Large Displacement Motion Transfer with Unsupervised Anytime Interpolation

Guixiang Wang[1], Jianjun Li[2]

[1]School of Computer Science and Engineering, Hangzhou Dianzi University
[2]School of Information Science and Technology, Hangzhou Normal University

ICML 2025

What is Motion Transfer?

- Aims to transfer pose from a driving video to an object in a source image, animating the source object.
- Hot research topic with applications in film animation, game production, and face exchange.
- Goal: Animate a still image by transferring pose from a driving video to generate a video with the same pose.
- Key challenges:
  - Accurately transfer motion patterns.
  - Maintain identity consistency.
  - Struggle with large displacement motions.

Problem Statement & Existing Methods

- **Problem:** Current unsupervised methods struggle with large displacement motions.
- **Existing Methods (Limitations):**
  - Supervised methods: Rely on prior knowledge (landmarks, 3D models), often fail with out-of-training data.
  - Unsupervised methods:
    - **FOMM, MRAA:** Use local linear affine transformations, struggle with non-linear object motion.
    - **TPSMM:** Uses Thin-plate splines, but keypoint detection is often inaccurate.
    - **CPABMM:** Uses continuous piecewise affine transformation, limited in finer motions, causing artifacts.
    - **CoP:** Based on chain-of-pose, difficult to obtain pose chains with different identity information or large displacement.

Proposed Method: Unsupervised Anytime Interpolation

- **Core Idea:** Decompose large displacement motion into many small displacement motions by inserting intermediate images.
- **Keypoint-based anytime interpolation:**
  - Estimates keypoint information of interpolated images based on source and driving image keypoints.
  - Assumes linear motion of each keypoint.
  - Uses a dense motion network to predict motion low from source to interpolated images.
  - Generates interpolated images using an image generation network.

Overall Architecture

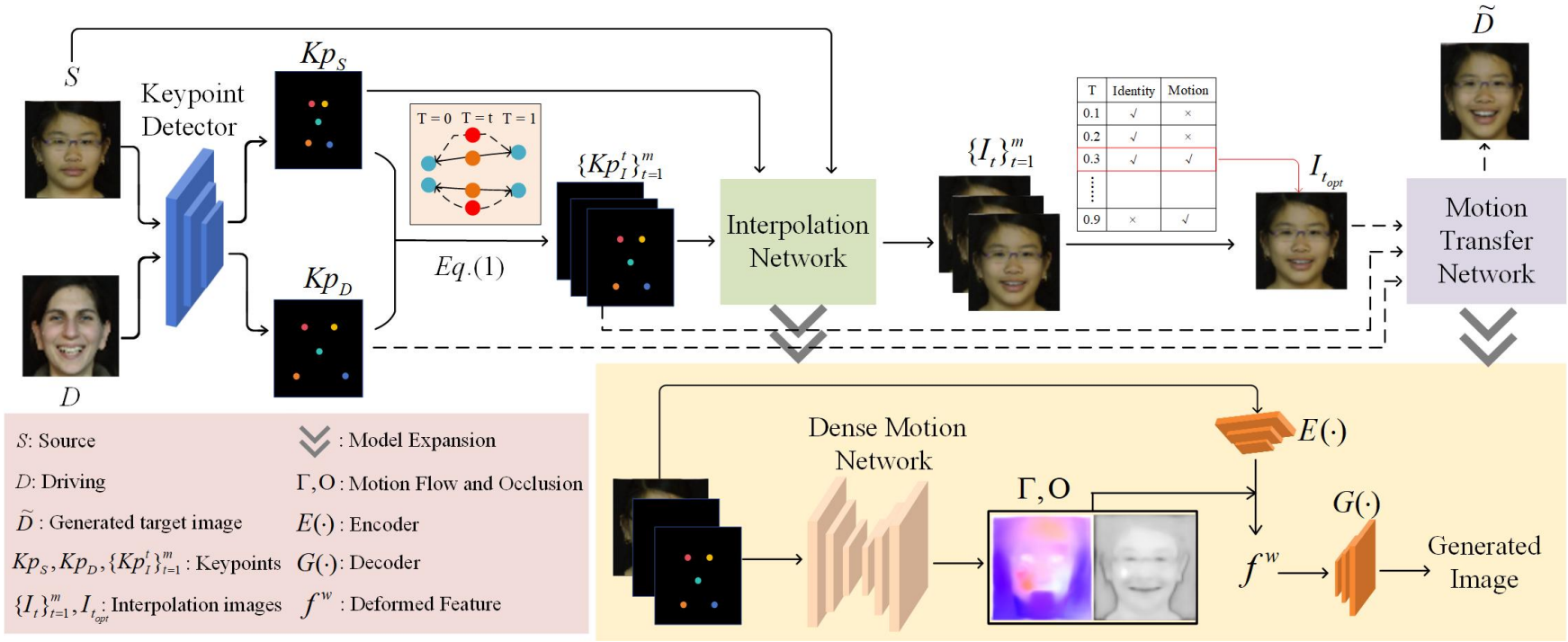

Figure: Overall framework of the proposed method.

Motion Transfer with Optimal Interpolated Image

- Traditional methods struggle with large displacement motion directly from source to driving.
- The proposed method uses an optimal interpolated image (I_{topt}) to replace the source image for motion transfer.
- **Optimal Interpolation Selector:** Selects an image from a series of interpolated images that satisfies two requirements:
  - Maintains identity information integrity.
  - Has small motion to the driving image.
- Small displacement motion estimation from interpolated to driving image is more accurate.

Example of Anytime Interpolation

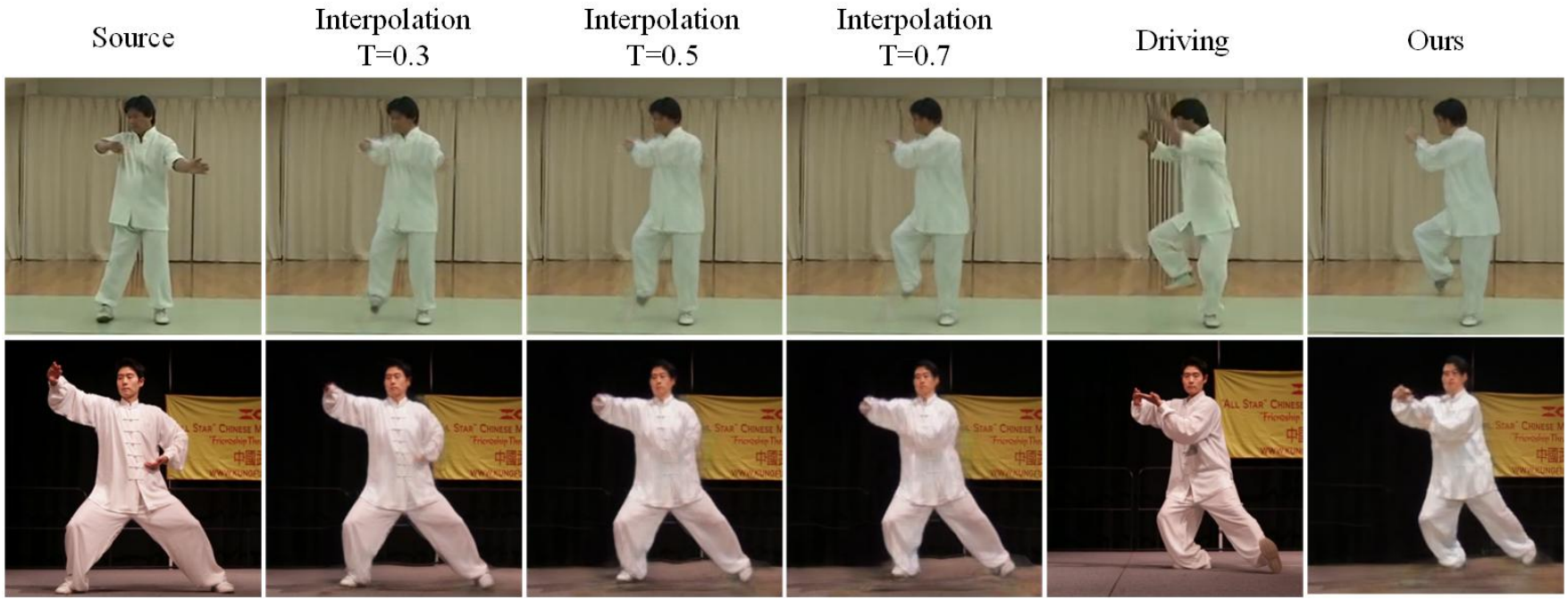

Figure: Interpolation results generated at different moments under the same identity

Example of Anytime Interpolation

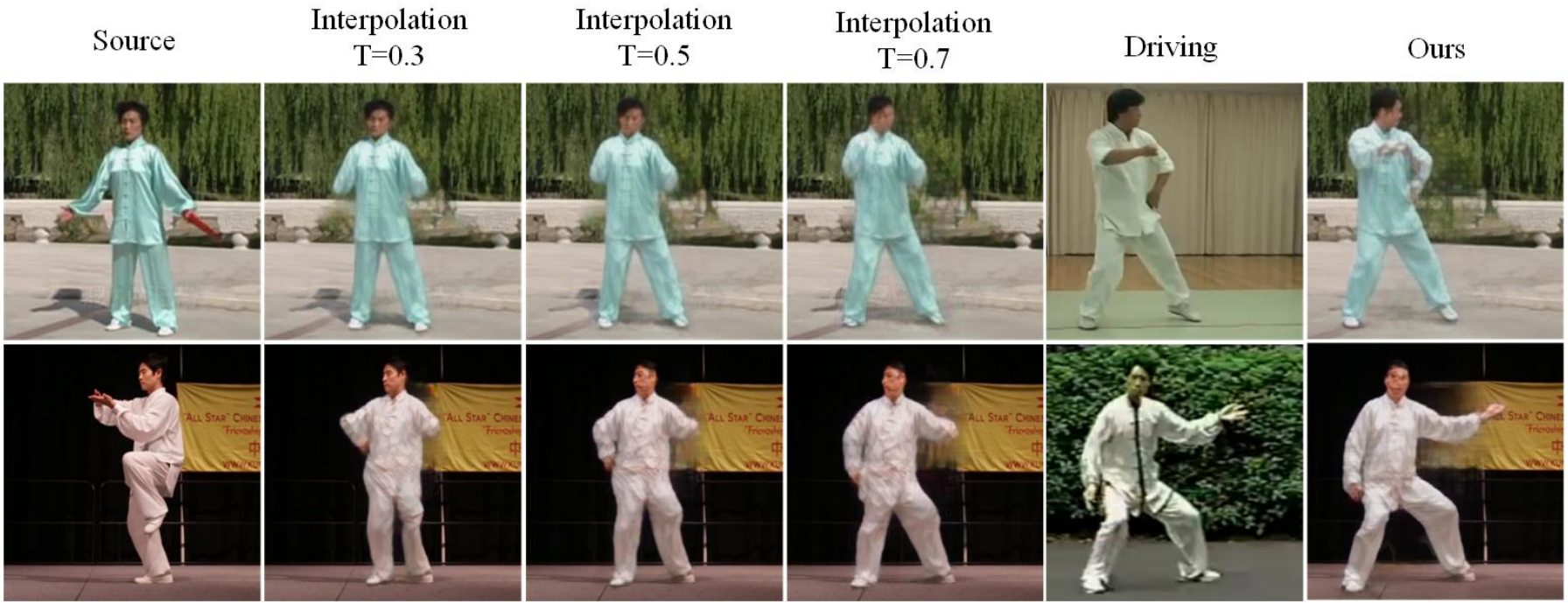

Figure: Interpolation results generated at different moments under the different identity.

Proposed Method: Bidirectional Training Strategy

- Proposed because there are no real images as labels for interpolation, ensuring meaningful optimal interpolated images.
- Consists of two pipelines: Source to Driving (S to D) motion transfer and Driving to Source (D to S) motion transfer.
- Adds constraints to ensure consistency between optimal interpolated images generated by both pipelines.

Bidirectional Training Strategy Diagram

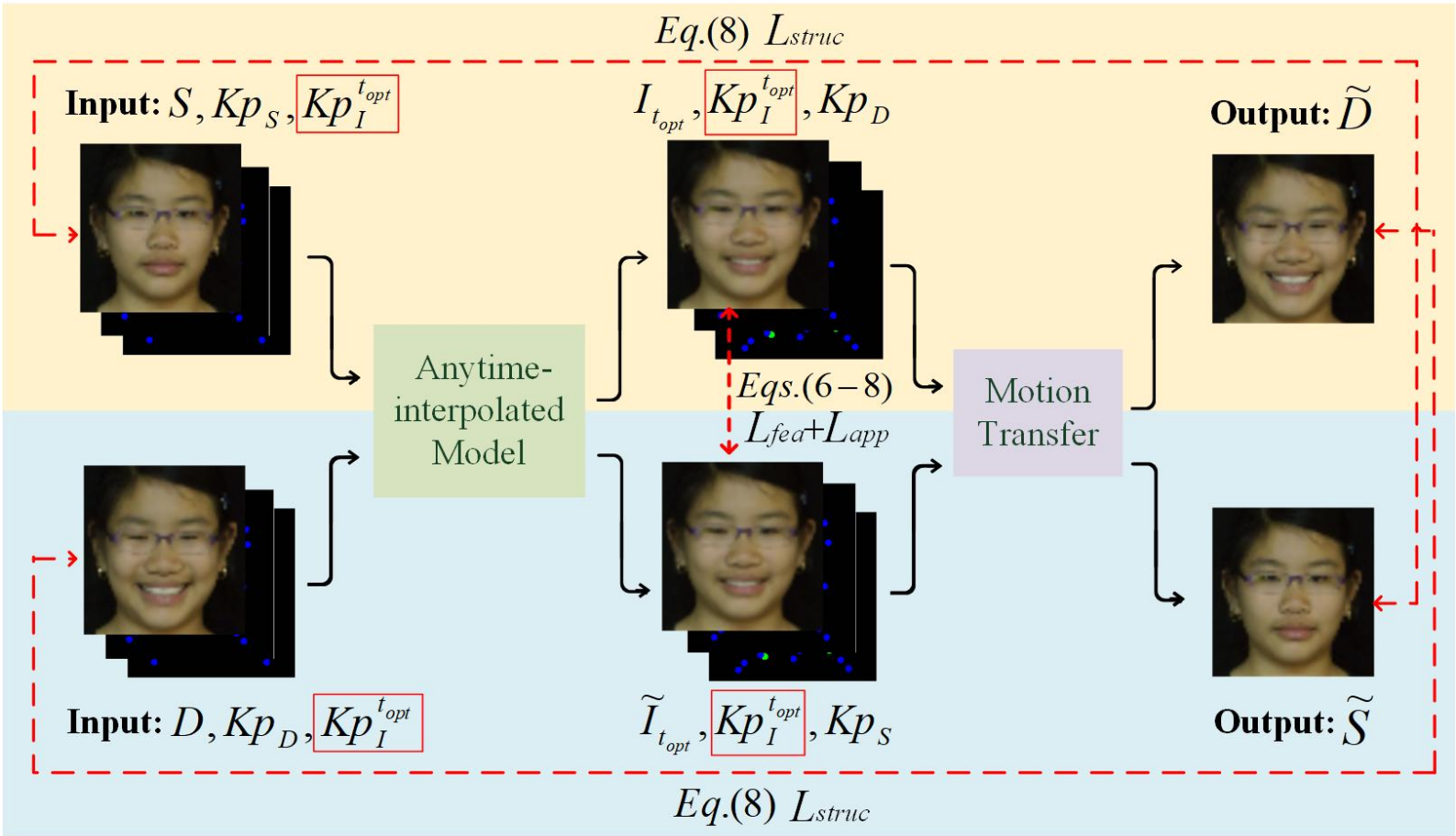

Figure: Bidirectional training strategy for consistency.

Proposed Method: Loss Functions

- Using pre-trained Vision Transformer (ViT) for constraints to encourage high-quality images.
- **Reconstruction loss** (L_{rec}): Ensures generated images are realistic (using VGG-19 network).
- **Feature consistency loss** (L_{fea}): Encourages optimal interpolated images to be consistent.
- **Appearance consistency loss** (L_{app}): Encourages consistent appearance of optimal interpolated images using ViT.
- **Structural consistency loss** (L_{struc}): Ensures structural consistency of optimal interpolated images, source, and driving images using ViT.

Experimental Results & Analysis

- **Datasets:** UvA-Nemo, Fashion, Tai-Chi-HD, and TedTalks (faces and human bodies).
- **Metrics:** L1, Average Keypoint Distance (AKD), Missing Keypoint Rate (MKR), Average Euclidean Distance (AED).
- **Video Reconstruction:**
  - Significant improvements in motion-related metrics.
  - Our model obtains a pose closer to the real image.
  - Qualitative results show more accurate large displacement motion estimation.
  - Note: May lead to some loss of appearance information and blurring in interpolated images, especially with backgrounds.

Quantitative Comparison: Video Reconstruction

	Tai-Chi-HD			Fashion			UvA-Nemo			TedTalks		
	L_1	(AKD, MKR)	AED	L_1	(AKD, MKR)	AED	L_1	AKD	AED	L_1	(AKD, MKR)	AED
X2Face	0.080	(17.65, 0.109)	0.270	-	-	-	0.031	3.539	0.221	-	-	-
FOMM	0.057	(6.65, 0.036)	0.172	0.013	(1.131, 0.006)	0.059	0.021	1.408	0.067	0.033	(7.07, 0.014)	0.163
MRAA	0.048	(5.41, 0.025)	0.149	-	-	-	0.017	1.323	0.060	0.026	(3.75, 0.007)	0.114
DAM	0.044	(4.79, 0.021)	0.146	**0.011**	(1.041, **0.004**)	**0.054**	-	-	-	-	-	-
MTIA	0.045	(4.67, 0.021)	0.148	-	-	-	-	-	-	0.026	(3.46, 0.007)	0.113
TPSMM	0.045	(4.57, 0.018)	0.151	**0.011**	(0.845, 0.005)	0.056	0.011	1.177	**0.050**	0.027	(3.39, 0.007)	0.124
CPABMM	**0.041**	(4.61, 0.021)	**0.117**	-	-	-	-	-	-	**0.022**	(3.21, 0.008)	**0.085**
Ours	0.047	(4.21, 0.014)	0.157	**0.011**	(0.800, **0.004**)	0.056	**0.010**	1.155	0.051	0.028	**(3.15, 0.005)**	0.136
Ours-V2	0.046	**(3.67, 0.013)**	0.149	**0.011**	**(0.771, 0.004)**	0.058	**0.010**	**0.853**	**0.050**	-	-	-

Table: Quantitative comparison of video reconstruction task on four diferent datasets.

Qualitative Comparison

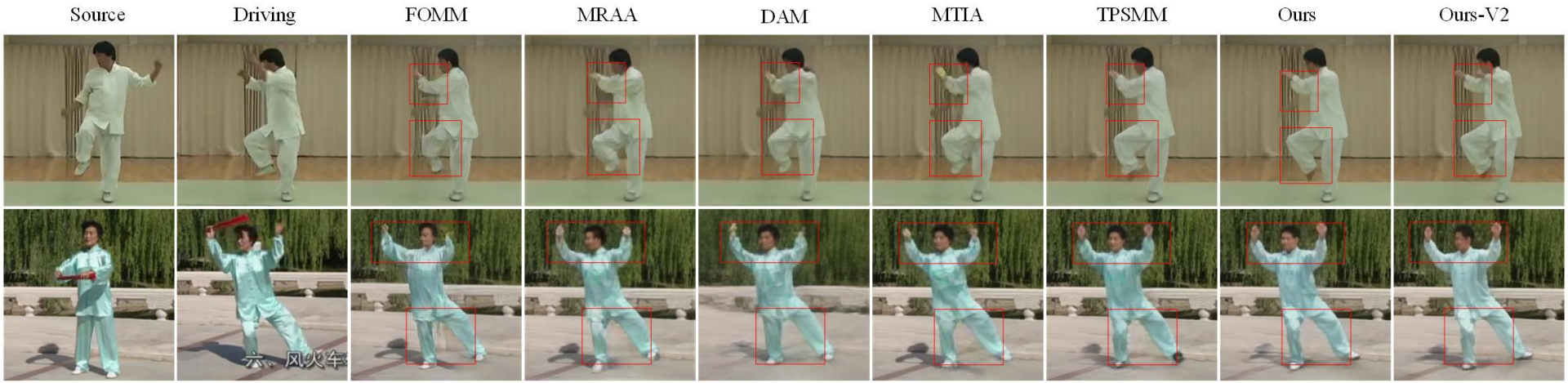

Figure: Some bad comparison methods cases on the Tai-chi-HD dataset, while our method shows high quality on video reconstruction task.

Qualitative Comparison

Figure: Qualitative comparison with comparison methods on image animation task TaiChiHD (above), Fashion (below).

Ablation Study Results

	L1	(AKD, MKR)	AED
TPSMM	**0.045**	(4.57,0.018)	**0.151**
+TPSMM, INT (Model1)	0.053	(5.56,0.026)	0.169
+Model1, BT (Model2)	0.048	(4.31,0.016)	0.158
+Model2, L_{app}, L_{struc} (Ours)	0.047	**(4.21,0.014)**	0.157

Table: Ablating the key components of the proposed method. INT is denoted as adding an unsupervised anytime interpolation method, and BT is denoted as adding a bidirectional training strategy.

Ablation Studies

- **Tai-Chi-HD dataset:**
  - Adding unsupervised anytime interpolation alone (Model1) decreases performance without constraints, leading to poor interpolated image generation.
  - Adding bidirectional training strategy (Model2) improves performance by constraining interpolated images.
  - Full model (Ours) with appearance and structure consistency losses shows optimal results in motion-related metrics.

Method Extension (Ours-V2)

- Extended to multi-view tasks using source image and optimally interpolated images for fused deformation features.
- Achieves significant improvements on most datasets, especially in AKD.
- Significantly improves appearance of generated images.

Conclusion

- Proposed a large displacement motion transfer model based on unsupervised anytime interpolation.
- Decomposes large displacement motion into small displacement motions, improving motion estimation accuracy.
- Introduced a bidirectional training strategy with ViT-based constraints (appearance and structure consistency) to narrow the solution space and improve interpolated image quality.
- Experiments show significant improvements in motion-related metrics compared to state-of-the-art methods.