W-DiT: Structure-Guided High-Frequency Propagation for Fashion Video Generation

Anonymous authors
Paper under double-blind review

Abstract

Generating fashion videos from reference images requires garment details to remain consistent under changing poses and cloth deformation. High-frequency product details, such as logos and fabric textures, should not be repeatedly regenerated across frames, as their appearance can progressively drift during motion. Our key insight is to generate low-frequency dynamics while propagating high-frequency product information directly from reference images along the garment's motion trajectories. Guided by this principle, we propose **W-DiT**, a video diffusion framework that couples generated garment motion with structure-guided detail propagation. First, *frequency disentanglement* separates low-frequency structure from high-frequency appearance and encodes reference details through a dedicated pathway that bypasses VAE compression. Second, *structure-guided alignment* matches low-frequency reference and target features to establish soft spatial correspondences that track the garment's evolving pose and deformation. Third, *high-frequency propagation* uses these correspondences to sample details directly from the fixed reference feature bank and inject them into each target frame, transporting product appearance along the motion trajectories. To support this task, we curate **Fashion-Video** from authorized Xiaohongshu images, comprising approximately 10,000 synthesized fashion videos and 30,000 associated images with textual conditioning. Our experiments show that W-DiT preserves fine-grained garment details across changing poses while maintaining motion quality.

1 Introduction

Fashion video generation aims to transform static reference images and textual instructions into dynamic product showcases. Beyond plausible motion, this task requires **consistent garment details**: a logo must preserve its shape, a printed pattern must remain attached to the same garment region, and a fabric texture must follow cloth deformation. Small changes to these attributes can alter the perceived product even when the video remains visually realistic. Preserving garment details throughout motion is therefore central to reference-guided fashion video synthesis.

Recent image-conditioned video generators combine visual conditioning with powerful temporal generative models Zhang et al. (2023); Hu (2024); Wan et al. (2025). VAE compression can attenuate fine spatial information before it reaches the video backbone Medi et al. (2025); Lai et al. (2025). Even when additional appearance features are provided, a generic conditioning mechanism leaves the model to infer both which reference details to use and where to reproduce them in each frame. Under articulated motion, folds, and changing viewpoints, the same pattern may consequently be reconstructed with different shapes or positions. Preserving appearance information at the input thus needs to be accompanied by a mechanism that aligns it with the evolving garment.

Our central idea is to generate low-frequency dynamics while **propagating high-frequency product information directly from reference images along garment motion**. Low-frequency structure provides cues about body configuration, garment extent, and deformation; high-frequency components carry information about logos, edges, and textures. Structural descriptors determine where reference details should contribute in each target frame, while high-frequency features supply their appearance content. This division connects detail reuse to the garment's evolving pose and shape.

Building on this idea, we propose **W-DiT**, a Wavelet-infused Diffusion Transformer with three components. *Frequency disentanglement* decomposes each reference image and encodes its high-frequency

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

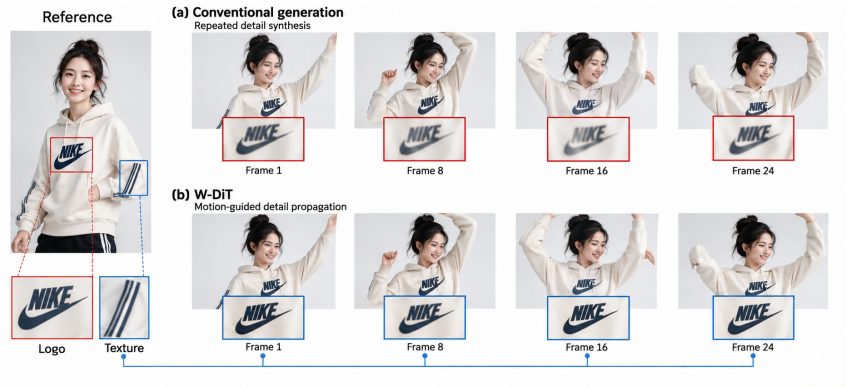

Figure 1: Reference-guided fashion video generation requires fine-grained garment details to remain consistent as pose and clothing geometry change.

subbands through a lightweight pathway that bypasses reference VAE compression. *Structure-guided alignment* computes cross-attention between reference and target structural descriptors, producing soft spatial correspondences for each frame. *High-frequency feature propagation* uses these weights to sample a fixed reference-detail bank and inject the aligned features into the DiT. Each frame therefore accesses the original reference features through correspondences that change with the generated structure. The video backbone integrates these features while jointly synthesizing garment appearance and motion.

To support the study of garment consistency, we curate **FashionVideo** from authorized Xiaohongshu images. The dataset contains approximately 10,000 synthesized videos and 30,000 associated images, constructed through prompt generation, reference-image editing, and video synthesis. It provides image and text conditions for fashion video generation, together with garment-editing annotations. Experiments evaluate garment fidelity and video quality, with controlled ablations examining high-frequency encoding and structure-guided sampling.

The main contributions are:

- **A frequency-decoupled video generation framework.** W-DiT couples garment motion synthesis with a dedicated reference-detail pathway that preserves high-frequency features outside VAE compression.
- **Structure-guided high-frequency propagation.** Soft structural correspondences guide sampling from a fixed reference bank, aligning garment details with changing poses and deformation.
- **FashionVideo and its evaluation.** The dataset provides approximately 10,000 synthesized videos and 30,000 associated images from authorized sources, supporting comparisons and component analyses of garment-detail preservation.

2 Related Work

2.1 Fashion-Specific Image-to-Video

Image-to-Video (I2V) aims to synthesize temporally coherent videos from one or more input images. Existing methods are commonly divided into **Keyframe-to-Video** and **Reference-Image-to-Video** paradigms. Keyframe-to-Video approaches Guo et al. (2023); Chen et al. (2023); Wang et al. (2023) treat the input image as the initial frame and generate subsequent frames through temporal modeling. Methods such as Stable Video Diffusion Blattmann et al. (2023) extend this paradigm with temporal attention or 3D convolutions, but often encounter identity drift or structural degradation in long sequences. More recent DiT-based models Brooks et al. (2024); Wan et al. (2025); Team et al. (2025); Zheng et al. (2024b) represent videos as spatio-temporal latent tokens and jointly denoise the sequence, supporting global consistency. Reference-Image-to-Video methods leverage the input image as a visual prior for identity or appearance while allowing flexible motion. Representative works include AnimateAnyone Hu (2024) and I2VGen-XL Zhang et al. (2023). Despite notable progress, maintaining long-term appearance consistency remains challenging.

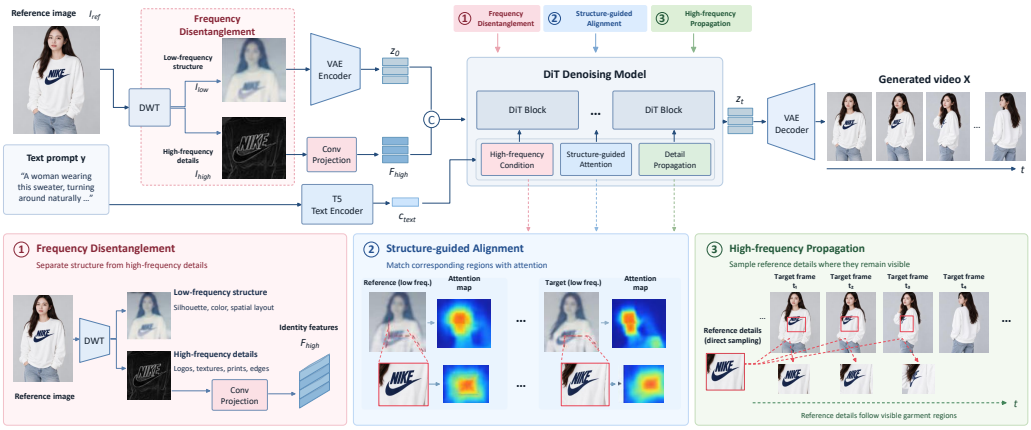

Figure 2: **Overview of W-DiT.** Structural correspondences guide sampling from a high-frequency reference bank, and aligned details are injected into the DiT through a learned gate. The lower panels illustrate frequency disentanglement, structure-guided alignment, and detail propagation.

Recent works have explored *fashion-specific* video synthesis using diffusion models. Representative approaches include DreamPose Karras et al. (2023) for pose-conditioned animation, Fashion-VDM Karras et al. (2024) and VITON-DiT Zheng et al. (2024a) for video virtual try-on, and Dress&Dance Chen et al. (2025) and ProFashion Kong et al. (2025) for multi-modal generation. M2HVideo Mu et al. (2025) addresses mannequin-to-human video synthesis, while Eevee Zeng et al. (2025) focuses on high-resolution close-up try-on. 3DV-TON Wei et al. (2025) constructs an image try-on result, reconstructs a textured 3D mesh, and animates it using source-video motion to guide video diffusion. W-DiT computes reference-to-target correspondences within the denoising process and uses them to sample garment details for motion generated from image and text conditions.

Frequency-guided detail preservation. HDW-SR Yang et al. (2026) uses wavelet decomposition and cross-frequency attention to guide detail restoration in image super-resolution. Our setting concerns garment details under changing video poses: structural features establish spatial correspondences, and those correspondences retrieve high-frequency values from a fixed reference bank.

2.2 Fashion Datasets

Fashion datasets such as GLAMI-1M Kosar et al. (2022), Fashion-Gen Rostamzadeh et al. (2018), and DeepFashion-MultiModal Jiang et al. (2022) primarily pair static images with text for classification, retrieval, and image generation. Video datasets, including UBC Fashion Zablotskaia et al. (2019) and Fashion-Text2Video Jiang et al. (2023), remain limited in scale and lack paired product-image conditioning. **FashionVideo** contains approximately 10,000 synthesized videos and 30,000 associated images from authorized Xiaohongshu imagery, with motion prompts and garment-editing annotations for studying reference-detail preservation.

3 Approach

W-DiT couples motion synthesis with reference-detail propagation in a pre-trained video diffusion backbone Wan et al. (2025). As shown in Fig. 2, frequency disentanglement builds separate structural and high-frequency reference features, structure-guided alignment determines their correspondence with the evolving video, and high-frequency propagation injects the aligned details into the DiT. The backbone jointly denoises the video, and its VAE decoder produces the output frames.

3.1 Problem Formulation

Given reference images $\mathcal{R} = \{I^n\}_{n=1}^{N}$ depicting the target garment and a motion prompt y, W-DiT generates a video $V = \{V_t\}_{t=1}^{T}$:

$$V \sim p_\theta(V \mid \mathcal{R}, y). \tag{1}$$

The number of references N may vary across examples; our experiments use up to five references ($1 \le N \le 5$). We use t for video time and $\tau \in [0, 1]$ for diffusion time; in the feature equations below, t indexes the temporal latent grid. For garment editing, an image editor first applies the requested change to a source image. The resulting target-outfit image then serves as the reference in Eq. (1).

3.2 Frequency Disentanglement

We apply a single-level 2D Haar wavelet transform independently to each reference image:

$$\left(L^n, B^n_{\mathrm{LH}}, B^n_{\mathrm{HL}}, B^n_{\mathrm{HH}}\right) = \mathrm{DWT}_{\mathrm{2D}}(I^n). \tag{2}$$

The low-frequency coefficients L^n retain coarse appearance and spatial layout. The three high-frequency bands capture directional edges and local texture variations. We encode them into paired structural and detail descriptors:

$$\begin{aligned} S^n &= P_S(E_{\mathrm{VAE}}(\mathcal{U}(L^n)))\,, \\ D^n &= E_{\mathrm{hf}}(\mathrm{Concat}[B^n_{\mathrm{LH}}, B^n_{\mathrm{HL}}, B^n_{\mathrm{HH}}])\,. \end{aligned} \tag{3}$$

Here $\mathcal{U}$ reconstructs a low-pass image at the input resolution by inverse DWT with zero detail bands. The frozen VAE encoder E_{VAE} and projection P_S produce structural descriptors $S^n \in \mathbb{R}^{M_n \times d}$. A lightweight convolutional encoder E_{hf} maps the concatenated detail bands to $D^n \in \mathbb{R}^{M_n \times d_h}$, bypassing the reference VAE encoder. The two outputs use the same spatial grid and token order, so $S^n(j)$ and $D^n(j)$ describe the same reference region.

Concatenating the references yields the structural bank $S = \mathrm{Concat}_n(S^n)$ and detail bank $D = \mathrm{Concat}_n(D^n)$, each with $M = \sum_n M_n$ entries. Both banks are computed once per reference set and remain fixed during sampling. Structural features supply coarse appearance conditioning through the backbone's reference pathway; the detail bank supplies high-frequency values.

3.3 Structure-Guided Alignment

The alignment module determines which reference regions supply details to each target location. At a selected DiT block, let $H_{\tau,t} \in \mathbb{R}^{P \times d_v}$ be the features at temporal position t and denoising time τ, with P spatial tokens. We omit the block index for clarity. Structural descriptors are extracted as

$$S^{\mathrm{tar}}_{\tau,t} = G_S(\mathcal{P}_{\mathrm{low}}(H_{\tau,t}), \tau)\,. \tag{4}$$

The operator $\mathcal{P}_{\mathrm{low}}$ applies spatial low-pass filtering to each channel on the target token grid. The adapter G_S projects the filtered features into $\mathbb{R}^{P \times d}$ and incorporates the diffusion-time embedding. At each denoising step, the adapter extracts target descriptors from the current video features.

Target descriptors provide queries and reference structural descriptors provide keys:

$$\begin{aligned} Q_{\tau,t} &= S^{\mathrm{tar}}_{\tau,t} W_Q, \qquad K = S W_K, \\ A_{\tau,t} &= \mathrm{softmax}_j\left(\frac{Q_{\tau,t} K^\top}{\sqrt{d_a}} + B\right). \end{aligned} \tag{5}$$

Here W_Q and W_K project to dimension d_a. The additive mask B excludes padded or unavailable reference tokens, and the softmax is taken over reference positions j. Each row of $A_{\tau,t} \in \mathbb{R}^{P \times M}$ specifies how one target location draws information from the reference regions. The paired detail bank D supplies the attention values.

The correspondence covers all reference images in the set. Spatial positions are encoded consistently for the structural and detail tokens within each reference, while the target features retain the backbone's spatial and temporal positional encoding. The frame-dependent weights select reference regions for detail injection.

3.4 High-Frequency Feature Propagation

Using the structural correspondence, we sample high-frequency features from the reference bank:

$$\widehat{D}_{\tau,t}(p) = \sum_{j=1}^{M} A_{\tau,t}(p, j) D(j), \qquad \text{or} \qquad \widehat{D}_{\tau,t} = A_{\tau,t} D. \tag{6}$$

The index j identifies a reference image and a spatial location within it. Thus, $\widehat{D}_{\tau,t}(p)$ combines reference details selected for position p, and $\widehat{D}_{\tau,t} \in \mathbb{R}^{P\times d_h}$ is aligned with the target token grid.

A learned gate controls the strength of detail injection:

$$\begin{aligned} g_{\tau,t} &= \sigma\big(G_g\big(S^{\mathrm{tar}}_{\tau,t},\, A_{\tau,t}S,\, e(\tau)\big)\big)\,, \\ H'_{\tau,t} &= H_{\tau,t} + g_{\tau,t} \odot P_D\Big(\widehat{D}_{\tau,t}\Big)\,. \end{aligned} \tag{7}$$

Here $A_{\tau,t}S$ is the reference structure aggregated with the same correspondence weights, $e(\tau)$ is the diffusion-time embedding, and σ is the sigmoid function. The gate $g_{\tau,t} \in (0,1)^{P\times 1}$ depends on the target structure, matched reference structure, and denoising time. The projection P_D maps the aligned details to dimension d_v, and the gate is broadcast over channels for the residual update.

The reference bank D remains fixed, while $A_{\tau,t}$ follows changes in the target structure. At each frame, these weights select and align reference features with the moving garment. The DiT integrates the sampled details through the gated residual update.

3.5 Training Objective

The trainable reference and propagation modules are optimized through the backbone's generative objective, with the VAE encoder and decoder frozen. Let $z_0 = E_{\mathrm{VAE}}(V)$ and $\epsilon \sim \mathcal{N}(0, I)$. For $z_\tau = (1-\tau)z_0 + \tau\epsilon$, the flow-matching objective is

$$\mathcal{L}_{\mathrm{gen}} = \mathbb{E}_{V,\mathcal{R},y,\tau,\epsilon}\left[\left\|v_\theta(z_\tau, \tau; \mathcal{R}, y) - (\epsilon - z_0)\right\|_2^2\right]. \tag{8}$$

During training, reference banks are computed from the conditioning images and alignment queries from the noisy video state z_τ. The clean video supplies the training target.

We retain the wavelet-domain adversarial loss as an auxiliary texture objective. Let $\mathcal{H}$ extract the three spatial high-frequency bands of a frame and D_{hf} be a patch discriminator. A predicted clean latent is obtained as $\widehat{z}_0 = z_\tau - \tau v_\theta(z_\tau, \tau; \mathcal{R}, y)$, and $\widehat{V}_t$ denotes a frame decoded from this estimate. The objectives are

$$\begin{aligned} \mathcal{L}_{\mathrm{hf_adv}} &= -\mathbb{E}\big[\log D_{\mathrm{hf}}(\mathcal{H}(\widehat{V}_t))\big], \\ \mathcal{L}_D &= -\mathbb{E}\big[\log D_{\mathrm{hf}}(\mathcal{H}(V_t))\big] \\ &\quad - \mathbb{E}\big[\log(1 - D_{\mathrm{hf}}(\mathcal{H}(\widehat{V}_t)))\big], \\ \mathcal{L}_{\mathrm{total}} &= \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{hf}}\mathcal{L}_{\mathrm{hf_adv}}. \end{aligned} \tag{9}$$

Expectations include sampled frames and discriminator patches. The discriminator update uses detached generated frames; the generator receives gradients through the frozen VAE decoder. The auxiliary loss encourages high-frequency texture realism, while the correspondence-weighted reference features provide garment-specific appearance conditioning. We retain $\lambda_{\mathrm{hf}} = 0.1$.

4 The FashionVideo Dataset

Fashion video generation requires paired visual and textual conditions that describe both garment appearance and motion. We curate **FashionVideo** from authorized Xiaohongshu images, using image editing and video synthesis to construct domain-specific training examples. All target videos are synthesized from these image-based inputs.

4.1 Dataset Composition

FashionVideo contains approximately **10,000 videos** and **30,000 associated images**. A generation example consists of a target video V, its available reference set $\mathcal{R}$, and a motion prompt y. Associated editing examples additionally record a source image, an edited image, and the instruction describing the garment change. These editing annotations belong to the same collection of video records.

Each reference set contains available views of the target garment. Editing records also retain the source outfit as context. Reference-guided generation uses the target-outfit images as appearance conditions, with up to five references per example in our experiments.

Table 1: Comparison on FashionVideo. W-DiT uses one or three references, as indicated; all baseline rows use one reference.

Method	Appearance Metrics			Motion Metrics		Fidelity Metrics		
	Aesthetic Quality ↑	Image Quality ↑	Subject Consist. ↑	Motion Smooth. ↑	Dynamic Degree ↑	Color Accuracy ↑	**Detail Fidelity (DFS) ↑**	**Face Identity (FaceID) ↑**
AnimateAnyone Hu (2024)	47.2	49.3	88.1	97.2	88.7	75.4	68.5	71.2
CogVideoX-5B-I2V Yang et al. (2025)	51.4	53.2	91.6	97.5	87.8	78.1	70.3	74.5
Keling Team et al. (2025)	52.0	61.7	93.3	99.4	64.8	80.2	73.8	76.1
Wan2.2 Wan et al. (2025)	52.3	63.1	93.8	99.6	88.1	81.5	72.1	79.8
W-DiT (1 reference)	53.5	64.8	94.7	99.5	89.7	89.2	82.7	85.9
W-DiT (3 references)	**54.7**	**65.9**	**95.6**	**99.6**	**91.5**	**91.0**	**85.3**	**88.6**

4.2 Prompt, Reference, and Video Construction

Gemini 2.5 Pro generates descriptions of garment appearance, scene, initial pose, and motion, including requested appearance changes for editing examples. Qwen-Image-Edit constructs target-outfit references from authorized source images, and Wan2.5 synthesizes videos conditioned on these references and prompts. Manual review removes examples with visible identity distortions, incorrect garment edits, or conspicuous generation artifacts.

5 Experiments

5.1 Evaluation Setup

Data and inputs. We reserve 10% of FashionVideo for testing, yielding approximately 1,000 videos. Each source image, its edits, and derived videos remain in the same split. Unless stated otherwise, quantitative comparisons and ablations use one reference and one motion prompt. Table 1 also reports W-DiT with three references, while Figure 3 uses up to five.

We select publicly accessible baselines that can be evaluated with the reference images and text prompts. To make the comparisons as consistent as the available implementations allow, we evaluate all methods on the same held-out examples. Baselines with available training code and weights are fine-tuned on the FashionVideo training split; otherwise, we use their released checkpoints or services. We match the output resolution, frame rate, and duration wherever supported, and apply the same frame sampling and metric pipeline to all outputs. Methods requiring additional conditions receive the inputs specified by their original protocols.

Metrics. We report scores on a 0–100 scale. VBench Huang et al. (2024) measures aesthetic quality, image quality, subject consistency, motion smoothness, and dynamic degree. Appearance is the arithmetic mean of the first three scores; Motion averages the last two.

The Detail Fidelity Score (DFS) measures reference-garment similarity. Garment regions are segmented with SAM Kirillov et al. (2023), and cosine similarities between DINOv2 Oquab et al. (2024) descriptors from the 11th Transformer block are averaged over evaluated frames. FaceID measures person identity using cosine similarity between ArcFace Deng et al. (2019) embeddings of reference and generated faces. All methods use the same frame sampling and region-processing procedure.

5.2 Implementation Details

W-DiT is initialized from Wan2.2-I2V-A14B Wan et al. (2025) and follows its default backbone and sampling configuration. The Wan2.2 control uses the same FashionVideo split, training budget, output resolution, frame count, and inference settings. The VAE is frozen. A three-layer convolutional encoder processes the single-level Haar detail bands. Structural attention and gated injection follow Equations 5–7, with $\lambda_{\mathrm{hf}} = 0.1$. Component ablations use one reference and matched training and inference settings. Detail-path variants share encoder capacity, attention width, and injection blocks.

5.3 Main Results

FashionVideo. Table 1 evaluates garment fidelity while monitoring overall video quality. With one reference, W-DiT improves DFS from 72.1 to 82.7 and FaceID from 79.8 to 85.9 over the



Figure 3: Qualitative examples with up to five reference images for W-DiT. Two references are displayed, and motion prompts are abbreviated.

Table 2: Evaluations on Fashion-Text2Video (left) and try-on video editing (right). The editing comparison conditions all generators on Qwen-Image-Edit outputs.

Method	Fashion-Text2Video				Try-on Editing		
	Appear.↑	Motion↑	DFS↑	FaceID↑	Appear.↑	Motion↑	FaceID↑
AnimateAnyone Hu (2024)	62.4	89.1	60.8	75.3	65.5	91.3	75.1
CogVideoX-5B-I2V Yang et al. (2025)	64.6	90.2	62.7	77.8	67.2	92.1	77.4
Keling Team et al. (2025)	67.9	91.5	66.1	80.6	66.1	93.5	79.2
Wan2.2 Wan et al. (2025)	68.1	91.8	65.4	82.2	67.6	93.8	82.3
KeyTailor He et al. (2025)	64.7	89.4	63.9	79.8	66.0	92.1	80.5
Dress&Dance Chen et al. (2025)	65.2	89.5	63.8	79.4	65.8	92.5	81.3
W-DiT (Ours)	**72.0**	**93.4**	**78.9**	**87.0**	**73.2**	**95.8**	**90.1**

Wan2.2 control, while retaining comparable Motion. The improvement is therefore concentrated in appearance preservation. Three references further raise DFS to 85.3 and FaceID to 88.6.

Qualitative comparison. To inspect garment preservation under pose changes, Figure 3 compares generated sequences with their references. W-DiT retains the light outerwear and red-and-black pattern during walking and turning, whereas Wan2.2 shows greater variation in garment shape and pattern placement.

Temporal detail fidelity. To examine detail degradation over time, we compute DFS over the first, middle, and final thirds of each video. Wan2.2 drops from 75.2 to 69.3, whereas W-DiT remains stable from 83.3 to 82.1, widening the gap from 8.1 to 12.8 points. This trend supports using a fixed reference-detail bank to preserve garment appearance throughout motion. Detailed results are reported in Appendix C.

External evaluation. Since FashionVideo contains synthesized target videos, we further evaluate on Fashion-Text2Video, a separate collection of captured fashion videos Jiang et al. (2023). Using each video's first frame as the reference for all methods, W-DiT requires no additional fine-tuning and achieves 78.9 DFS and 87.0 FaceID, exceeding the best competing scores by 12.8 and 4.8 points, respectively (Table 2). It also leads in Appearance and Motion, showing that the garment-fidelity advantage extends to captured videos.



Figure 4: Garment-editing examples with abbreviated prompts.

Table 3: Component ablations on FashionVideo using one reference.

Model variant	DFS ↑	Appearance ↑	Motion ↑	FaceID ↑
Wan2.2	72.1	69.7	93.9	79.8
W-DiT w/o HF Reference Path	75.5	70.1	94.1	82.8
W-DiT with Full-band Reference Features	80.2	70.6	94.4	84.8
HF + Standard Attention	78.4	70.4	94.2	84.1
W-DiT (Full)	**82.7**	**71.0**	**94.6**	**85.9**
W-DiT w/o Adaptive Gating	80.6	70.3	94.3	84.6
W-DiT w/o HF Adversarial Loss	81.6	70.7	94.5	85.5

5.4 Evaluation on Try-On Video Editing

This experiment tests whether detail propagation remains effective after an upstream garment edit. Qwen-Image-Edit first produces the target garment; each generator then receives the same edited reference and motion prompt. In the right block of Table 2, W-DiT improves Appearance, Motion, and FaceID over Wan2.2 by 5.6, 2.0, and 7.8 points. Figure 4 shows examples, and Appendix B documents both input settings.

5.5 Ablation Study

Quantitative ablations. Table 3 isolates the contributions of high-frequency encoding, structural correspondence, gating, and auxiliary supervision. All variants use one reference.Removing the detail path reduces DFS by 7.2 points despite retaining structural conditioning, showing that *reference details contribute beyond coarse appearance cues*. Replacing wavelet bands with RGB features under matched encoder capacity lowers DFS, supporting the benefit of *explicit high-frequency encoding*. Replacing structural queries and keys with target hidden and high-frequency reference features, while preserving the attention values and injection pathway, reduces DFS by 4.3 points with little change in Motion. This result supports *structure-guided selection of reference details*. Uniform injection and removal of the high-frequency adversarial loss cause smaller declines, with gating contributing more than auxiliary supervision. Motion remains comparable across these variants, indicating that improved fidelity does not come from suppressing motion.

Visual ablation. To examine the effect of structure-guided alignment on garment details, Figure 5 compares Wan2.2, HF + Standard Attention, and the full W-DiT across two reference-guided sequences. On the black jacket, the chest print changes in shape and internal arrangement in the baselines, particularly in later frames, while W-DiT retains a more consistent layout. On the sweater, W-DiT better preserves the large motif as the subject turns. These examples illustrate the detail-fidelity differences measured in Table 3 and show the contribution of structural matching beyond adding a high-frequency reference path.



Figure 5: Visual ablation of detail preservation. Red boxes indicate enlarged garment regions.



Figure 6: Reference-region alignment across changing poses. Red boxes mark target scarf regions; lower panels show the corresponding attention over the reference image.

Correspondence visualization. To examine whether structural correspondence retrieves relevant reference regions across changing poses, Figure 6 visualizes attention from the target scarf regions. Across three poses, attention remains concentrated on the reference scarf, supporting structure-guided detail alignment.

5.6 Human Evaluation and Metric Reliability

To verify whether the measured gains are perceptible, we compare W-DiT and Wan2.2 on matched references and prompts. Among non-tied judgments, W-DiT is preferred for garment texture (76.5%), identity (79.3%), and overall quality (71.2%). Appendix A details the protocol and reports agreement between DFS, FaceID, and human judgments.

6 Conclusion

We present **W-DiT**, a structure-guided high-frequency propagation framework for reference-guided fashion video generation. The key to faithful fashion video generation is to synthesize changing garment motion while using that structure to align and reuse fine appearance details from the reference images. Experiments show improved garment-detail fidelity over the matched Wan2.2 control while maintaining motion quality. We also introduce **FashionVideo**, comprising approximately 10,000 synthesized videos and 30,000 associated images derived from authorized Xiaohongshu imagery. The method remains limited by correspondence errors, missing reference views, the output VAE's reconstruction capacity, and artifacts inherited from synthetic training targets. Future work will focus on visibility-aware correspondence and evaluation across more diverse captured videos.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets (specifically in curating the FashionVideo dataset, as detailed in Section 4). We have not used generative AI tools for developing conceptual models, formulating hypotheses, designing experiments, implementing methods, cleaning data, or interpreting results, and mathematical proofs are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Jun-Kun Chen, Aayush Bansal, Minh Phuoc Vo, and Yu-Xiong Wang. Dress&dance: Dress up and dance as you like it - technical preview, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Qingdong He, Xueqin Chen, Yanjie Pan, Peng Tang, Pengcheng Xu, Zhenye Gan, Chengjie Wang, Xiaobin Hu, Jiangning Zhang, and Yabiao Wang. The devil is in the details: Enhancing video virtual try-on via keyframe-driven details injection. *arXiv preprint arXiv:2512.20340*, 2025.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: text-driven controllable human image generation. *ACM Trans. Graph.*, 41(4), July 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530104.

Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22747–22757, October 2023.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22623–22633, 2023.

Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, Chris Lee, and Ira Kemelmacher-Shlizerman. Fashion-vdm: Video diffusion model for virtual try-on. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687623.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Xianghao Kong, Qiaosong Qi, Yuanbin Wang, Anyi Rao, Biaolong Chen, Aixi Zhang, Si Liu, and Hao Jiang. Profashion: Prototype-guided fashion video generation with multiple reference images, 2025.

Vaclav Kosar, Antonín Hoskovec, Milan Šulc, and Radek Bartyzal. Glami-1m: A multilingual image-text fashion dataset. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

Bolin Lai, Xudong Wang, Saketh Rambhatla, James M Rehg, Zsolt Kira, Rohit Girdhar, and Ishan Misra. Toward diffusible high-dimensional latent spaces: A frequency perspective. *arXiv preprint arXiv:2511.22249*, 2025.

Tejaswini Medi, Hsien-Yi Wang, Arianna Rampini, and Margret Keuper. FAVAE-effective frequency aware latent tokenizer. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*, 2025.

Xiangyu Mu, Dongliang Zhou, Jie Hou, Haijun Zhang, and Weili Guan. From mannequin to human: A pose-aware and identity-preserving video generation framework for lifelike clothing display. *arXiv preprint arXiv:2510.16833*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge, 2018.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

Min Wei, Chaohui Yu, Jingkai Zhou, and Fan Wang. 3dv-ton: Textured 3d-guided consistent video try-on via diffusion models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9345–9354, 2025.

Chao Yang, Boqian Zhang, Jinghao Xu, and Guang Jiang. Hdw-sr: High-frequency guided diffusion model based on wavelet decomposition for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23462–23472, 2026.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation, 2019.

Jianhao Zeng, Yancheng Bai, Ruidong Chen, Xuanpu Zhang, Lei Sun, Dongyang Jin, Ryan Xu, Nannan Zhang, Dan Song, and Xiangxiang Chu. Eevee: Towards close-up high-resolution video-based virtual try-on. *arXiv preprint arXiv:2511.18957*, 2025.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. Viton-dit: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint*, 2024a.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024b.

A Human Evaluation

A.1 Study Design

This appendix details the human evaluation in Section 5.6, using pairwise comparisons with a no-preference option.

Participants. We recruited 20 human evaluators with normal or corrected-to-normal vision. Evaluators were not affiliated with this project and were naive to the purpose of the study.

Stimuli. We randomly sampled 200 video pairs from the FashionVideo test set, each pair consisting of one video generated by W-DiT and one by Wan2.2 (the strongest baseline), matched on the same reference image and text prompt. Each evaluator was assigned 50 pairs in randomized order.

Task. For each pair, evaluators were asked to judge along three dimensions independently:

- **Garment Texture Fidelity**: "Which video better preserves the texture and pattern of the reference garment?"
- **Identity Preservation**: "In which video does the person's facial identity remain more consistent with the reference image?"
- **Overall Quality**: "Which video is of higher overall visual quality?"

For each question, evaluators selected one of three options: *Left*, *Right*, or *No Preference*. The assignment of W-DiT and Wan2.2 to left/right was randomized and balanced across pairs.

Compensation. All evaluators were compensated at a rate exceeding the local minimum wage.

A.2 Results

Table 4: Human pairwise evaluation results. Preference rates are computed over all judgments excluding "No Preference" responses. Inter-annotator agreement (Fleiss' κ) is reported for each dimension.

Dimension	W-DiT Preferred	Wan2.2 Preferred	Fleiss' κ
Garment Texture Fidelity	76.5%	23.5%	0.61
Identity Preservation	79.3%	20.7%	0.58
Overall Quality	71.2%	28.8%	0.55

Among judgments selecting a method, W-DiT is preferred for texture fidelity, identity preservation, and overall quality. Fleiss' κ ranges from 0.55 to 0.61.

Table 5: Pearson correlation between automated metrics and human preference scores across 200 video pairs.

Metric	Dimension	Pearson r
DFS	Garment Texture Fidelity	0.74
FaceID	Identity Preservation	0.78

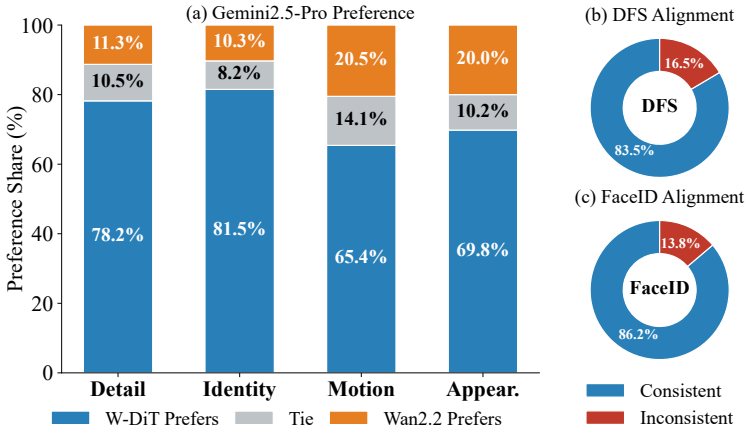

Figure 7: Metric validation: (a) Gemini preference evaluation; (b–c) agreement with DFS and FaceID.

A.3 Correlation with Automated Metrics

Table 5 reports correlations of $r = 0.74$ between DFS and human texture judgments and $r = 0.78$ between FaceID and human identity judgments across the 200 evaluated pairs. These positive associations support using the metrics within this sample. Their interpretation depends on the recorded construction of per-pair metric differences and human preference scores.

A.4 Gemini-Based Evaluation

As an independent perceptual check, the judge evaluates 1,000 generated videos and prefers W-DiT in 78.2% of detail-fidelity and 81.5% of identity comparisons (Figure 7).

B Try-On Video Editing: Fair Comparison Input

B.1 Setting

The current editing application supplies the Qwen-Image-Edit output, with the target garment already applied, to the video model as its reference. Table 6 compares source-image and edited-reference inputs. The edited-reference setting follows Section 5.4.

B.2 Results

This comparison checks that the editing gains do not arise from unequal upstream inputs. With the same edited reference, W-DiT† achieves the best Appearance, Motion, and FaceID scores.

C Temporal Detail Fidelity

We divide the generated frames into three equal temporal intervals. DFS is first averaged within each interval for every video and then averaged across the test set. Table 7 shows that W-DiT maintains garment fidelity throughout the sequence, while Wan2.2 degrades as generation progresses.

Table 6: Garment-editing results under two input settings. W-DiT† and the baselines receive Qwen-Image-Edit outputs; the source-image variant receives the original image.

Method	Input Reference	Appear.↑	Motion↑	FaceID↑
AnimateAnyone Hu (2024)	QwenEdit image	65.5	91.3	75.1
CogVideoX-5B-I2V Yang et al. (2025)	QwenEdit image	67.2	92.1	77.4
Keling Team et al. (2025)	QwenEdit image	66.1	93.5	79.2
Wan2.2 Wan et al. (2025)	QwenEdit image	67.6	93.8	82.3
KeyTailor He et al. (2025)	QwenEdit image	66.0	92.1	80.5
Dress&Dance Chen et al. (2025)	QwenEdit image	65.8	92.5	81.3
W-DiT (source-image setting)	Original image	70.7	95.1	88.5
W-DiT† (Ours)	QwenEdit image	**73.2**	**95.8**	**90.1**

Table 7: Temporal DFS on FashionVideo under single-reference conditioning. Values report the cross-video mean and standard error.

Segment	Wan2.2	W-DiT	ΔDFS
Early	75.2 ± 0.5	83.3 ± 0.4	$+8.1$
Middle	71.8 ± 0.6	82.7 ± 0.5	$+10.9$
Late	69.3 ± 0.6	82.1 ± 0.5	$+12.8$

D Try-On Video Editing: Prompt Generation Instructions

The following system prompt is used for editing-caption construction. Given an edited reference image, a source video, a short caption, and an edit instruction, it describes the source video's original garment relative to the edited image and expands the observed motion. These annotations belong to FashionVideo's editing records; they do not define a separate additional 10,000-video dataset. The reference image in this reverse-captioning procedure is an edited observation, whereas the current video-generation reference in Section 3 depicts the target outfit to be preserved.

System Prompt for Try-On Editing Caption Generation

System Prompt
You are an expert in video generation instruction writing. Given a reference image, a source video, a Short Caption (`$text:{caption}`), and an edit prompt (`$edit:{edit_prompt}`), your task is as follows.
Core Logic: The reference image is the result of applying the "`$edit`" instruction to the original video. Using the video as the ground truth, and the `$edit` instruction as a clue, restore the subject's original garment appearance and expand the motion description based on video details.

Rules:
1. Visual Feature Reverse Restoration (Priority: B > A)

- **Case B (no `$edit`):** Directly state that "the subject's garment is fully consistent with the reference image."
- **Case A (with `$edit`):** Identify all modifications described in `$edit` and explicitly "undo" them in the description. *Example: if `$edit` is "change to a red dress" but the video shows blue, describe it as "the dress in the reference image is changed to blue."*

2. Dynamic Motion Expansion (Stage-by-Stage Description):

- Do not simply repeat the Short Caption. Observe the video and decompose the action into: **[Initial Static Pose] + [Dynamic Evolution over 5 Seconds]**.
- The description must include changes in body weight distribution, gesture details, and facial expressions.

3. Output Structure (strictly enforced):

- **Sentence 1 (Environment & Anchor):** Describe the scene environment and state that the subject's facial features strictly follow the reference image.

- **Sentence 2 (Garment Restoration):** Describe the garment state based on Case A or B.
- **Sentences 3–4 (Full Motion Description):** First describe the precise pose at the start of the video, then describe the coherent motion evolution over the following 5 seconds.
- **Length:** 150–200 words; professional and visually descriptive language required.

Output Examples:
Case A (reverse restoration applied):
The scene environment and the subject's facial features strictly follow the reference image. Garment: the beige long-sleeve knit top in the reference image is replaced with a white sleeveless tweed dress. Initial pose: the woman stands sideways to the camera, her right hand lightly pinching the strap of her bag, weight slightly shifted to her left leg, expression elegant and natural. Motion: as the camera slowly pushes forward, she gracefully turns toward it; the hem of the dress sways lightly with the movement; she then raises her left hand to gently brush aside the loose strands of hair near her ear, her eyes gazing warmly into the lens with strong camera presence.

Case B (no edit instruction):
The scene environment and the subject's facial features fully follow the reference image. Garment: fully consistent with the white slip dress and matching straw hat in the reference image. Initial pose: the woman sits slightly sideways on the edge of a light-colored sofa, hands naturally folded on her knees, head gently turned to the right, long hair falling over her shoulders. Motion: she slowly adjusts her sitting posture, leaning slightly forward; her arms then extend gradually to both sides and lift upward, fingers spread; finally she raises her head toward the camera and reveals a calm, serene smile.