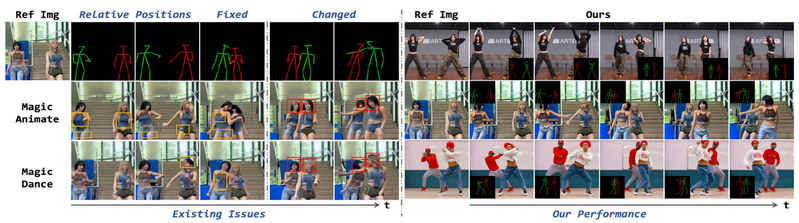
EverybodyDance: Bipartite Graph–Based Identity Correspondence for Multi-Character Animation

Haotian Ling, Zequn Chen, Qiuying Chen
Donglin Di, Yongjia Ma, Hao Li, Chen Wei
Zhulin Tao, Xun Yang

理想汽车
Li Auto

NEURAL INFORMATION PROCESSING SYSTEMS

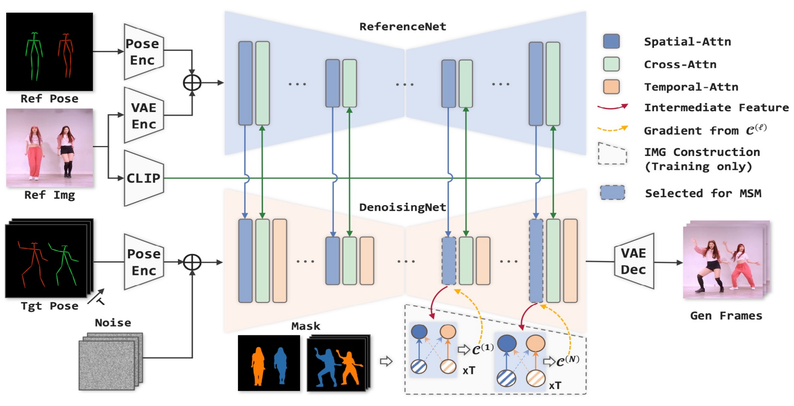
EverybodyDance: Enforcing Identity Correspondence in Multi-Character Animation



Maintaining Identity Correspondence (IC) in pose-driven, multi-character animation is a non-trivial challenge, especially when characters swap positions. We introduce EverybodyDance, a systematic solution that models characters as two node sets in a weighted bipartite graph, which we term the Identity Matching Graph (IMG). We propose a novel Mask–Query Attention (MQA) to compute edge weights (affinity) between character pairs. Our key insight is to formalize IC correctness as a graph structural metric and optimize it directly during training. Supported by targeted strategies like identity-embedded guidance, our method substantially outperforms SOTA baselines in both IC and visual fidelity, evaluated on our new Identity Correspondence Evaluation (ICE) benchmark.

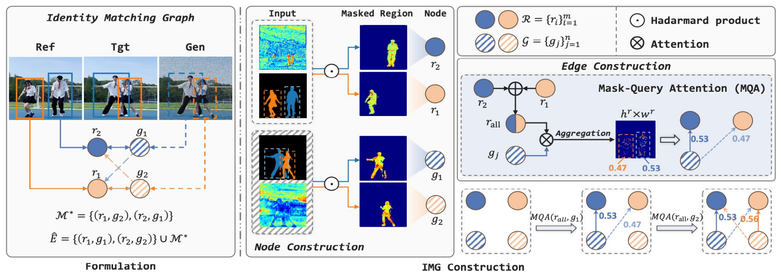
The left panel highlights the challenges of extending existing methods to multi-character scenarios. The yellow box indicates feature interference between characters, while the red box marks identity mismatches.

Pipeline of EverybodyDance



EverybodyDance adapts the Animate Anyone architecture by incorporating **Identity-Embedded Guidance (IEG)**, using color-coded skeletons for both the reference (Ref Pose) and target (Tgt Pose) inputs to uniquely mark each character. The **ReferenceNet** fuses appearance information from the reference image (Ref Img) with the corresponding reference IEG (Ref Pose), binding character identity to create identity-aware features. These features are then injected into the **DenoisingNet** via Spatial-Attention , guiding it to generate frames that follow the target IEG (Tgt Pose) while maintaining the correct identity. During training only, intermediate features and masks are used to construct the **Identity Matching Graph (IMG)**, which enforces correct IC through a dedicated matching loss.

Illustration of Identity Matching Graph



We formalize the problem by modeling reference characters (Ref) and generated characters (Gen) as the two node sets of a weighted bipartite graph. The ground-truth (Tgt) frame defines the correct, or target, edges, which must be correctly identified even when characters swap positions. Graph nodes are constructed by element-wise multiplication of intermediate U-Net features with their corresponding segmentation masks. To efficiently compute the edge weights (affinities), we propose Mask-Query Attention (MQA). Each generated node then acts as a query to this aggregated representation, allowing a single attention operation to efficiently compute the affinity scores between that one generated character and all original reference characters.

Target Strategy

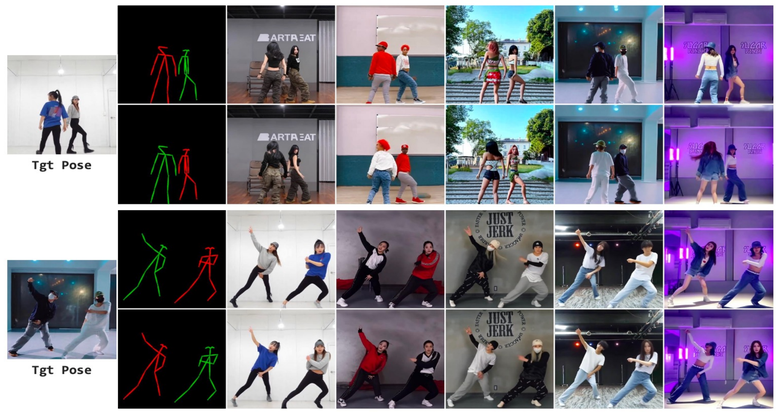
To enhance the robustness of our framework, we introduce two targeted strategies that work synergistically with IMG. First, Multi-Scale Matching (MSM) constructs and optimizes the IMG loss at multiple layers of the U-Net. This enforces correct IC across the entire feature hierarchy. Second, Pre-Classified Sampling (PCS) addresses the long-tail data distribution where challenging samples, such as character position swaps, are rare. PCS ensures that the model receives sufficient emphasis on these critical, infrequent scenarios during training.

Quantitive Evaluation

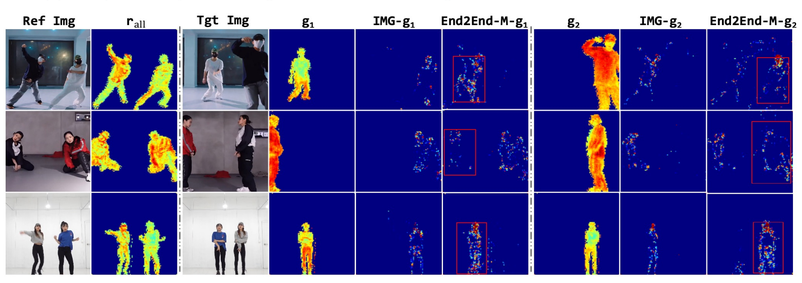
Method	Frame Quality					Video Quality	
	SSIM↑	PSNR*↑	LPIPS↓	L1↓	FID↓	FID-VID↓	FVD↓
AnimateAnyone [9]	0.616	14.97	0.339	5.16E-05	59.19	32.057	364.85
AnimateAnyone* [9]	0.596	14.67	0.342	5.35E-05	54.71	31.274	358.31
MimicMotion [18]	0.621	15.00	0.338	5.48E-05	60.77	26.490	381.69
MagicDance [12]	0.508	13.81	0.424	1.33E-04	53.30	47.127	471.71
MagicAnimate [10]	0.614	13.95	0.369	6.36E-05	76.28	42.257	521.67
UniAnimate [17]	0.623	15.66	0.328	3.41E-05	44.38	26.696	295.56
EverybodyDance	**0.654**	**16.93**	**0.304**	**2.86E-05**	**40.19**	**23.584**	**225.06**

Quantitative analysis on the ICE benchmark demonstrates that EverybodyDance substantially outperforms all state-of-the-art baselines across both frame and video quality metrics. To ensure these gains stem from our proposed framework rather than dataset-specific biases, we fine-tuned the Animate Anyone backbone (denoted AnimateAnyone*) on our custom dataset . The results clearly show that AnimateAnyone* still fails to match the performance of EverybodyDance, confirming that our method's significant improvements are attributable to our novel architecture and targeted strategies, not merely the training data.

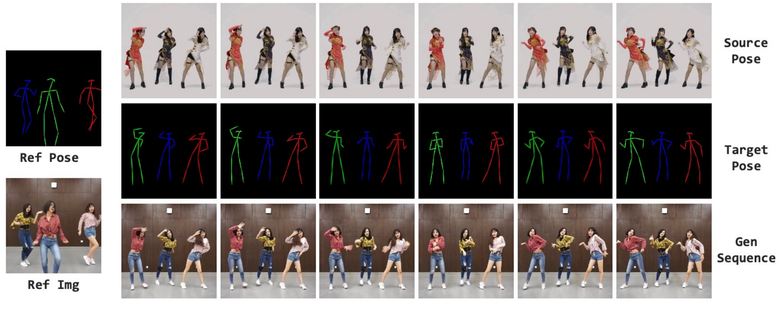
Visualization



We apply a motion template from a source video to animate diverse reference images. Crucially, we test the model's controllability by reassigning character positions, which is achieved by simply reordering the color-coded identities in the target IEG. As the results show, the model correctly renders the sequence with the characters' relative positions swapped, precisely following the new IEG assignment.



We display the affinity scores of each generated character (e.g., the person on the left, the person on the right) over the aggregated attention map. The attention maps from our method (labeled 'IMG-...') are sharply localized. This demonstrates that each generated character correctly attends to its corresponding reference identity, even during challenging position swaps (as seen in the middle row). In contrast, the baseline's maps (labeled 'End2End-M-...') exhibit feature contamination, with attention incorrectly leaking to the non-corresponding character (highlighted in red boxes).



Furthermore, our method demonstrates robust generalizability and is readily scalable to scenarios involving more characters.