Dynamic Try-On: Taming Video Virtual Try-on with Dynamic Attention Mechanism

Jun Zheng[1], Jing Wang[1], Fuwei Zhao[2], Xujie Zhang[1], Xiaodan Liang[1,*]

[1]Shenzhen Campus of Sun Yat-sen University *[2]ByteDance*

Background

Core Task: Video Virtual Try-On aims to transfer a garment onto a person in a video, preserving their identity and motion.

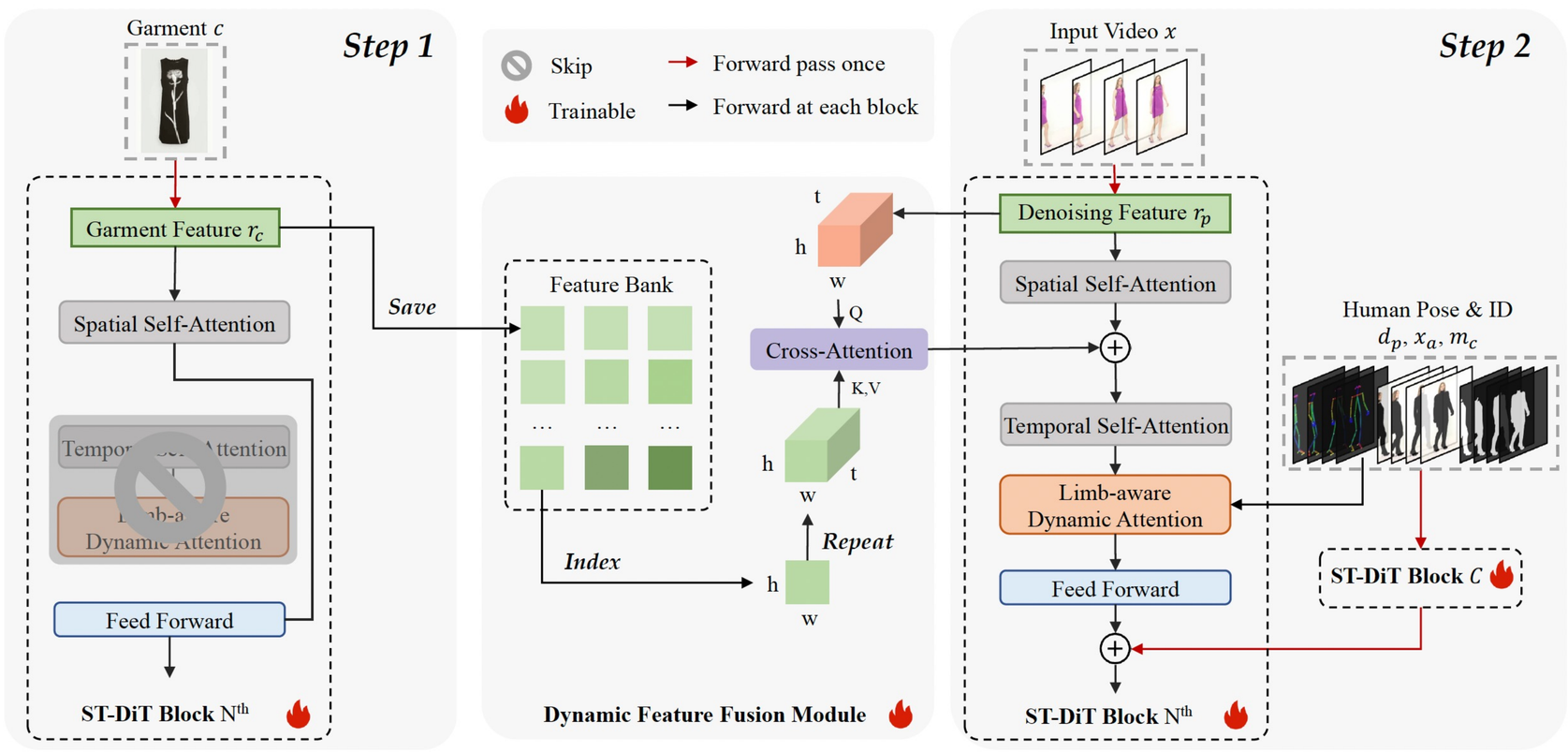
Key Challenges: Existing methods struggle to generate temporally coherent results for ***complex human motions*** while preserving ***high-fidelity clothing details***.

Proposed Solution

Dynamic Feature Fusion Module (DFFM):
A lightweight module that repurposes the main DiT backbone to encode garment features.

Limb-aware Dynamic Attention Module (LDAM):
A novel attention layer that leverages human pose information to explicitly model the motion of individual limbs across frames.

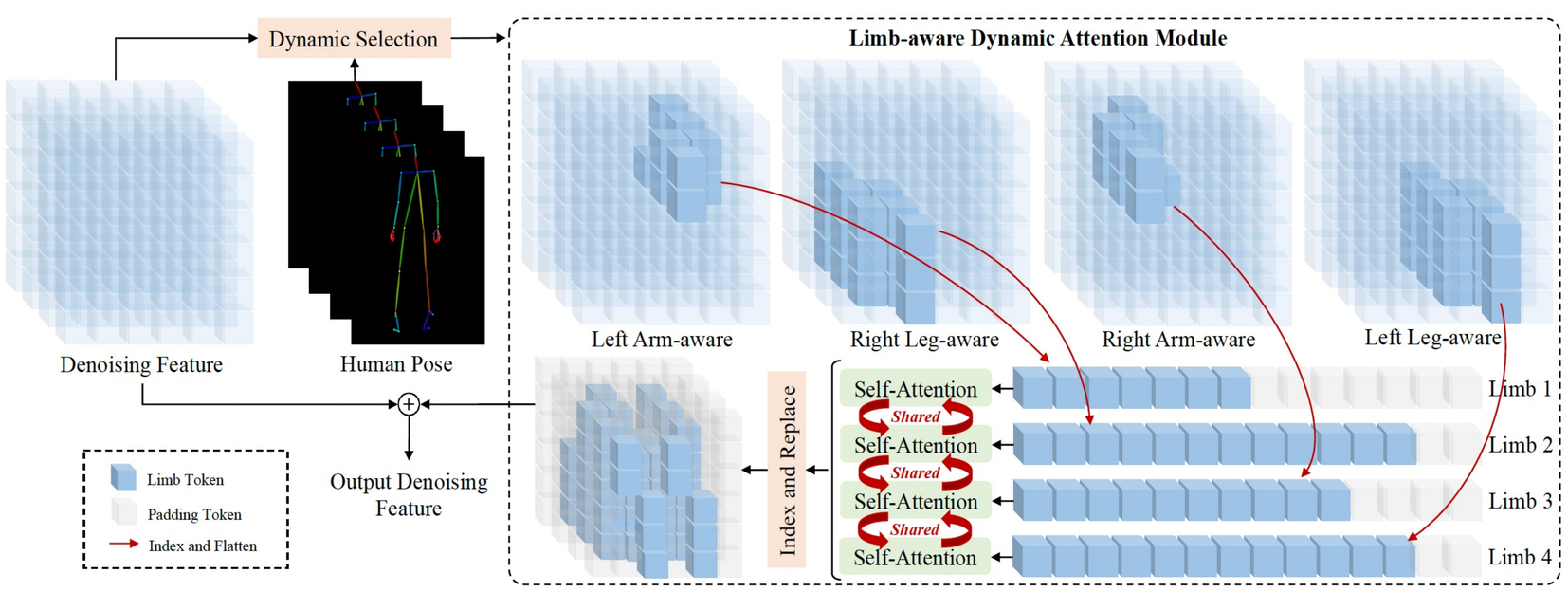
Method



Step 1: Garment Feature Extraction:
A single garment image is passed through the DiT backbone once. Its intermediate features are extracted and stored in a ***Feature Bank***.

Step 2: Dynamic Try-On Generation:
During the denoising process, the model:

1. Injects garment details using our efficient ***DFFM for high-fidelity texture***.
2. Utilizes the human pose sequence to guide the ***LDAM for superior temporal consistency***, ensuring smooth limb movements.



Limb-aware Dynamic Attention Module
LDAM dynamically selects and groups tokens corresponding to human limbs (e.g., left arm, right leg).

Shared self-attention is then applied within each group, ensuring that each limb is rendered consistently throughout the video sequence. This is critical for challenging motions.

Evaluation Results

Method	SSIM↑	LPIPS↓	VFID↓	FVD↓
PBAFN [9]	0.870	0.157	4.516	-
StableVITON [19]	0.914	0.132	6.291	220.05
OOTDiffusion [3]	0.863	0.154	7.852	205.03
FW-GAN [5]	0.675	0.283	8.019	-
ClothFormer [16]	0.921	0.081	3.967	-
Tunnel Try-on [45]	0.913	**0.054**	3.345	-
StableVITON + MA	0.888	0.145	3.655	66.24
OOTDiffusion + MA	0.851	0.159	4.465	89.17
ViViD [8]	0.913	0.133	2.961	66.14
Dynamic Try-On	**0.924**	0.098	**2.246**	**57.49**

Performance on VVT Dataset

Method	20 blocks	28 blocks	36 blocks	44 blocks
w/o DFFM	42.6G	58.3G	74.0G	OOM
w/ DFFM	35.4G	48.5G	61.3G	74.2G

Computational Efficiency: Our DFFM reduces training memory by ***~17%*** (58.3G → 48.5G for a 28-block model) compared to standard encoder-based approaches.

Garment Preservation Paradigm		Additional Attention Layers			Training Memory Cost of Different Resolutions		Evaluation Metrics			
w/o DFFM	w/ DFFM	None	3D Full Attention	LDAM	192 × 256	384 × 512	SSIM↑	LPIPS↓	VFID↓	FVD↓
✓		✓			58.3G	OOM	0.918	**0.092**	2.493	63.53
	✓	✓			48.5G	73.8G	0.915	0.104	2.487	66.25
	✓		✓		78.6G	OOM	0.918	0.105	2.451	60.73
	✓			✓	57.6G	79.3G	**0.924**	0.098	**2.246**	**57.49**

Ablation Study: The LDAM module improves all video metrics, demonstrating higher performance and efficiency than a standard (and more costly) 3D Full Attention layer.

See Dynamic Try-On in Action!
Scan for video demos and project page