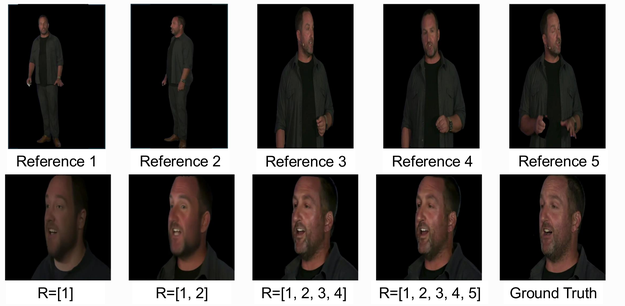
Free-viewpoint Human Animation with Pose-correlated Reference Selection

Fa-Ting Hong[1,2] Zhan Xu[2] Haiyang Liu[2] Qinjie Lin[3] Luchuan Song[2]
Zhixin Shu[2] Yang Zhou[2] Duygu Ceylan[2] Dan Xu[1]

[1]HKUST [2]Adobe Research [3]Northwestern University

Project

CVPR Nashville JUNE 11-15, 2025



Comparison between results with different number of reference images
Reference 3, 4 and 5 provide different facial details

Previous Works:

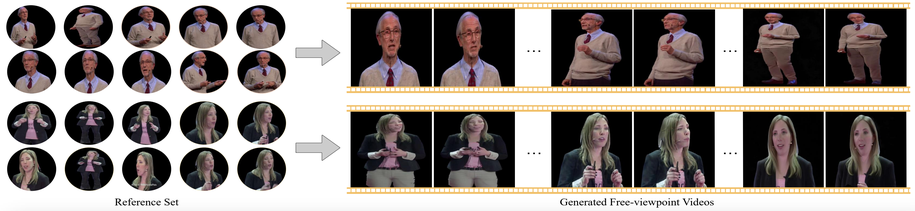
- Single-view references limit realism due to fixed viewpoints.
- They lack sufficient detail for novel view synthesis.
- Large viewpoint changes reveal unseen regions that are hard to synthesize

Pose-correlated Reference Selection Network:

- Multiple reference images.
- Adaptive Reference Selection: Remove useless token.
- Pose Correlation Mechanism: Identify the informative token.
- New Dataset

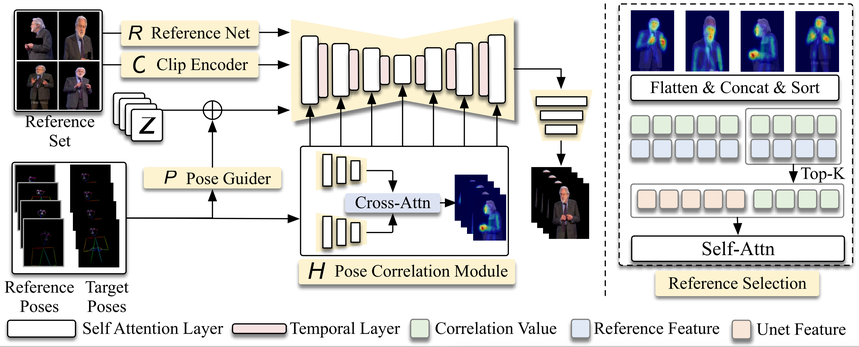
Dataset	Dataset Statistics					Identity Split	
	Identities	Clips	Total Duration (hrs)	Camera Distance Changes	Viewpoint Changes	Train IDs	Test IDs
DyMVHumans [38]	33	1,964	5.376	✗	✓	30	3
MSTed (Ours)	1,084	15,260	29.923	✓	✓	1,000	84

Table 1. Comparison between our proposed MSTed dataset and the existing DyMVHumans dataset [38]. MSTed offers a significantly larger scale with 1,084 identities and 15,260 clips spanning approximately 30 hours of video. Furthermore, MSTed includes diverse real-world variations in camera distances and viewpoints, making it more representative of in-the-wild scenarios.



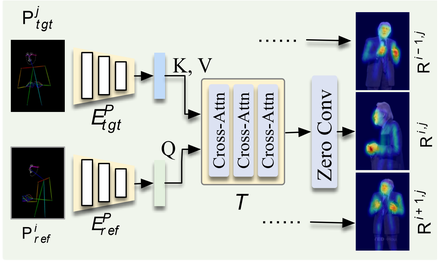
Framework: Multiple Reference + Reference Selection

- **Input:** Reference set, Target pose Sequences
- **Output:** A new video that the character moves aligned with the input target poses.



Pose Correlation Mechanism

- Paired target pose and reference pose
- N Target pose × M Reference image
  → M × N correlation map



Reference Selection

- Flatten reference feature & correlation map

$$\mathbf{r}_j = \mathbf{r}_{1,j}||\mathbf{r}_{2,j}||\dots||\mathbf{r}_{N,j},$$
$$\mathbf{f}_l = \mathbf{f}_l^1||\mathbf{f}_l^2||\dots||\mathbf{f}_l^N,$$

- Sort the correlation value
- Select Top-K

$$\mathbf{f}_l^{K_l} = \{f_l^i \mid i \in \text{argsort}(\mathbf{r}_j)[:K_l]\},$$
$$\mathbf{r}_j^{K_l} = \{r_j^i \mid i \in \text{argsort}(\mathbf{r}_j)[:K_l]\},$$

- Compensated Reference Feature Sampling

$$\mathbf{f}_{\text{random},l}^j = \mathcal{S}_{uni}(\mathbf{f}_l, K_l),$$

Quantitative Results

Model	$\mathcal{L}_1 \downarrow$	PSNR ↑	LPIPS ↓	MOVIE ↓	FVD ↓
MagicAnimate [34]	154.02	27.92	0.5984	119.33	35.08
AnimateAnyone [9]	113.69	29.38	0.5458	94.93	33.10
Champ [40]	81.69	30.87	0.4618	67.84	25.68
Ours(R=1)	**78.91**	**32.18**	**0.2045**	**56.53**	**20.88**
Ours(R=2)	**74.20**	**32.49**	**0.1869**	**55.60**	**7.044**

Table 2. Quantitative results comparision on the MSTed dataset. The results show in table indicate that our method perform better. And the results can be better with the increase of reference number. "R=2" means our model uses 2 reference images.

Model	$\mathcal{L}_1 \downarrow$	PSNR ↑	LPIPS ↓	MOVIE ↓	FVD ↓
MagicAnimate [34]	132.64	27.87	0.6583	109.74	51.433
AnimateAnyone [9]	56.30	33.58	0.2179	42.56	12.300
Champ [40]	61.65	30.87	0.4388	63.21	45.762
Ours(R=1)	**56.47**	**34.22**	**0.1660**	**39.70**	**9.047**
Ours(R=2)	**50.37**	**34.78**	**0.1435**	**39.21**	**8.782**
Ours(R=5)	**49.43**	**34.84**	**0.1400**	**39.10**	**7.656**
Ours(R=10)	**35.27**	**35.02**	**0.1383**	**34.35**	**5.459**

Table 3. Quantitative results on DyMVHumans dataset [38]. Our model can accept 10 reference images and get good results.

Qualitative Results