68.67**	81.33	**85.33**	47.67	**71.33**	**82.33**	**63.06**
		2+8	11	17.98	16.39	66.00	81.33	85.00	**50.67**	70.67	81.33	61.96
		2+38	11	18.43	16.70	65.00	**82.33**	**85.33**	50.33	**71.33**	79.33	62.76
		2+114	1	17.77	17.05	65.67	**82.33**	**85.33**	49.33	70.00	80.00	62.00
Oracle	Pair+Dream Machine	2+1	11	5.71	5.84	81.67	93.67	95.67	72.33	90.00	96.33	80.08
		2+3	11	5.80	5.00	81.33	94.33	95.00	73.33	91.00	96.67	81.19
		2+8	11	6.81	6.00	81.33	91.67	94.00	71.67	87.67	96.00	78.20
		2+38	11	7.42	7.10	78.33	91.33	93.67	65.33	84.33	94.67	75.26
		2+114	1	9.21	9.68	74.33	89.33	92.67	59.67	77.33	90.67	70.70

Table 2. **Ablation study on the number of input images to the pose estimator on ScanNet dataset.** "# Images" denotes the total number of images provided to the DUSt3R pose estimator, where 2 images are from the original pair and the remaining images are sampled from the generated video. Using 5 images, as used in the main paper, shows the best performance. "# Samples" indicates the sampling iterations per video. For the experiment with 2+114 images, only one sampling was conducted instead of 11, since the video consists of 114 frames in total.

Pose estimator	Input data	Cambridge Landmarks					ScanNet								
		MRE↓	R_{acc} ↑			AUC_{30} ↑	MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			AUC_{30} ↑
			5°	15°	30°				5°	15°	30°	5°	15°	30°	
SIFT+N.N.	Pair	97.64	15.17	22.41	24.48	20.49	112.95	48.99	2.06	3.44	5.50	23.02	25.09	31.62	1.82
LOFTR		30.30	31.38	56.55	70.00	51.63	64.46	45.49	8.33	17.00	22.00	27.00	28.33	35.33	6.43
DUSt3R		13.28	63.45	87.24	88.97	77.23	21.31	24.72	65.33	76.33	79.00	48.33	68.33	73.67	60.34
MASt3R		36.55	28.62	64.83	74.14	55.69	24.35	17.93	44.00	73.33	79.67	38.00	67.33	77.67	55.10
Ours-Avg. (DUSt3R)	DynamiCrafter	13.22	60.00	86.90	89.66	76.36	19.97	18.87	62.33	78.67	83.00	45.33	67.33	74.33	58.84
	CogVideoX	13.35	54.83	85.86	90.00	74.91	22.04	20.21	55.67	73.00	77.33	39.33	63.67	71.33	54.90
	Runway	12.49	47.59	84.14	90.69	72.93	22.87	18.96	57.33	73.67	79.00	36.67	64.33	72.67	54.77
	Dream Machine	21.85	31.38	69.66	80.00	59.39	22.44	19.82	50.33	67.00	75.00	36.67	59.33	72.33	53.00
Ours-Medoid (DUSt3R)	DynamiCrafter	12.70	65.17	88.97	90.34	79.00	18.96	16.42	68.00	82.33	84.33	48.67	71.67	80.33	62.14
	CogVideoX	11.16	64.48	91.03	92.41	80.40	19.41	17.01	66.33	81.33	83.00	51.33	71.33	81.00	62.08
	Runway	10.78	64.83	91.03	94.14	80.59	19.93	16.31	67.67	81.33	84.33	51.00	72.33	80.67	61.83
	Dream Machine	11.96	57.93	89.66	92.76	78.67	17.65	15.88	68.67	81.33	85.33	47.67	71.33	82.33	63.06
Ours-Avg. (MASt3R)	DynamiCrafter	41.11	10.00	34.48	61.03	33.93	26.31	20.47	38.00	63.33	72.00	21.33	54.00	73.00	46.66
	CogVideoX	35.09	13.45	49.66	66.21	43.23	27.73	20.61	41.00	62.67	71.00	22.67	57.00	70.67	46.54
	Runway	36.75	9.31	45.86	64.48	40.34	26.81	19.96	31.00	62.00	71.67	19.00	54.33	75.00	46.04
	Dream Machine	36.23	12.76	49.66	60.69	40.97	26.29	20.71	32.00	60.67	73.00	20.00	56.00	71.33	44.67
Ours-Medoid (MASt3R)	DynamiCrafter	31.43	34.83	70.00	76.55	60.03	21.97	16.48	53.00	75.67	80.00	40.67	70.33	80.00	57.90
	CogVideoX	28.12	42.76	74.14	80.00	64.97	21.88	16.29	56.00	75.33	80.33	41.33	70.33	82.67	57.67
	Runway	29.04	42.07	72.76	78.97	63.57	21.68	15.28	50.33	75.67	81.67	41.00	70.00	83.33	57.19
	Dream Machine	27.47	34.48	74.14	80.69	63.14	19.91	15.05	53.00	78.67	83.00	41.00	70.33	82.33	58.28
Oracle	All Video Models	3.65	90.69	96.55	98.28	92.08	5.80	5.00	81.33	94.33	95.00	73.33	91.00	96.67	81.19

Table 3. **Camera pose estimation results on outward-facing datasets (Cambridge Landmarks and ScanNet).** We evaluate the pairwise pose estimation task using our method based on two pose estimators DUSt3R and MASt3R. We consider two variants of selection heuristics: averaging poses from randomly sampled frames (Avg.) and selecting the most self-consistent video using our minimal medoid distance metric (Medoid). Our method consistently outperforms both DUSt3R and MASt3R when using input pairs alone across three video generators. We also present an Oracle baseline that selects the best possible relative pose recovered from all generated videos.

Pose estimator	Input data	DL3DV-10K									NAVI								
		MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			$AUC_{30°}$↑	MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			$AUC_{30°}$↑
				5°	15°	30°	5°	15°	30°				5°	15°	30°	5°	15°	30°	
SIFT+N.N.	Pair	76.64	46.80	18.06	28.09	33.44	31.77	33.11	36.45	12.11	107.46	45.10	4.67	6.67	7.33	16.33	17.00	19.00	3.20
LOFTR		35.92	41.76	37.67	52.33	61.00	40.00	41.00	45.33	23.53	71.34	51.21	6.67	14.33	19.00	24.67	25.33	29.33	4.88
DUSt3R		10.72	13.08	39.67	87.33	94.00	55.33	83.67	89.00	66.99	8.65	7.88	68.67	92.67	94.67	69.00	92.33	95.00	78.66
MASt3R		4.13	3.88	83.67	98.00	99.33	88.33	95.33	97.00	87.22	5.59	5.23	71.67	94.33	98.00	69.67	96.00	98.00	80.84
Ours-Avg. (DUSt3R)	DynamiCrafter	10.45	11.30	37.33	88.67	95.00	49.33	83.67	89.67	65.76	8.39	7.86	57.00	91.33	96.00	56.00	89.00	97.00	74.33
	CogVideoX	10.74	11.47	34.00	86.33	95.00	45.67	82.33	89.33	64.14	9.32	8.72	50.00	92.00	94.67	49.00	88.33	96.00	72.38
	Runway	10.27	10.86	38.67	88.67	95.33	50.33	83.00	90.33	66.32	8.45	8.14	55.33	90.00	94.67	48.33	88.00	96.33	72.79
	Dream Machine	10.40	11.17	35.33	86.67	94.33	46.67	83.33	89.67	64.59	8.58	8.22	55.33	91.00	95.00	56.00	89.67	95.67	74.11
Ours-Medoid (DUSt3R)	DynamiCrafter	10.02	9.13	38.33	87.33	95.67	58.33	87.00	93.00	67.97	8.26	6.57	68.00	92.67	95.67	69.00	91.67	96.67	78.78
	CogVideoX	9.88	9.77	38.00	88.00	97.00	57.00	87.00	92.33	68.14	8.74	7.27	62.67	93.00	95.67	64.00	92.33	95.67	77.51
	Runway	9.49	8.81	41.33	90.33	96.67	57.33	86.67	92.33	69.44	8.08	6.24	67.67	93.67	96.00	67.67	93.33	97.00	79.02
	Dream Machine	9.13	8.72	41.33	90.33	96.33	57.67	86.33	94.67	69.11	7.85	6.51	69.33	93.67	95.33	71.00	93.00	95.67	79.06
Ours-Avg. (MASt3R)	DynamiCrafter	5.64	5.91	63.00	98.33	99.33	76.00	92.67	95.67	79.93	8.75	9.64	50.33	88.33	95.33	35.33	83.33	96.00	66.71
	CogVideoX	8.05	8.56	34.00	96.00	98.33	51.67	88.33	94.33	70.02	10.02	10.5	43.67	86.67	94.00	28.67	81.67	95.67	64.43
	Runway	7.25	7.11	55.67	92.33	98.00	65.00	88.67	95.67	74.68	9.05	9.21	44.00	87.00	95.67	34.00	84.67	96.67	68.11
	Dream Machine	7.30	7.59	59.67	92.00	96.67	64.33	86.33	94.67	73.06	9.37	9.29	43.67	89.33	95.33	38.00	86.33	95.67	68.39
Ours-Medoid (MASt3R)	DynamiCrafter	4.49	4.04	81.33	98.67	99.33	86.33	95.67	97.67	85.86	5.29	5.61	69.00	96.67	98.67	63.00	95.67	98.67	80.21
	CogVideoX	4.98	4.73	73.00	99.00	99.33	86.00	94.67	97.00	83.70	5.85	6.23	66.00	96.33	98.33	60.33	94.00	98.00	78.49
	Runway	4.17	4.01	81.67	99.00	99.33	87.33	96.00	97.33	86.79	5.28	5.20	72.67	96.33	98.67	69.00	97.00	98.67	81.63
	Dream Machine	4.30	4.21	80.67	99.00	99.33	85.33	94.67	97.00	85.88	5.66	5.45	70.00	97.33	98.33	70.00	96.00	98.33	81.42
Oracle	All Video Models	1.35	1.05	97.67	100.00	100.00	96.33	99.33	100.00	95.83	2.23	1.67	94.33	99.33	100.00	94.33	100.00	100.00	92.90

Table 4. **Camera pose estimation results on center-facing datasets (DL3DV-10K and NAVI).** MASt3R demonstrates significantly improved performance on these center-facing datasets compared to outward-facing ones. We evalute our method based on two pose estimators DUSt3R and MASt3R. Our method obtains comparable results on the DL3DV-10K dataset and slightly better performance on the NAVI dataset, demonstrating that using a video model does not hinder performance even when DUSt3R and MASt3R are already strong.

Yaw range	# Pairs	Pose estimator	Input data	MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			$AUC_{30°}$↑
						5°	15°	30°	5°	15°	30°	
[0°, 50°]	200	DUSt3R	Pair	11.33	22.50	76.00	83.50	89.00	43.50	67.50	78.00	60.20
		Ours	Dream Machine	**9.12**	**15.75**	**78.00**	**87.50**	**90.00**	**45.00**	**70.50**	**82.00**	**61.82**
		Oracle	Dream Machine	2.72	4.77	88.50	97.50	99.00	71.50	91.50	97.50	84.52
[0°, 25°]	100	DUSt3R	Pair	9.29	20.76	82.00	87.00	**91.00**	41.00	74.00	80.00	61.97
		Ours	Dream Machine	**8.41**	**15.30**	**85.00**	**89.00**	89.00	**44.00**	**75.00**	**82.00**	**62.43**
		Oracle	Dream Machine	2.12	4.77	91.00	99.00	99.00	74.00	93.00	96.00	85.20
[25°, 50°]	100	DUSt3R	Pair	13.36	24.25	70.00	80.00	87.00	**46.00**	61.00	76.00	58.43
		Ours	Dream Machine	**9.83**	**16.20**	**71.00**	**86.00**	**91.00**	**46.00**	**66.00**	**82.00**	**61.20**
		Oracle	Dream Machine	3.33	4.78	86.00	96.00	99.00	69.00	90.00	99.00	83.83

Table 5. **Camera pose estimation results on large overlapping pairs with yaw changes in the range [0°, 50°] on the ScanNet dataset.** Our method demonstrates improved performance over DUSt3R on input pairs alone, in scenarios with significant overlapping regions.

Yaw range	# Pairs	Pose estimator	Input data	MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			$AUC_{30°}$↑
						5°	15°	30°	5°	15°	30°	
[65°, 180°]	200	DUSt3R	Pair	**83.48**	58.93	**20.50**	28.50	31.50	19.00	26.00	31.50	20.88
		Ours	Dream Machine	83.94	**37.81**	18.50	**30.00**	**33.00**	**20.00**	**31.00**	**44.00**	**21.28**
		Oracle	Dream Machine	36.94	11.91	38.00	51.50	57.00	41.50	71.50	89.00	39.50
[65°, 110°]	95	DUSt3R	Pair	59.24	50.44	**31.58**	44.21	46.32	**25.26**	35.79	43.16	31.30
		Ours	Dream Machine	**56.35**	**33.16**	28.42	**48.42**	**50.53**	24.21	**38.95**	**51.58**	**32.53**
		Oracle	Dream Machine	15.98	11.30	60.00	75.79	77.89	46.32	70.53	89.47	56.11
[110°, 180°]	105	DUSt3R	Pair	**105.41**	66.61	**10.48**	**14.29**	**18.10**	13.33	17.14	20.95	**11.46**
		Ours	Dream Machine	108.89	**42.02**	9.52	13.33	17.14	**16.19**	**23.81**	**37.14**	11.11
		Oracle	Dream Machine	55.91	12.46	18.10	29.52	38.10	37.14	72.38	88.57	24.48

Table 6. **Camera pose estimation results on non-overlapping pairs with yaw changes in the range [65°, 180°] on the ScanNet dataset.** The performance of DUSt3R and our method significantly drops in this challenging non-overlapping scenario. While our method obtains better translation estimation, it exhibits slightly worse rotation estimation compared to DUSt3R.

Yaw range	# Pairs	Pose estimator	Input data	MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			$AUC_{30°}$↑
						5°	15°	30°	5°	15°	30°	
[0°, 50°]	200	DUSt3R	Pair	4.28	11.04	79.00	95.50	98.00	49.00	89.00	93.00	73.60
		Ours	Dream Machine	**3.24**	**8.16**	**81.50**	**97.50**	**99.50**	**49.00**	**90.50**	**96.00**	**76.17**
		Oracle	Dream Machine	1.68	3.16	93.00	100.00	100.00	85.50	98.00	99.00	89.80
[0°, 25°]	67	DUSt3R	Pair	2.78	10.87	91.73	96.99	98.50	42.86	87.22	93.23	73.31
		Ours	Dream Machine	**1.87**	**8.48**	**94.74**	**99.25**	**100.00**	**45.86**	**88.72**	**96.24**	**76.17**
		Oracle	Dream Machine	1.04	3.62	98.50	100.00	100.00	81.95	97.74	98.50	89.82
[25°, 50°]	133	DUSt3R	Pair	7.25	11.37	53.73	92.54	97.01	**61.19**	92.54	92.54	74.18
		Ours	Dream Machine	**5.96**	**7.54**	**55.22**	**94.03**	**98.51**	55.22	**94.03**	**95.52**	**76.17**
		Oracle	Dream Machine	2.95	2.26	82.09	100.00	100.00	92.54	98.51	100.00	89.75

Table 7. **Camera pose estimation results on large overlapping pairs with yaw changes in the range [0°, 50°] on DL3DV-10K.** DUSt3R already performs strongly on this center-facing dataset, and Our method still achieves slight improvements over DUSt3R.

Yaw range	# Pairs	Pose estimator	Input data	MRE↓	MTE↓	R_{acc} ↑			t_{acc} ↑			$AUC_{30°}$↑
						5°	15°	30°	5°	15°	30°	
[90°, 180°]	200	DUSt3R	Pair	19.20	15.00	**32.50**	79.00	85.50	52.50	86.00	87.00	65.07
		Ours	Dream Machine	**16.06**	**9.62**	31.50	**82.00**	**89.50**	**53.50**	**88.50**	**91.50**	**66.37**
		Oracle	Dream Machine	8.18	3.77	68.00	92.00	95.00	86.00	96.50	97.50	82.18
[90°, 110°]	158	DUSt3R	Pair	17.81	14.73	**30.38**	79.11	86.71	48.73	87.34	87.97	64.64
		Ours	Dream Machine	**14.66**	**9.17**	28.48	**82.28**	**91.14**	**50.63**	**89.87**	**93.04**	**66.20**
		Oracle	Dream Machine	6.35	3.11	67.09	93.67	96.84	86.08	97.47	98.73	83.14
[110°, 180°]	42	DUSt3R	Pair	24.42	15.99	40.48	78.57	80.95	**66.67**	80.95	83.33	66.67
		Ours	Dream Machine	**21.31**	**11.30**	**42.86**	**80.95**	**83.33**	64.29	**83.33**	**85.71**	**66.98**
		Oracle	Dream Machine	15.05	6.26	71.43	85.71	88.10	85.71	92.86	92.86	78.57

Table 8. **Camera pose estimation results on pairs with large yaw changes in the range [90°,180°] on DL3DV-10K.** The center-facing nature of this dataset ensures overlapping regions despite significant viewpoint changes, enabling DUSt3R to produce reasonable estimations. Our method obtains better pose estimation results over DUSt3R.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, 2017. 1

[2] Zhengcong Fei. CogVideoX-Interpolation: Keyframe Interpolation with CogVideoX, 2024. https://github.com/feizc/CogvideX-Interpolation [Accessed: (October 2024)]. 1, 2

[3] 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*, 2024. 1

[4] 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. 1, 2