DUALCAST: DUAL-IMAGE CONDITIONED DIFFUSION TRANSFORMERS FOR DIGITAL HUMAN E-COMMERCE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

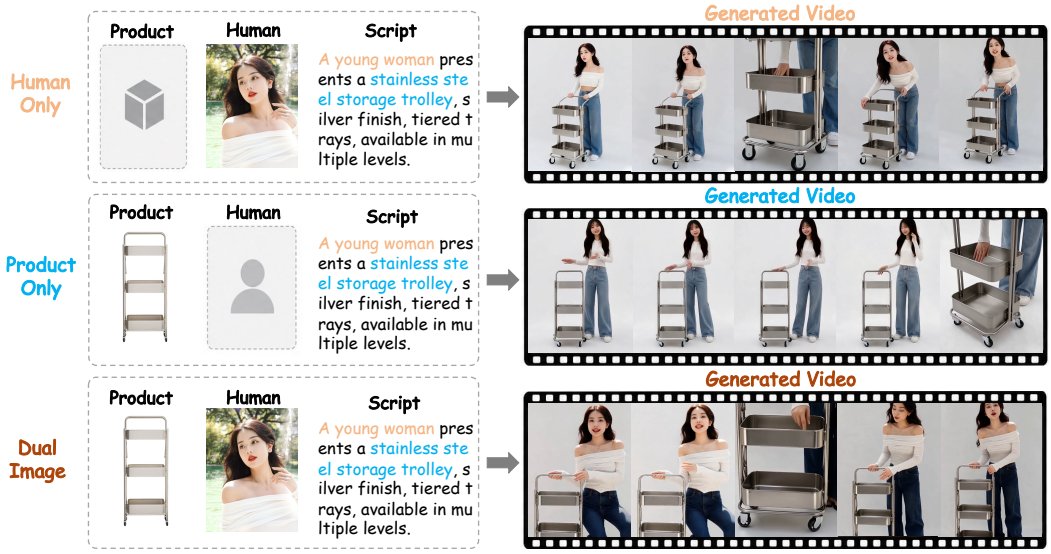

Figure 1: Three inference modes of one fine-tuned DualCast checkpoint, from top to bottom: human-only, product-only, and dual-image. An omitted visual reference is supplied semantically through the script rather than matched to an unavailable image.

ABSTRACT

Digital human e-commerce video generation requires preserving product appearance, human identity, and overall video style during dynamic demonstrations. Although recent models support multiple reference images, their general conditioning schemes do not explicitly distinguish the asymmetric roles of "what is presented" and "who presents it." Consequently, generated videos may favor a single reference, resulting in product distortion, identity drift, or single-reference collapse. To address this issue, we propose **DualCast**, a role-addressed dual-image conditioning framework for video generation. By explicitly modeling the distinct roles of product and human conditions, DualCast reduces role ambiguity and feature competition between references. First, a dual-spatiotemporal slot injection mechanism VAE-encodes the product and human images separately and places their latent features at the front and rear boundaries of the DiT sequence, enabling sequence-level condition separation and directed injection. Second, a mask-guided region-weighted loss strengthens supervision over product, face, and product–face overlap regions, improving generation fidelity in critical areas. Third, image condition dropout enables a single model to support dual-image, product-only, and human-only generation. In addition, we construct a curated dataset of 10K e-commerce videos for model training. Experimental results demonstrate the strong overall performance of DualCast.

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.

1 INTRODUCTION

Digital human e-commerce video generation is conditioned on specific human and product references. Given human and product images and a text script, the goal is to generate a coherent promotional video in which a digital human introduces, displays, or demonstrates the specified product. Unlike generic text-to-video (T2V) generation, this task requires semantic and temporal coherence alongside consistent product appearance, human identity, and overall video style. Fine-grained product attributes, including logos, packaging geometry, textures, and colors, must be faithfully reproduced, while human identity should remain stable across motion and viewpoint changes. Product deformation, identity drift, or rendering errors in product–face overlap regions can significantly reduce the usability of generated videos.

Recent diffusion transformers (DiTs) and large-scale video generation models Peebles & Xie (2023); Wan et al. (2025) have demonstrated strong temporal modeling capabilities. Multi-image conditioning further enables joint control of human and product content Jiang et al. (2025); Fei et al. (2025); Zhang et al. (2025). Reference-conditioned models can combine multiple images through concatenation or shared attention. Without an explicit role convention, these interfaces may treat the supplied images as interchangeable visual conditions. Such an undifferentiated interface does not explicitly distinguish the asymmetric roles of "what is presented" and "who presents it," forcing the model to infer each reference's target role from a mixed condition space. The resulting feature competition and attention bias may cause one reference to dominate, leading to product distortion, human identity drift, or single-reference collapse. This ambiguity intensifies when product and human features affect the same local regions.

To address these issues, we propose **DualCast**, a role-addressed dual-image conditioning framework for video generation. Its core idea assigns fixed, distinguishable positions to product and human conditions in the DiT sequence, establishing explicit correspondence between references and roles. Specifically, a dual-spatiotemporal slot injection mechanism independently VAE-encodes the product and human images and places their latent features at the front and rear boundaries of the DiT sequence, respectively. These boundary positions provide role addresses within the shared sequence. Shared CLIP and text tokens also supply semantic role cues; positional separation does not imply independent attention. The boundary slots serve only as conditions; training loss and decoding are restricted to the intermediate video latents, leaving the output video length unchanged.

In addition to condition injection, a uniform training objective may provide insufficient supervision for local regions critical to visual fidelity. Standard flow-matching or noise-prediction losses assign equal weights to all latent locations, limiting the contribution of product, facial, and product–face overlap regions that occupy only small portions of a frame. We therefore introduce a mask-guided region-weighted loss that constructs a spatial weighting map from product and face masks and strengthens supervision over product, facial, and product–face overlap regions. This objective increases the contribution of fidelity-critical features during optimization, while loss normalization controls total spatial weight across samples with different mask areas.

Practical applications also require flexible input configurations. Users may provide both product and human images or only one reference, while training separate models for different configurations is inefficient. We thus introduce an image condition dropout strategy. During fine-tuning, the product or human condition is randomly removed with predefined probabilities, and the corresponding reference features and region masks are disabled. This strategy enables a single DualCast model to support dual-image, product-only, and human-only inference. Even when one reference is absent, the model can still generate a complete digital human promotional video. In addition, we construct a curated dataset of 10K e-commerce video clips with paired product–human references and precomputed region masks for model training and evaluation. The evaluation shows leading perceptual scores and the best average rank across nine metrics, alongside favorable reference and video-quality results.

Our main contributions are summarized as follows:

- **Role-Addressed Task Formulation:** We formulate digital human e-commerce video generation as a role-addressed dual-image conditioning problem that explicitly distinguishes the asymmetric roles of product and human references.

- **Dual-Spatiotemporal Slot Injection:** We inject independently VAE-encoded product and human latent features into the front and rear non-generative boundary slots of the DiT sequence, using fixed positional priors to identify the two condition roles within shared attention.
- **Region-Weighted Supervision:** We construct a mask-guided region-weighted loss from product and face masks to strengthen supervision over product, facial, and product–face overlap regions.
- **Flexible Inference and E-Commerce Dataset:** We unify dual-image, product-only, and human-only inference through image condition dropout and construct a curated dataset containing 10K e-commerce video clips.

2 Related Work

Video generation and reference conditioning. Diffusion and flow matching underpin recent image and video synthesis (Ho et al., 2020; Rombach et al., 2022; Ho et al., 2022; Lipman et al., 2023). DiTs (Peebles & Xie, 2023) scale these approaches to spatiotemporal latents. Wan (Wan et al., 2025) supplies our pretrained video prior. ControlNet, IP-Adapter, and DreamBooth offer complementary conditioning and personalization mechanisms (Zhang et al., 2023a; Ye et al., 2023; Ruiz et al., 2023); Tune-A-Video and AnimateDiff extend reference-driven generation to video (Wu et al., 2023b; Guo et al., 2024).

VACE, SkyReels-A2, Kaleido, and Phantom support reference-based video synthesis (Jiang et al., 2025; Fei et al., 2025; Zhang et al., 2025; Liu et al., 2025). Kaleido's reference-aware positional design is directly relevant to role separation. Movie Weaver binds references to concepts using anchored prompts and concept embeddings (Liang et al., 2025); MAGREF combines masked guidance with subject disentanglement (Deng et al., 2026). HuMo uses temporal VAE-reference injection within a multimodal human-generation framework (Chen et al., 2025). DualCast specializes reference conditioning to fixed product/presenter roles at opposite latent boundaries, combining this convention with regional e-commerce supervision and a shared interface for complete or partial reference inputs.

Human animation and object interaction. Motion transfer, talking faces, and pose-conditioned animation emphasize human appearance and dynamics (Siarohin et al., 2019; Prajwal et al., 2020; Zhang et al., 2023b; Hu et al., 2023; Xu et al., 2024; Lin et al., 2025). Product demonstrations additionally require recognizable objects under occlusion and manipulation. Our regional loss gives product and face appearance explicit emphasis within the video objective, complementing the pretrained motion prior with task-specific supervision. Commercial systems are included as end-to-end comparators through their supported reference interfaces, with training-data differences stated in the comparison protocol.

3 Method

To address feature entanglement in dual-subject synthesis, we propose DualCast, a role-addressed framework (Fig. 2). It comprises three core designs: (1) dual-spatiotemporal slot injection for role assignment (Sec. 3.1); (2) a mask-guided region-weighted loss for region fidelity (Sec. 3.2); and (3) joint training strategies for multi-mode deployment (Sec. 3.3). We also construct a large-scale e-commerce video dataset for training (Sec. 3.4).

3.1 Dual-Spatiotemporal Slot Injection

The Wan2.2-I2V backbone supplies a single-image interface. We extend it with a fixed two-role latent convention. In digital human e-commerce generation, product and presenter conditions are asymmetric: the former specifies what is presented, whereas the latter specifies who presents it. We therefore assign both references to fixed boundary positions in the latent sequence. Let $I_p, I_h \in \mathbb{R}^{3\times H\times W}$ denote the product and presenter images, respectively, and V a target video of T frames. We encode V as

$$x_0 = \mathcal{E}_{\text{VAE}}(V) \in \mathbb{R}^{C_v\times T'\times h\times w}, \tag{1}$$

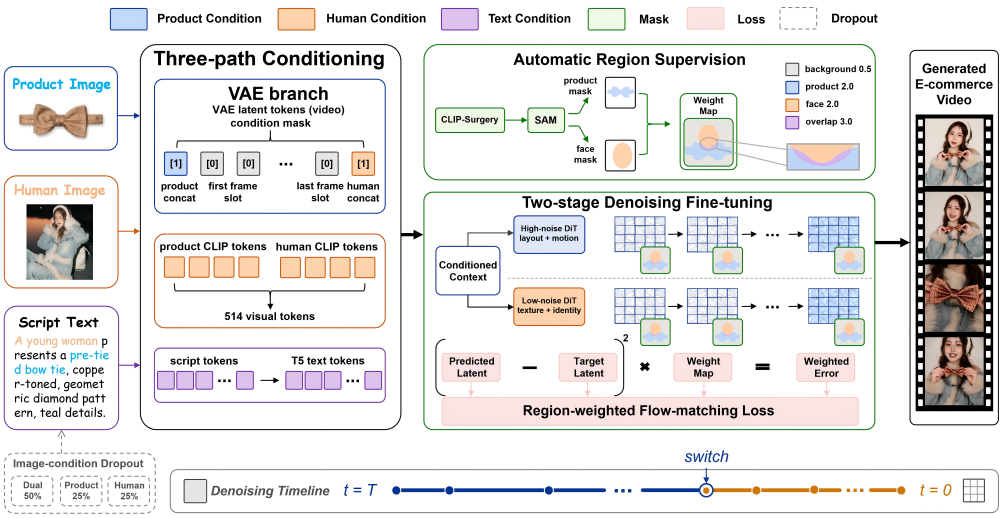

Figure 2: **Overview of DualCast.** The product and presenter images are independently VAE-encoded and placed at the front and rear boundaries of the extended latent sequence, respectively. The stated mode probabilities are 50/25/25 for dual/product/human inputs. Expert switching is inherited from Wan2.2. Only intermediate video latents are denoised, supervised, and decoded. Both reference CLIP embeddings combine with the text prompt as cross-attention context. Product and face masks define region-weighted supervision, while condition dropout supports the three reference-availability patterns within the same checkpoint and role convention.

where T' is the latent temporal length. At flow-matching timestep τ, the noisy video latent is

$$x_\tau = (1-\tau)x_0 + \tau\epsilon. \tag{2}$$

The references are independently encoded as

$$z_p = \mathcal{E}_{\text{VAE}}(I_p), \qquad z_h = \mathcal{E}_{\text{VAE}}(I_h), \tag{3}$$

where $z_p, z_h \in \mathbb{R}^{C_v \times 1 \times h \times w}$. We construct the extended latent sequence as

$$\bar{x}_\tau = z_p \oplus_t x_\tau \oplus_t z_h \in \mathbb{R}^{C_v \times (T'+2) \times h \times w}, \tag{4}$$

where $\oplus_t$ denotes temporal concatenation. The product latent occupies the front boundary at $t = 0$, while the presenter latent occupies the rear boundary at $t = T' + 1$.

We define a slot indicator $M_s \in \{0,1\}^{1 \times (T'+2) \times h \times w}$, with ones at both boundaries and zeros at intermediate positions. M_s only marks the fixed boundary slots and remains unchanged under condition dropout; the product and presenter roles follow their fixed front–rear order. The indicator is concatenated with the extended latent along the channel dimension before patchification:

$$\tilde{x}_\tau = \text{Concat}_c(\bar{x}_\tau, M_s) \in \mathbb{R}^{(C_v+1) \times (T'+2) \times h \times w}. \tag{5}$$

The boundary positions serve only as non-generative conditions. If $v_\theta(\tilde{x}_\tau, C, \tau)$ denotes the DiT prediction over the extended sequence, we retain only its intermediate positions:

$$v_\theta^{\text{gen}} = \Pi_{\text{gen}}\left[v_\theta(\tilde{x}_\tau, C, \tau)\right], \tag{6}$$

where Π_{gen} selects positions $1, \ldots, T'$. Training loss is computed only over these positions. During sampling, only the intermediate video latents are updated by the denoising scheduler, while boundary latents z_p and z_h remain fixed. Only the resulting intermediate latents are passed to the VAE decoder. Consequently, the reference slots add no output frames, and the generated video retains the intended length of T decoded frames.

CLIP Semantic Context. The product and presenter images are independently processed by a CLIP vision encoder:

$$f_p = \text{CLIP}(I_p), \qquad f_h = \text{CLIP}(I_h). \tag{7}$$

Each embedding contains 257 tokens. After projection into the text-context dimension, the visual token sequences are concatenated with the T5 representation C_{txt} of the prompt:

$$C = \text{Proj}(f_p) \oplus \text{Proj}(f_h) \oplus C_{\text{txt}}. \tag{8}$$

Unlike the VAE boundary latents, CLIP embeddings do not occupy temporal or generative slots. They provide auxiliary semantic information exclusively through DiT cross-attention. Explicit role assignment is realized by the fixed latent boundary positions, while the CLIP pathway supplies additional semantic features. Fixed CLIP-token order and role-specific text also convey role information; boundary position is therefore not the only role cue.

Role identity and available evidence. The two boundary positions encode a stable product/presenter convention across all input modes. A missing image changes the evidence supplied at its position, while the indicator still identifies the position's role. The shared attention mechanism can therefore use both visual references when they are present and the text/video prior when one is absent. This separation supports a single input interface for all three modes without treating the retained reference as a different kind of entity. It also distinguishes role assignment from independent generation: the two references still interact through a common video model.

3.2 Mask-Guided Region-Weighted Loss

Standard flow-matching objectives uniformly penalize spatiotemporal locations, potentially underemphasizing product and facial regions critical for e-commerce video generation. We therefore introduce a mask-guided region-weighted loss to strengthen supervision on these regions.

For each target video frame, we independently obtain a product mask M_p and a face mask M_f using CLIP Surgery (Li et al., 2023) and SAM (Kirillov et al., 2023). Their intersection is defined after both masks are resized and binarized at the generated latent resolution. We construct the regional weight map as

$$W_{txy} = \begin{cases} \lambda_{\text{overlap}}, & M_p^{txy} = 1 \wedge M_f^{txy} = 1, \\ \lambda_p, & M_p^{txy} = 1 \wedge M_f^{txy} = 0, \\ \lambda_f, & M_p^{txy} = 0 \wedge M_f^{txy} = 1, \\ \lambda_{\text{bg}}, & M_p^{txy} = 0 \wedge M_f^{txy} = 0. \end{cases} \tag{9}$$

We set $\lambda_{\text{overlap}} = 3.0$, $\lambda_p = 2.0$, $\lambda_f = 2.0$, and $\lambda_{\text{bg}} = 0.5$. The overlap receives the highest weight because errors in these locations may simultaneously affect product appearance and facial identity.

Under Eq. 2, the target flow velocity is

$$u_\tau = \frac{\partial x_\tau}{\partial \tau} = \epsilon - x_0. \tag{10}$$

The region-weighted objective is computed only over the T' generated latent positions:

$$\mathcal{L}_{\text{RW}} = \mathbb{E}_{x_0,\epsilon,\tau}\left[\omega(\tau)\frac{\sum_{c,t,x,y} W_{txy}\left((v_\theta^{\text{gen}})^{ctxy} - u_\tau^{ctxy}\right)^2}{C_v \sum_{t,x,y} W_{txy} + \varepsilon}\right]. \tag{11}$$

Here, v_θ^{gen} denotes the prediction restricted to the intermediate generative positions, $\omega(\tau)$ follows the timestep-weighting schedule of the base model, and ε is a numerical stability constant. Normalization controls total weight across samples with different region sizes, retaining regional emphasis within a common loss scale. Appendix E details its behavior for disjoint, overlapping, and absent product or face regions, together with the role of mask quality.

Interpreting the normalized regional objective. Let a_b, a_p, a_f, a_o denote the fractions of mutually exclusive background, product-only, face-only, and overlap latent locations, summing to one. The average spatial weight and a region's share of total spatial weight are

$$\bar{W} = .5a_b + 2a_p + 2a_f + 3a_o, \qquad \pi_r = \frac{\lambda_r a_r}{\bar{W}}. \tag{12}$$

Product and face locations receive four times the per-location background weight; overlap receives six times that weight. The overlap is assigned weight 3 directly, rather than adding two weights of

2. For the illustrative composition $(.80, .10, .08, .02)$, the respective normalized weight shares are 48.78%, 24.39%, 19.51%, and 7.32%. These are consequences of the loss formula for a hypothetical mask, not measured mask frequencies. They show how relatively small reference regions can receive substantial supervision while background still contributes to overall scene quality.

With equal errors at every location, normalization makes the objective independent of mask size up to the numerical stabilizer. With unequal errors, the distribution of mask area changes which errors receive emphasis. Empty masks recover uniform weighting, whereas incorrect nonempty masks can emphasize the wrong content. This distinction explains the objective's intended benefit and why mask quality remains a separate consideration. Appendix E discusses the mask-to-latent alignment requirements.

3.3 Multi-Mode Dropout and Split-Noise Training

We introduce image condition dropout for multi-mode deployment and use a backbone-aligned fine-tuning schedule that preserves the noise-range specialization of the pretrained dual-expert DiT.

Image Condition Dropout. To support three input configurations with one model, we sample dual-reference, product-only, and human-only conditions with probabilities of 50%, 25%, and 25%, respectively, i.e., a ratio of 2:1:1. For dual-reference training, both boundary latents (z_p, z_h) and CLIP context tokens (f_p, f_h) are retained. For product-only training, z_h and f_h are zeroed and the face mask M_f is disabled; for human-only training, z_p and f_p are zeroed and the product mask M_p is disabled. The regional weight map W is then recomputed from the remaining masks, preventing absent conditions from affecting region-weighted supervision. This joint training uses one checkpoint for complete and partial inputs while retaining consistent product and presenter roles.

Backbone-Aligned Split-Noise Fine-Tuning. Wan2.2 employs two pretrained DiT experts specialized for different noise ranges. We preserve this existing specialization during fine-tuning rather than treating the two experts as a newly introduced architecture. Following Eq. 2, a larger τ corresponds to a higher noise level. The high-noise expert is fine-tuned only over $\tau \in [0.358, 1.0]$, focusing on global composition and macroscopic motion, while the low-noise expert is fine-tuned over $\tau \in [0, 0.358]$, focusing on local texture, product details, and facial appearance. Inference follows the original expert-switching boundary of the backbone.

3.4 Dataset Construction

The training unit combines two reference assets, a presentation script, and a generated demonstration clip. Figure 3 summarizes the curation path. Keeping reference provenance separate from target-video provenance clarifies the learning signal: real product or presenter photographs can condition a synthesized target, and the target still carries the generation model's visual prior.

4 Experiments

4.1 Experimental Setup

Evaluation Set and Implementation. We construct separate 150-case sets for the main comparison and ablation studies by independently pairing product and presenter images from a held-out asset pool. All evaluation products and presenter identities are excluded from training. For each pair, Qwen3.6-27B generates an e-commerce prompt describing presentation content and human–product interaction. Results average seeds 0, 42, and 100 (Appendix F) with shared references and prompts. Videos contain 81 frames at 720×560; unsupported methods use the closest compatible aspect ratio. DualCast is initialized from Wan2.2-I2V-14B (Wan et al., 2025). The high- and low-noise stages are trained separately for 8 hours each on 128 NVIDIA A100 GPUs using $720 \times 560 \times 81$ videos. Dual-reference, product-only, and human-only conditions are sampled at a 2:1:1 ratio. Region-loss weights are 0.5, 2.0, 2.0, and 3.0 for background, face, product, and product–face overlap, respectively.

Baselines. We compare four open-source multi-reference methods: Kaleido (Zhang et al., 2025), Phantom (Liu et al., 2025), VACE (Jiang et al., 2025), and SkyReels-A2 (Fei et al., 2025). We also evaluate four closed-source systems: Wan2.7 (Alibaba Cloud, 2026b), SkyReels-v4 (Chen et al.,

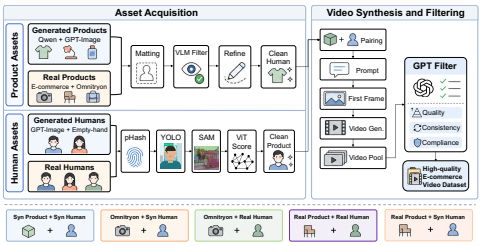

Figure 3: **Construction pipeline of the curated 10K e-commerce video dataset.** Product and presenter assets are collected or synthesized, quality-filtered, and paired to generate e-commerce-oriented prompts and initial frames. Candidate videos are subsequently filtered according to visual quality, temporal consistency, and prompt compliance. Product and face masks are precomputed for region-weighted training. Detailed source-use, filtering, privacy, and quality-control policies are provided in Appendix C.

We curate 10,000 clips from generated and real-world reference assets (Figure 3). GPT-Image-2 produces initial frames and Wan2.7 generates videos; GPT-5.5 filters candidates for visual quality, temporal consistency, and prompt compliance. The retained mixture is 3,000 synthetic product–presenter samples, 2,800 Omnitryon-product samples, and 4,200 real e-commerce-product samples, with both real and synthetic presenter references. Appendix C details acquisition, filtering, and the five-way composition. Real reference assets do not make the synthesized target clips real captured demonstrations. Evaluation products and presenter identities are excluded from training according to the reported asset split. That separation addresses direct identity reuse, while shared generators and filtering still induce a teacher/domain dependence. Wan2.7 is both a data generator and a comparator, so matching its scores is not evidence of independent distributional generalization.

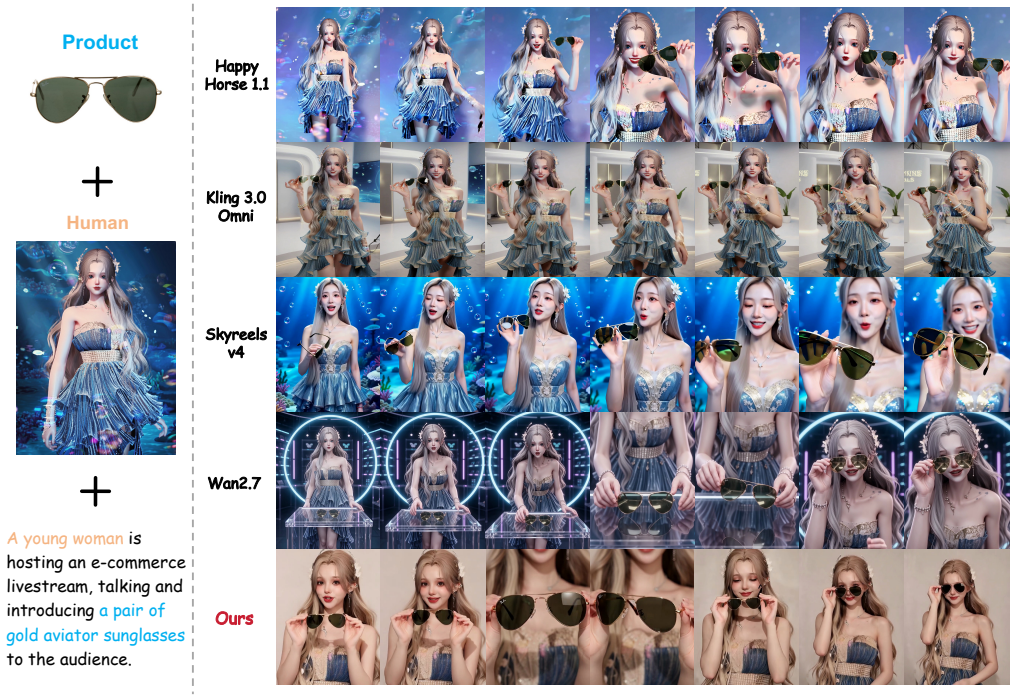

Figure 4: Selected qualitative comparison outputs. Product, face, and style differences are visible in individual cases. These panels do not estimate failure rates or measure style disentanglement.

2026), HappyHorse 1.1 (Alibaba Cloud, 2026a), and Kling3.0-Omni (Kling Team et al., 2025). All methods receive the same reference assets and role-specific prompts through supported interfaces (Appendix F). DualCast is domain fine-tuned; no matched-data, matched-compute fine-tuning of the comparison systems is established. Wan2.2-I2V is a backbone reference, with a conditioning interface that differs from the role-specific dual-reference configuration.

Evaluation Metrics. ViCLIP measures prompt–video alignment; S2V-C measures reference-to-video subject consistency using Grounding DINO (Liu et al., 2024), SAM (Kirillov et al., 2023), and CLIP. Qwen3.5-397B-A17B rates EcomStyle, FaceID, ProdID, Interaction, and Motion on a 0–5 scale normalized to $[0, 1]$ (Appendix B). Ablation ID-Ret instead uses ArcFace similarity and is

Table 1: **Quantitative comparison against state-of-the-art methods.** Metrics span semantic consistency, perceptual quality (0–1, ↑), and video quality (VBench). Open-source and closed-source methods are grouped separately; Avg Rank↓ is computed over all nine metrics across all methods, with ties assigned their average ranks. † denotes our base model. **Bold** indicates the best result; highlighted cells denote our top-ranked metrics.

Method	Open	Semantic & Consistency		Perceptual Quality (0–1)				Video Quality (VBench)			Avg Rank↓
		ViCLIP↑	S2V-C↑	EcomStyle↑	FaceID↑	Interaction↑	Motion↑	Dynamic Deg.↑	Aesthetic↑	Imaging↑	
Kaleido Zhang et al. (2025)	✓	0.220	0.516	0.878	0.891	0.826	0.896	0.500	0.499	0.726	8.39
Phantom Liu et al. (2025)	✓	0.227	0.493	0.912	0.916	0.877	0.904	0.650	0.510	0.751	6.83
VACE Jiang et al. (2025)	✓	0.210	0.596	0.904	0.939	0.888	0.922	0.750	0.513	0.755	5.61
SkyReels-A2 Fei et al. (2025)	✓	0.197	0.575	0.865	0.928	0.844	0.872	0.787	0.490	0.740	7.94
Wan2.2-I2V-14B† *(Base)* Wan et al. (2025)	✓	0.197	**0.717**	0.530	0.913	0.255	0.703	0.775	0.512	0.709	8.06
DualCast (Ours)	✓	**0.237**	0.596	**0.973**	**0.978**	**0.967**	**0.974**	**0.988**	**0.524**	0.751	**1.72**
Wan2.7 Alibaba Cloud (2026b)	×	0.219	0.563	0.944	0.952	0.938	0.949	0.963	**0.524**	0.759	3.78
SkyReels-v4 Chen et al. (2026)	×	0.220	0.573	0.926	0.935	0.923	0.932	0.975	0.507	0.735	5.56
HappyHorse 1.1 Alibaba Cloud (2026a)	×	0.216	0.593	0.949	0.957	0.940	0.952	0.963	0.514	0.753	3.50
Kling3.0-Omni Kling Team et al. (2025)	×	0.215	0.581	0.938	0.950	0.944	0.945	0.975	0.518	**0.767**	3.61

Table 1 reports feature/video metrics and judge ratings. DualCast has the highest ViCLIP (.237) and Dynamic Degree (.988), ties the highest aesthetic score (.524), and achieves S2V-C .596. Its S2V-C exceeds the strongest closed comparator by .003, ties VACE, and falls below the Wan2.2 backbone's .717. Imaging quality (.751) also trails several systems. These are distinct trade-offs rather than uniform dominance.

DualCast leads all four perceptual dimensions: EcomStyle .973, FaceID .978, Interaction .967, and Motion .974. Its margins over the strongest comparators are .024, .021, .023, and .022, respectively. These gains cover presentation style, presenter identity, product interaction, and motion. DualCast also achieves the best nine-metric average rank of 1.72, followed by HappyHorse 1.1 at 3.50 and Kling3.0-Omni at 3.61. ProdID is evaluated separately in the ablations. Perceptual ratings use six mid-clip frames, complemented by full-clip human assessment. Rank summarizes ordering, without measuring effect size or significance (Appendix D).

Figures 4 and 5 complement the numerical results with selected outputs, showing how presenter appearance, product visibility, and their temporal presentation contribute to the evaluated task.

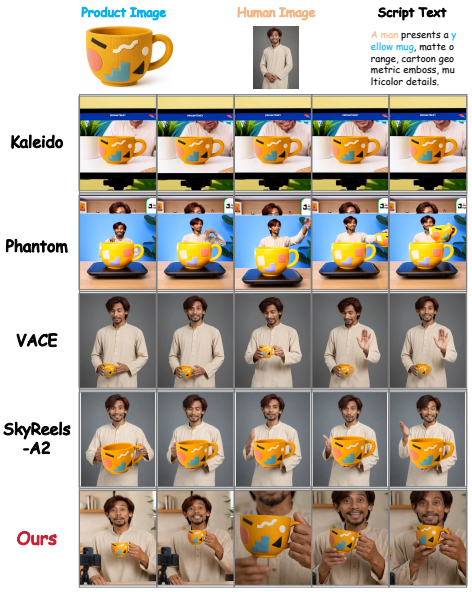

Figure 5: Selected temporal examples of presenter and product visibility. The frames illustrate why reference similarity is considered alongside role-specific perceptual ratings when assessing a two-subject presentation.

not interchangeable with the ordinal FaceID rating. Dynamic Degree measures motion prevalence; Aesthetic and Imaging Quality supply complementary visual measures, following VBench (Huang et al., 2024) and DOVER (Wu et al., 2023a). Appendix D gives their interpretation.

4.2 Comparison with Evaluated Systems

Table 1 compares prompt alignment, reference consistency, presentation judgments, and general video quality. The separate columns retain the distinction between retaining the supplied subjects and producing a coherent, visually plausible video of the requested presentation.

4.3 Ablation Study

Table 2 examines conditioning, supervision, and expert-training components. Table 3 evaluates the shared checkpoint with complete or partial inputs and the corresponding fidelity metrics.

Table 2: Component ablations on the separate evaluation set. ProdID is normalized to $[0, 1]$. Green cells show gains of the full model over each removal.

Model Variant	Condition Fidelity			Visual Quality		Avg.↑
	ID-Ret↑	S2V-C↑	ProdID↑	Aesthetic↑	Imaging↑	
DualCast (Full)	**0.511**	**0.589**	**0.953**	**0.517**	**0.748**	**0.664**
w/o Dual-End	0.070	0.508	0.943	**0.517**	0.743	0.556
Gain of Full	**+630.0%**	**+15.9%**	**+1.1%**	±0.0%	**+0.7%**	**+19.3%**
w/o Region-Loss	0.447	0.582	0.912	0.506	**0.748**	0.639
Gain of Full	**+14.3%**	**+1.2%**	**+4.4%**	**+2.2%**	±0.0%	**+3.8%**
w/o Two-Stage	0.349	0.570	0.935	0.502	0.737	0.619
Gain of Full	**+46.4%**	**+3.3%**	**+1.9%**	**+3.0%**	**+1.5%**	**+7.3%**

Table 2 reports ID-Ret .511 for the full model versus .070 without Dual-End conditioning. This contrast concerns the complete conditioning configuration; it does not isolate position from information or capacity. Removing regional weighting reduces ID-Ret to .447 and ProdID from .953 to .912, with Imaging unchanged at .748. High-noise-only training gives ID-Ret .349 and Imaging .737. The results favor the full recipe without proving that its weights or inherited expert configuration are optimal.

The joint model supports partial inputs (Table 3). Dual-reference inference improves ID-Ret over both single-mode-trained comparators, although its .953 ProdID is below the human-only-trained comparator's .978. Product-only and human-only inference yield .955 ProdID and .498 ID-Ret. Averages omit inapplicable metrics, so they do not rank different tasks on one common scale. Each reference role is present in 75% of training conditions; all three inference modes retain the same role convention. Reference availability selects the conditioning mode without changing slot roles, and the per-mode columns retain the criteria applicable to the supplied subject.

Table 3: **Ablation study on training and inference strategies.** We compare single-modality baselines with our joint training strategy under single- and dual-modal inference. EcomStyle and ProdID are normalized to $[0, 1]$ by dividing their raw scores by 5. Green highlights improvements over the baseline. † Means omit unavailable entries.

Training	Inference	ID-Ret↑	EcomStyle↑	ProdID↑	Avg.↑
Product-Centric Comparison					
Baseline (Product-Only)	Dual	0.426	0.975	0.943	0.781
Ours (Joint)	Product-Only	—	0.993	0.955	0.974†
Ours (Joint)	**Dual**	**0.511**	**0.988**	**0.953**	**0.817**
Δ vs. Baseline (Dual)		**+20.0%**	**+1.3%**	**+1.1%**	**+4.6%**
Human-Centric Comparison					
Baseline (Human-Only)	Dual	0.391	0.988	0.978	0.785
Ours (Joint)	Human-Only	0.498	0.995	—	0.747†
Ours (Joint)	**Dual**	**0.511**	0.988	0.953	**0.817**
Δ vs. Baseline (Dual)		**+30.7%**	±0.0%	−2.6%	**+4.0%**

Human assessment. Full-clip human judgments complement automatic and sampled-frame metrics. Table 4 reports preferences for product consistency, presenter identity, and video quality, followed by their Overall summary.

Table 4: **User-study preference rates (%).** Values above 50% favor DualCast; ties count as 0.5. Appendix A details the protocol.

Evaluation Metric	vs. Kaleido	vs. Phantom	vs. VACE	vs. SkyReels-A2
Product Consistency	63.6%	67.2%	52.0%	67.3%
Identity Preservation	76.8%	79.8%	60.2%	61.6%
Video Quality	71.4%	72.2%	67.7%	82.4%
Overall Preference	70.6%	73.1%	60.0%	70.4%

The 50 participants each assess 20 of 50 cases, producing 15,000 scalar scores and 1,000 participant–case comparisons per baseline and dimension. Overall preferences are 60.0–73.1%. Identity and video-quality preferences exceed 60% against every comparator; product preference is closest to parity against VACE. These descriptive rates differ from the protocol's signed-rank statistic over 50 participant means. Appendix A explains aggregation and dependence from shared cases.

5 Conclusion

DualCast adapts video diffusion to the distinct roles of a product and a presenter. Fixed boundary reference latents assign stable role addresses, region-weighted supervision emphasizes product and facial content, and reference dropout supports dual-image, product-only, and human-only generation with one checkpoint. DualCast leads all four perceptual dimensions, achieves the best nine-metric average rank of 1.72, and receives favorable human-preference ratings. Component and input-mode studies support the relevance of the conditioning and training configuration. The contribution is a unified treatment of role assignment, localized supervision, and partial reference inputs for this task.

Limitations and future work. The current setting uses short clips with one product and one presenter. Fine logos, occlusion, and long-duration identity preservation remain challenges; real-video evaluation and temporally dense fidelity measures are useful next steps.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and consistency checking. The authors take responsibility for the final manuscript, including its mathematical derivations, implementation descriptions, experimental values, citations, and claims.

ETHICS STATEMENT

Generating videos from product and presenter references raises risks of impersonation, misleading endorsements, and inaccurate product depiction. Deployment should require appropriate permission to use a person's likeness, respect product-image rights, and clearly disclose synthetic content. Our perceptual study reports aggregate preferences; participant identifiers should remain protected. The asset-processing and study protocols are described in Appendices C and A. Removing account metadata does not itself establish consent or a redistribution license for the underlying visual assets.

REFERENCES

Alibaba Cloud. HappyHorse Reference-to-Video API Reference. Alibaba Cloud Model Studio Documentation, 2026a. URL https://help.aliyun.com/zh/model-studio/happyhorse-reference-to-video-api-reference. Accessed: 2026-07-22.

Alibaba Cloud. Wan 2.7 Reference-to-Video API Reference. Alibaba Cloud Model Studio Documentation, 2026b. URL https://help.aliyun.com/zh/model-studio/wan-video-to-video-api-reference. Accessed: 2026-07-22.

Guibin Chen, Dixuan Lin, Jiangping Yang, Youqiang Zhang, Zhengcong Fei, Debang Li, Sheng Chen, Chaofeng Ao, Nuo Pang, Yiming Wang, et al. Skyreels-v4: Multi-modal video-audio generation, inpainting and editing model. *arXiv preprint arXiv:2602.21818*, 2026.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning. *arXiv preprint arXiv:2509.08519*, 2025.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. Magref: Masked guidance for any-reference video generation with subject disentanglement. In *International Conference on Learning Representations*, 2026.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, and Yahui Zhou. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. URL https://arxiv.org/abs/2504.02436.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.04725.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *ICLR workshop on deep generative models for highly structured data*, 2022.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. URL https://arxiv.org/abs/2311.17117.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17982.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. URL https://arxiv.org/abs/2503.07598.

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 *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2304.02643.

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. URL https://arxiv.org/abs/2512.16776.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *arXiv preprint arXiv:2304.05653*, 2023. URL https://arxiv.org/abs/2304.05653.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. *arXiv preprint arXiv:2502.07802*, 2025.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13847–13858, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. URL https://arxiv.org/abs/2502.11079.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2303.05499.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2212.09748.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, 2020. URL https://arxiv.org/abs/2008.10010.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.10752.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2208.12242.

Aliaksandr Siarohin, Stephane Lathuiliere, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/2003.00196.

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.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. *arXiv preprint arXiv:2211.04894*, 2023a. URL https://arxiv.org/abs/2211.04894.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *IEEE/CVF International Conference on Computer Vision*, 2023b. URL https://arxiv.org/abs/2212.11565.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. URL https://arxiv.org/abs/2308.06721.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023a. URL https://arxiv.org/abs/2302.05543.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b. URL https://arxiv.org/abs/2211.12194.

Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, Tiankun Cao, Cheng Wang, Xiaotao Gu, Jie Tang, Dan Guo, and Meng Wang. Kaleido: Open-sourced multi-subject reference video generation model. *arXiv preprint arXiv:2510.18573*, 2025. URL https://arxiv.org/abs/2510.18573.

A User Study Protocol

This section describes the human-evaluation protocol used for Table 4. We conduct a within-subjects, blind comparison between DualCast and four baselines: Kaleido, Phantom, VACE, and SkyReels-A2.

A.1 Participants

Fifty evaluators participated in the study. Each evaluator registered with a nickname that was converted into an anonymous hashed identifier. No personally identifiable information was collected. All evaluators received the same instructions and used an identical evaluation interface.

A.2 Stimuli and Procedure

The study uses a fixed pool of 50 cases selected before evaluation. Each case contains a product reference, a presenter reference, a text prompt, and videos generated by the five methods. All videos are normalized to 720×560 resolution, 81 frames, a five-second duration, and no audio.

Each evaluator is assigned 20 cases using a deterministic random seed derived from the anonymous identifier. This allows an evaluator to resume the same cases after re-entering the website. For each case, all five videos are displayed side by side and labeled only as "Video 1" through "Video 5." Their left-to-right order is independently randomized for each evaluator and case to reduce positional bias.

A.3 Rating Criteria

Evaluators independently rate every video on a five-point ordinal scale, where 1 denotes the worst performance and 5 denotes the best, according to the following dimensions:

1. **Product Consistency:** similarity between the generated product and the reference product;
2. **Identity Preservation:** identity consistency between the generated presenter and the reference portrait;
3. **Video Quality:** overall visual quality, motion naturalness, and the absence of visible artifacts.

The study collects 15,000 ratings in total: 50 evaluators $\times$ 20 cases $\times$ 5 methods $\times$ 3 dimensions.

A.4 Aggregation and Inferential Scope

For a given evaluator, case, dimension, and baseline, let s_{ours} and s_b denote the scores of DualCast and the baseline. The pairwise outcome is defined as

$$w = \begin{cases} 1, & s_{\text{ours}} > s_b, \\ 0.5, & s_{\text{ours}} = s_b, \\ 0, & s_{\text{ours}} < s_b. \end{cases} \tag{13}$$

The preference rate is the mean of w over all paired comparisons, multiplied by 100%. Thus, a rate above 50% favors DualCast, while ties contribute 0.5. Each baseline and evaluation dimension contains 1,000 evaluator–case comparisons for computing preference rates. The evaluation protocol specifies averaging the twenty paired score differences within each participant and applying a two-sided Wilcoxon signed-rank test to the fifty participant-level means. That analysis concerns score differences, whereas the displayed preference rate averages the sign-based w. A rate of 52% does not determine the signed-rank result. Shared cases induce an additional dependence across participants; the 15,000 scalar ratings cannot be used as an independent-trial sample size. Without participant/case records and test outputs, we retain the reported preference rates as descriptive summaries and do not attach significance stars or reconstruct confidence intervals.

B MULTIMODAL LLM-AS-A-JUDGE PROTOCOL

B.1 OVERVIEW

Embedding-based metrics measure feature-space similarity but cannot fully capture task-specific properties such as e-commerce presentation style, human–product interaction, or motion naturalness. We therefore employ Qwen3.5-397B-A17B as a multimodal judge. Given sampled video frames and the corresponding human and product references, the judge independently scores five dimensions on an integer scale from 0 to 5, where higher is better.

B.2 EVALUATION DIMENSIONS

EcomStyle. This metric evaluates whether the video follows the visual and presentational conventions of e-commerce broadcasting, including a camera-facing presenter, clear product presentation, and appropriate composition. A score of 5 indicates a convincing e-commerce presentation, 3 indicates weak selling intent, and 0 indicates an unrelated or invalid scene.

FaceID. This metric evaluates whether the generated presenter has the same facial identity as the human reference. The judge focuses on facial structure and proportions while ignoring clothing, hairstyle, accessories, background, pose, and minor illumination changes. Scores of 5, 3, and 0 indicate the same identity, an uncertain identity, and a different identity, respectively.

ProdID. This metric evaluates whether the generated product matches the reference in category, shape, color, texture, and branding or packaging, while allowing viewpoint and scale changes. Scores of 5, 3, and 0 indicate the same product, an uncertain match, and a different or missing product, respectively.

Interaction. This metric evaluates whether the presenter actively holds, points to, displays, or explains the target product. A score of 5 indicates a natural selling interaction, 3 indicates a weak association, and 0 indicates that the human and product are unrelated.

Motion. This metric evaluates temporal dynamics, including lip movement, gesture variation, pose changes, and the absence of freezing, jitter, or ghosting. Scores of 5, 3, and 0 indicate natural motion, limited motion, and nearly static or severely distorted motion, respectively.

B.3 JUDGING PROCEDURE

For each generated video, we uniformly sample six temporally ordered frames from the middle portion of the clip. The frames and two reference images are resized so that their longest side does not exceed 512 pixels and are JPEG-encoded. A single multimodal prompt contains the shared five-axis rubric, the sampled frames, the product reference, and the human reference. Method names are never included in the prompt.

The judge returns a JSON object containing five integer scores and a one-sentence justification. All axes are evaluated in one call but are explicitly instructed to be scored independently. Invalid responses are retried up to five times. Parsed scores are clamped to $[0, 5]$, and a missing or unparseable score is assigned 0.

B.4 AGGREGATION AND FAIRNESS CONTROLS

For method m and evaluation axis a, the normalized score is

$$S_{m,a} = \frac{1}{5N} \sum_{i=1}^{N} s_{m,a}^{(i)}, \tag{14}$$

where $s_{m,a}^{(i)} \in [0, 5]$ is the judge score for case i and $N = 150$ within each run. We average the resulting run-level scores over the three seeds in Appendix F. The main comparison and ablation study use separate held-out case sets. Dividing by 5 maps the reported results to $[0, 1]$ while preserving the ordering between methods. A missing video receives 0 on all axes and remains in the average. For single-reference inference, reference-dependent axes are marked as inapplicable and excluded from the corresponding average.

All methods are evaluated using the same case set, frame-sampling procedure, input resolution, prompt template, scoring rubric, and judge model. The judge is blind to method identity, and the rubric requests separate axis scores. Co-judging keeps the visual input and assessment context shared across axes; statistical dependence and visual/domain bias remain considerations when interpreting the resulting ratings.

C Dataset Construction Details

C.1 Asset Sources

We construct the dataset using both generated and real-world product and presenter assets. For products, Qwen3.6-27B prompts GPT-Image-2 to synthesize 1,500 images, while real product assets are collected from e-commerce platforms and Omnitryon. For presenters, GPT-Image-2 generates 1,500 empty-hand portraits, which are supplemented with filtered real-world portraits.

C.2 Asset Filtering

Product images are processed using BiRefNet for foreground extraction and a VLM-based filter covering eight quality dimensions. Their masks and boundaries are subsequently corrected manually. Human images undergo pHash deduplication, YOLOv8-based single-subject filtering, SAM-based background assessment, and aesthetic filtering using a SCUT-FBP5500-finetuned ViT.

C.3 Video Synthesis and Quality Control

For each product–presenter pair, Qwen3.6-27B generates an e-commerce prompt describing spatial layout and temporal dynamics. GPT-Image-2 synthesizes the initial frame, and Wan2.7 generates the corresponding video. GPT-5.5 then filters candidate videos according to visual quality, temporal consistency, and prompt compliance. Product and face masks are precomputed for region-weighted training.

C.4 Dataset Composition

The final dataset contains 10,000 videos: 3,000 synthetic samples, 800 Omnitryon products paired with synthetic presenters, 2,000 Omnitryon products paired with real presenters, 3,000 real products paired with real presenters, and 1,200 real products paired with synthetic presenters.

C.5 Source Use and Privacy

Real-world assets are stored using anonymous internal identifiers without associated account names or unrelated metadata. Only visual content and derived masks required for model training are retained. The training and evaluation asset pools are strictly separated at both the product-instance and presenter-identity levels.

D Evaluation Definitions and Score Interpretation

Different fidelity measurements. ID-Ret measures identity through ArcFace-based embedding similarity; FaceID is an integer Qwen judgment normalized by five. Embedding similarity and semantic judgments respond differently to pose, crop quality, and occlusion. They provide complementary identity diagnostics, and neither score is a calibrated probability of correct identity.

S2V-C uses detected/segmented subjects and CLIP features. It can be high when one salient reference dominates; the backbone's .717 therefore does not by itself establish superior retention of both entities. Product and presenter contributions depend on the reference-combination and frame-pooling rules. ViCLIP measures prompt/video alignment, while aesthetic and imaging measures address other visual properties. Dynamic Degree rewards detectable motion; larger values need not imply better motion or fewer artifacts.

Perceptual dimensions and aggregation. The five-axis judge protocol evaluates style, face, product, interaction, and motion together. Table 1 reports style, face, interaction, and motion; ProdID appears in the ablations on their separate evaluation set. DualCast ranks first in all four main-table perceptual dimensions. Normalization by five preserves each dimension's ordering but does not calibrate different dimensions to a common unit. Their separate columns show which aspects of presentation improve; an average over different sets of applicable dimensions would describe different tasks.

For a method m and metric j, the tie-adjusted rank is $r_{mj} = 1 + \#\{k : s_{kj} > s_{mj}\} + \frac{1}{2}\#\{k \neq m : s_{kj} = s_{mj}\}$. Equal weighting gives $R_m = \frac{1}{9}\sum_j r_{mj}$. Average rank summarizes relative ordering across the nine metrics; individual scores retain the magnitude and direction of each quality difference.

Temporal coverage. Six ordered frames, resized to at most 512 pixels on the long side, may omit brief identity changes, early reference copy-through, jitter, or small text. A high sampled-frame score therefore leaves those unobserved portions unconstrained. Identical prompts and frame rules improve comparability without establishing human validity; the seed check in Appendix F addresses run variability, while cross-judge agreement and frame-sampling coverage are separate questions.

E Conditioning and Regional-Loss Analysis

Role convention and missing references. The convention is product at the front and presenter at the rear. Intermediate video positions alone are supervised, updated, and decoded. The slot indicator remains one at both boundary positions even when a missing reference latent is zeroed. Product-only mode zeros presenter VAE and CLIP features and disables the face mask; human-only mode does the converse. Thus slot identity and reference availability are distinct. Conditional dropout does not swap the role convention.

With probabilities $(.50, .25, .25)$, each role is present with probability .75, both are present with probability .50, and at least one reference is always present. The high-/low-noise split at .358 preserves the backbone's experts and is not an additional learned routing mechanism.

Regional weights and normalization. Equation 12 gives each region's share of total spatial weight, not its realized share of the loss unless regional errors are equal. Empty overlap contributes no weight. If both masks are empty, all locations have weight .5 and normalization recovers the uniform loss up to the stabilizer. Incorrect nonempty masks instead direct weight toward incorrect pixels; normalization cannot correct them.

Frame masks must be mapped to the same temporal/spatial lattice as the video VAE latents before applying this equation. Spatial resizing alone does not specify how temporally compressed positions combine frame masks. Temporal pooling and binarization determine which frame regions contribute to each latent position; the spatial weight formula alone does not resolve that temporal alignment. No hand mask is used; hand–product contact is supervised only insofar as it falls within the existing regions and video prior.

Analytical overhead. For T' latent frames with P spatial tokens each, two reference frames increase sequence length from $T'P$ to $(T'+2)P$, a factor $1+2/T'$. A dense self-attention term scales by $(1 + 2/T')^2$ under the same tokenization, and the semantic context adds $2 \times 257 = 514$ visual tokens. The extra indicator changes patch-projection input width. These are structural counts, not measured latency, peak memory, or throughput; caching and attention kernels change realized cost.

F Experimental Configuration and Comparison Scope

Random seeds and result aggregation. Numerical experimental results in the main-text tables are averaged over runs with seeds 0, 42, and 100. Main-comparison and ablation results retain their separate 150-case evaluation sets. Seed 1000 supplies an additional stability check and is excluded from the reported means. The results remain stable across the tested seeds. Average rank is computed from the reported mean metric values using the tie rule in Appendix D; it is distinct

from averaging ranks computed independently within each run. Sample counts, training settings, and model specifications are fixed descriptors rather than averaged scores.

Table 5: Training and evaluation settings reported for DualCast.

Item	Reported setting
Initialization	Wan2.2-I2V-14B; existing high-/low-noise experts
Training data	10,000 curated e-commerce clips
Video tensor	720×560, 81 frames
Compute schedule	128 NVIDIA A100 GPUs; 8 hours per expert stage
Noise intervals	High $[.358, 1]$; low $[0, .358]$
Input modes	Dual/product/human probabilities .50/.25/.25
Loss weights	Background .5; product 2; face 2; overlap 3
Semantic context	257 CLIP tokens per reference, plus T5 text
Judge inputs	Six mid-clip frames; long side at most 512 pixels
Human study	50 participants; 50-case pool; 20 cases per participant

Attribution of the existing comparisons. The cross-system comparison combines reference interfaces, pretrained priors, training distributions, and inference implementations. Shared reference assets and prompts do not equalize these factors; native aspect-ratio fallbacks further change the pixel workload. The backbone row identifies a starting point, while the component removals compare configurations within the adapted pipeline. The "w/o Dual-End" label does not identify which latent features and capacities are retained, so its large identity delta is not a positional-effect estimate. Likewise, high-noise-only training changes exposure to the backbone experts; no optimality of .358 or the loss weights follows.

The data pipeline uses Wan2.7 as teacher and comparator. Excluding product instances and presenter identities prevents direct reference reuse under the stated split, but cannot eliminate teacher style, generator artifacts, or filtering preferences. Evaluation on real reference photos is also distinct from evaluation against real captured demonstration targets. Seed stability under this setting does not by itself establish robustness to new domains.

Qualitative and failure scope. The illustrations show selected cases. Aggregate scores are stable across the tested seeds, but a mouth deformation in one comparator output remains an instance-level observation. Product text, small packaging details, occluded faces, fingers, fast manipulation, and reference copying at clip boundaries can all escape sparse-frame evaluation. No frequency is assigned to these failure modes. Source anonymization removes associated account metadata but does not itself establish licensing or consent for every underlying visual asset.