FASTINTERACT: FEW-STEP HUMAN-OBJECT INTERACTION VIDEO GENERATION VIA ATTENTION PERTURBATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step distillation accelerates video diffusion models, but preserving coherent human–object interaction (HOI) while maintaining reference product consistency remains challenging. We introduce **A**ttention **P**erturbation (**AP**) to provide interaction guidance for distribution matching distillation (DMD). AP compares teacher predictions by selectively enabling or blocking target-video queries' access to conditioning keys and values, with inputs and all other attention connections fixed. An auxiliary HOI representation predicted by the student supplies an interaction-enhancement response, and a second contrast using product-reference attention provides a preservation direction. We coordinate these responses through a one-sided projection that removes the component of the enhancement response opposing the preservation direction before augmenting the DMD target. We further curate a multimodal HOI video dataset ***InteractVerse*** from product-oriented livestreams, with aligned person and product references, text, audio, motion, and spatial annotations. Experiments and ablations demonstrate that our method improves interaction plausibility and reference consistency over vanilla DMD.



Figure 1: Comparison of the 40-step teacher (top) and our 4-step student (bottom) for HOI video generation, conditioned on the same person and product references, audio, and text. Despite using only four sampling steps, our student better preserves product identity and visual details, including printed text, with colors more faithful to the product reference.

1 INTRODUCTION

Recent video diffusion model advances enable increasingly realistic audio-driven digital humans (Chen et al., 2025; Yang et al., 2025), potentially supporting livestream commerce, narrated presentations, and remote education. Yet lifelike appearance and accurate lip synchronization

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.

alone cannot satisfy product demonstrations requiring coherent human–object interactions (Xu et al., 2026). This makes human–object interaction (HOI) video synthesis important for advancing digital human generation. Such synthesis requires coordinated hand movements, precise object manipulation, and physically plausible contact (Liu et al., 2025a; Luo et al., 2026). High inference latency substantially hinders practical deployment. Video diffusion transformers' long spatiotemporal token sequences make repeated denoising network forward passes computationally expensive (Team Wan et al., 2025). Higher-order numerical solvers can reduce network evaluations without retraining, but numerical discretization errors may limit sampling accuracy at very low evaluation budgets (Lu et al., 2022). Distillation provides a complementary route, learning to synthesize high-quality outputs in few steps. Distribution matching distillation (DMD) trains a student to match a multi-step teacher's output distribution (Yin et al., 2024b;a). Recent avatar distillation methods address degraded lip synchronization and facial dynamics (Huang et al., 2026) and identity drift during long video generation (Huang et al., 2025a). Models have begun combining object manipulation with few-step synthesis (Meituan LongCat Team et al., 2026). Nevertheless, explicitly preserving hand–object contact and interaction structure during distillation remains underexplored, despite being essential for extending efficient avatar generation to convincing product demonstrations. Preserving HOI during few-step distillation is challenging: interaction fidelity depends on precise contact geometry and coordinated hand and object motion. Contact cues localized around fingers and object boundaries must remain consistent as poses, viewpoints, and occlusions change. Standard DMD matches the student's video distribution to the teacher's (Yin et al., 2024b;a), but does not explicitly prioritize these interaction-specific relationships. With generation compressed into few steps, the student may retain plausible human and product appearances without preserving their spatiotemporal coupling. Related failures in HOI video synthesis include unstable hand structure and physically implausible hand–object interpenetration (Luo et al., 2026). These failures motivate explicit interaction guidance during distillation to retain structural relationships underlying plausible HOI.

To address this challenge, we introduce *attention perturbation* (AP) to construct teacher responses to selected conditioning information. Inspired by PAG (Ahn et al., 2024), AP compares the teacher's video predictions with attention from target-video queries to the keys and values of selected conditioning tokens enabled or blocked, while keeping its weights, inputs, noisy states, and timestep fixed. We build on CoInteract (Luo et al., 2026), which jointly trains an RGB video and auxiliary HOI structure streams. The latter represents human–object spatial relationships using human silhouettes and object appearance. In our distillation framework, the student predicts this representation alongside the video, and AP perturbs attention from target-video queries to HOI conditioning keys and values derived from the student's prediction. The resulting prediction difference provides an interaction-enhancement response, which we use to augment the DMD teacher target. Interaction enhancement must also preserve the product being demonstrated. A response that strengthens interaction structure can also alter product appearance, creating a potential conflict with reference consistency. We therefore construct a second AP response by contrasting teacher predictions with attention from target-video queries to product-reference tokens enabled or blocked. Using this response as a preservation direction, we apply a one-sided projection that removes the component of the HOI response opposing it in prediction space. The coordinated response then augments the distillation target, reducing the conflict between interaction enhancement and product consistency.

We further curate a multimodal HOI video dataset from product-oriented livestreams. It aligns interaction clips with separate person and product references, text, audio, motion and pose representations, and spatial annotations, providing a resource for controllable HOI video generation (Sec. 3).

Our contributions are threefold:

- We introduce AP to construct interaction-enhancement responses by perturbing the target video's attention to HOI conditioning tokens, and use these responses to augment DMD for four-step HOI video generation.
- We develop a one-sided projection coordinating interaction enhancement with product consistency by removing conflicting components from the interaction-enhancement response.
- We curate a large-scale, interaction-complete HOI video dataset with aligned semantics, disentangled person and product references, temporal dynamics, and interaction-sensitive spatial supervision.

2 Related Work

2.1 Human-centric Video Generation

Reference-based avatars target identity-preserving head animation (Zhang et al., 2023; Xia et al., 2026), while audio-driven portrait methods improve speech synchronization and expressive motion through audio–visual alignment (Xu et al., 2024; Wang et al., 2025). Video diffusion extends human animation to full-body and long-form generation (Lin et al., 2025; Gao et al., 2025; Yang et al., 2025; Meituan LongCat Team et al., 2026). HuMo (Chen et al., 2025) combines text, appearance, and audio for flexible human-centric generation. Product demonstrations require stable hand–object contact and faithful product appearance, motivating HOI modeling.

HOI-Swap (Xue et al., 2024) replaces objects in existing interaction videos while adapting hand motion. AnchorCrafter (Xu et al., 2026) and HunyuanVideo-HOMA (Huang et al., 2025b) use pose, geometry, or trajectory conditions; Phantom (Liu et al., 2025d) supports subject-consistent reference-based generation. InteractAvatar (Zhang et al., 2026b) combines perception, interaction modeling, and audio-conditioned generation for text-directed interactions. CoInteract (Luo et al., 2026) jointly trains RGB and auxiliary HOI streams, using human silhouettes and object appearance to supervise spatial relationships and contact. We build on this to guide few-step distillation.

2.2 Distribution Matching Distillation

DMD (Yin et al., 2024b) accelerates diffusion generation, matching student and teacher distributions through score differences at noise levels. DMD2 (Yin et al., 2024a) removes paired regression and stabilizes training with two-timescale updates and real-data adversarial supervision. Decoupled DMD (Liu et al., 2025b) separates classifier-free guidance augmentation from distribution-matching regularization, motivating task-specific guidance. For streaming avatars, LiveAvatar (Huang et al., 2025a) and DynaForcing (Huang et al., 2026) address identity drift and degraded facial dynamics; SoulX-FlashTalk (Shen et al., 2025) combines bidirectional chunk attention with retrospective correction. StreamChar (Tian et al., 2026) separates step compression from online rollout training; TurboTalk (Liu et al., 2026) distills four-step into one-step generation. AptAvatar (Zhang et al., 2026a) anchors two-step distillation to a frozen four-step generator with self-generated history replay; JoyStreamer (Wang et al., 2026) uses twin-teacher DMD and timestep-dependent CFG for text control and audio–visual synchronization. We prioritize preserving HOI structure and product consistency during distillation. Inspired by Perturbed-Attention Guidance (Ahn et al., 2024), we contrast teacher predictions by selectively modifying target-video attention to HOI and product-reference tokens. A one-sided projection coordinates these responses before DMD target augmentation, transferring interaction guidance into a four-step student.

3 InteractVerse: An Interaction-Complete HOI Video Dataset

3.1 Motivation and Design Goals

Existing HOI video datasets often trade off scale for supervision richness. As summarized in Table 1, HOIGen-1M (Liu et al., 2025c) provides over one million HOI videos with sample-specific descriptions, but limited explicit supervision for participant appearance, temporal dynamics, and interaction structure. In contrast, controllable HOI datasets offer richer conditioning at a much smaller scale. AnchorCrafter-400 (Xu et al., 2026) includes person and object references with temporal and spatial guidance, while SparseHOI-5K (Xie et al., 2026) further includes synchronized audio and explicit hand–object observations. However, these datasets contain only hundreds to several thousand curated interactions and do not jointly provide sample-specific semantics, disentangled appearance references, and aligned audio, temporal, and interaction-sensitive spatial supervision.

We refer to this unified representation as *interaction completeness*. Rather than representing an HOI sample only as an RGB sequence paired with text or a few control signals, an interaction-complete sample describes the same interaction from four complementary aspects: *interaction semantics*, *participant appearance*, *temporal dynamics*, and *interaction-sensitive spatial structure*. These factors characterize what interaction occurs, who and what participate, how it evolves over time, and where

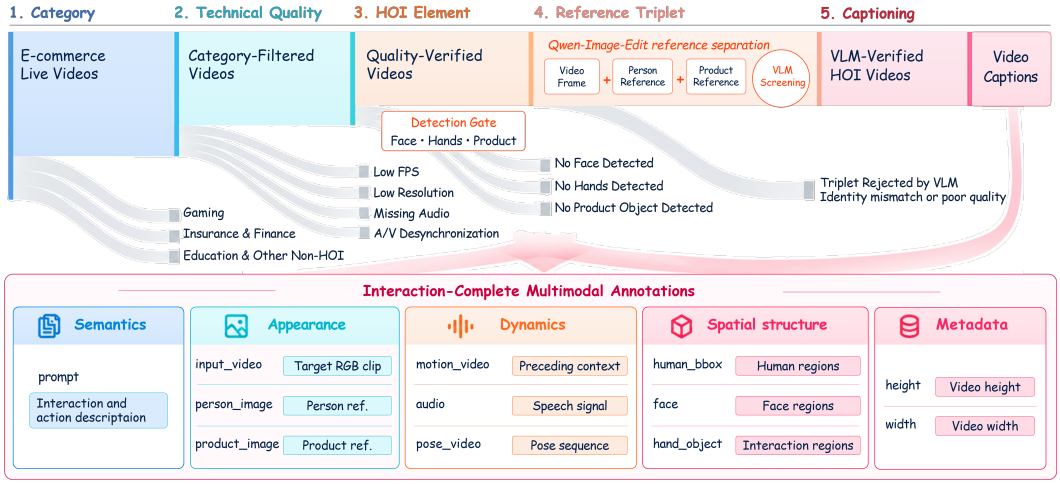

Figure 2: Overview of the construction and annotation of InteractVerse. Each HOI event is represented by aligned semantic, appearance, temporal, and spatial supervision, while a coarse-to-fine curation process ensures interaction relevance and reference consistency.

Table 1: Comparison with representative HOI video datasets. Person Ref. and Object Ref. denote explicit appearance references; Temporal denotes auxiliary temporally aligned motion signals beyond the target RGB video; and Spatial denotes explicit interaction-sensitive spatial supervision.

Dataset	Videos	Dur. (h)	Caption	Person Ref.	Object Ref.	Audio	Temporal	Spatial
HOIGen-1M	>1M	2.2K	✓	–	–	–	–	–
AnchorCrafter-400	356	2	–	✓	✓	–	✓	✓
SparseHOI-5K	4,850	10	–	✓*	✓*	✓	✓	✓
InteractVerse	**37,477**	**64.1**	✓	✓	✓	✓	✓	✓

*SparseHOI-5K derives appearance cues from an object-free subject reference and object-region observations, rather than providing a separately constructed person–product reference pair.

the interaction takes place. Such a representation is particularly important for controllable HOI generation, where a model must preserve reference appearance while reproducing plausible motion.

Following this principle, we construct InteractVerse, an interaction-centric dataset collected from real-world e-commerce livestreams. InteractVerse contains **37,477** curated HOI clips totaling approximately **64.1 hours**. Each sample forms a temporally continuous interaction segment paired with a sample-specific textual description, synchronized audio, disentangled person and product references, human pose sequences, and interaction-aware spatial annotations. These modalities are factorized while remaining aligned to the same HOI event. They provide explicit supervision for appearance preservation, temporal evolution, audio–visual correspondence, and spatially grounded interaction modeling within a unified sample. As shown in Table 1, InteractVerse combines tens of thousands of curated interactions with broad supervision coverage, bridging large-scale caption-centric datasets and smaller datasets designed for fine-grained HOI control. We will release InteractVerse and its annotations to facilitate future research on controllable HOI video generation.

3.2 Dataset Construction and Multimodal Annotations

As illustrated in Fig. 2, we curate product-centered livestream clips with visible presenters, hands, and products, filtering poor technical quality, weak HOI content, and audio–visual desynchronization. For each clip, Qwen-Image-Edit extracts separate person and product references, which a vision-language model (VLM) verifies against the source frame. We align interaction captions with synchronized audio, preceding motion context, human pose trajectories, and body, face, and hand–object regions. All modalities describe the same continuous HOI event and are independently accessible. Curation and annotation details are provided in Appendix B.

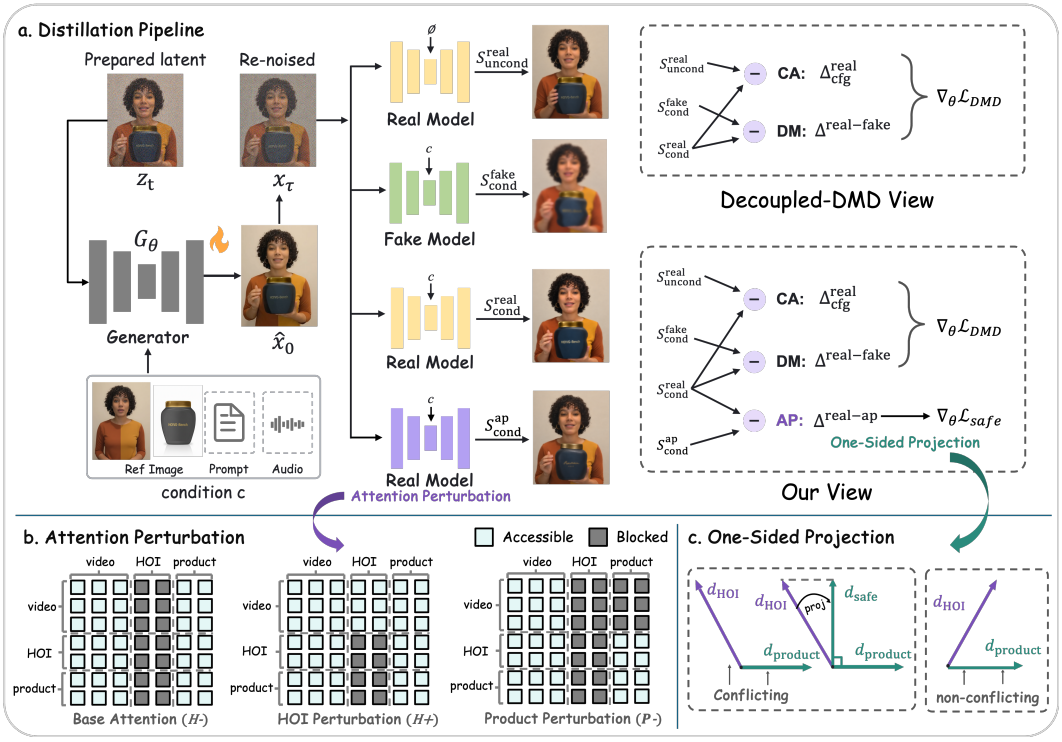

Figure 3: Overview of AP-augmented distillation for HOI video generation. (a) Distillation pipeline combining distribution matching, CFG augmentation, and AP guidance. (b) Attention perturbations for HOI enhancement and product-reference preservation. (c) One-sided projection coordinating the enhancement and preservation responses.

4 METHOD

We augment distribution matching distillation (DMD) with attention perturbation (AP), coordinating enhancement and preservation responses through a one-sided projection. Figure 3 illustrates the framework and its instantiation for HOI video generation.

4.1 REVISITING DMD AND ITS DECOUPLED FORMULATION

Following DMD2 (Yin et al., 2024a), let $\hat{x}_0 = G_\theta(z, c)$ be the student's clean target-video latent prediction with conditioning c and base randomness z. Re-noising gives

$$x_t = F(\hat{x}_0, t; \epsilon) = \alpha_t \hat{x}_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \alpha_t, \sigma_t > 0. \tag{1}$$

The noised teacher and student densities have scores $s_{\text{real}} = \nabla_{x_t} \log p_{\text{real},t}$ and $s_{\text{fake}} = \nabla_{x_t} \log p_{\text{fake},t}$. Conditioning and $p_{\text{fake},t}$'s dependence on θ are omitted. For weights $w_t \geq 0$, the DMD objective (Yin et al., 2024b) $\mathcal{L}_{\text{DMD}} = \mathbb{E}_{t,c}[w_t D_{\text{KL}}(p_{\text{fake},t} \| p_{\text{real},t})]$ has gradient

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}\left[w_t \alpha_t J_\theta^\top (s_{\text{real}} - s_{\text{fake}})\right], \qquad J_\theta = \frac{\partial G_\theta(z, c)}{\partial \theta}. \tag{2}$$

Under the assumptions in Appendix C, this identity holds for exact scores, estimated by a frozen teacher and an online denoiser trained on student samples.

Classifier-free guidance (CFG) (Ho & Salimans, 2022) gives the teacher score $s^{\text{cfg}}_{\text{real}} = s_{\text{real}} + (\gamma - 1)(s_{\text{real}} - s^-_{\text{real}})$, where $\gamma \geq 1$ and s^-_{real} uses reference condition c^-. Following Decoupled DMD (Liu et al., 2025b), keeping the fake score unguided yields

$$s^{\text{cfg}}_{\text{real}} - s_{\text{fake}} = \underbrace{s_{\text{real}} - s_{\text{fake}}}_{\Delta_{\text{DM}}} + (\gamma - 1) \underbrace{(s_{\text{real}} - s^-_{\text{real}})}_{\Delta_{\text{CA}}}. \tag{3}$$

Decoupled DMD interprets CFG augmentation (CA) as the driving term and distribution matching (DM) as a stabilizing regularizer. We retain DM and add AP guidance: whereas CA changes conditioning inputs, AP changes attention access to conditioning tokens while keeping inputs fixed.

4.2 Attention Perturbation

PAG (Ahn et al., 2024) and SAG (Hong et al., 2023) construct guidance from internal attention perturbations and attention-selected input degradation, respectively. In contrast, our AP selectively enables or blocks target-video attention to specific conditioning keys and values while keeping all inputs fixed. The prediction difference provides guidance for distillation.

Let $\mathcal{I}_{\text{vid}}$ index the target-video queries shared by all AP responses, and $\mathcal{I}^k_{\text{cond}}$ the conditioning tokens for pathway k. These may encode external conditions or auxiliary predictions. At each intervened layer and head, the blocked teacher pass applies the additive mask

$$(M_k)_{ij} = \begin{cases} -\infty, & i \in \mathcal{I}_{\text{vid}},\ j \in \mathcal{I}^k_{\text{cond}}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

before softmax, in addition to the visibility rules shared by both passes. The enabled pass omits only M_k; all other restrictions still apply. Both passes share teacher weights, inputs, noisy states, timestep, and all other visibility rules, retaining all tokens while recomputing hidden states. Each affected query retains visible keys over which attention is renormalized.

Let $\mu^{k,+}_{\text{real}}$ and $\mu^{k,-}_{\text{real}}$ be the teacher's clean target-video latent predictions with access to condition k allowed and blocked by M_k, respectively. Let μ_{real} and μ_{fake} denote the clean target-video latent predictions of the teacher under base attention and of the online denoiser. The score conversion is

$$s_{\text{real}} = \frac{\alpha_t \mu_{\text{real}} - x_t}{\sigma_t^2}, \qquad s_{\text{fake}} = \frac{\alpha_t \mu_{\text{fake}} - x_t}{\sigma_t^2}. \tag{5}$$

Applying the same conversion to both pathway configurations defines

$$d_k = \mu^{k,+}_{\text{real}} - \mu^{k,-}_{\text{real}}, \qquad \Delta^k_{\text{AP}} = s^{k,+}_{\text{real}} - s^{k,-}_{\text{real}} = \frac{\alpha_t}{\sigma_t^2} d_k. \tag{6}$$

With enhancement weight $\lambda \geq 0$, AP augments the teacher score:

$$s^{\text{aug}}_{\text{real}} = s^{\text{cfg}}_{\text{real}} + \lambda \Delta^k_{\text{AP}}, \tag{7}$$

$$s^{\text{aug}}_{\text{real}} - s_{\text{fake}} = \Delta_{\text{DM}} + (\gamma - 1)\Delta_{\text{CA}} + \lambda \Delta^k_{\text{AP}}. \tag{8}$$

$\gamma - 1$ and λ independently control CFG augmentation and AP augmentation. Under the exact-score and normalizability assumptions in Appendix C.2, this unprojected score admits a density-ratio interpretation as $\nabla_{x_t} \log p_{\text{aug},t}$ at fixed conditions and auxiliary inputs.

4.3 One-Sided Projection

AP can affect multiple video attributes. We use two conditioning token groups to obtain enhancement and preservation responses $h = d_k$ and $r = d_\ell$. Following gradient surgery (Yu et al., 2020), we remove the component of h opposing r:

$$\mathcal{P}_r(h) = \operatorname*{arg\,min}_{v:\, \langle v, r \rangle \geq 0} \frac{1}{2}\|v - h\|_2^2 = \begin{cases} h, & \langle h, r \rangle \geq 0, \\ h - \dfrac{\langle h, r \rangle}{\|r\|_2^2} r, & \langle h, r \rangle < 0. \end{cases} \tag{9}$$

We compute the projection per sample over the full target-video latent tensor, retaining h when either response is near zero or r fails a reliability test (Appendix C.3). The result h_{safe} satisfies $\langle h_{\text{safe}}, r \rangle \geq 0$ when projection is used; this constrains the added response, without guaranteeing preservation of an output attribute by the full update. In clean VAE-latent space, the coordinated target and update direction are

$$\begin{aligned} \mu_{\text{target}} &= \mu_{\text{real}} + (\gamma - 1)(\mu_{\text{real}} - \mu^-_{\text{real}}) + \lambda h_{\text{safe}}, \\ d_{\text{total}} &= \mu_{\text{target}} - \mu_{\text{fake}}, \end{aligned} \tag{10}$$

where μ^-_{real} is the CFG reference prediction. The preservation response determines the projection direction and is not separately added to the target. We use d_{total} for the DMD update; implementation details are provided in Appendix C.4.

4.4 HOI INSTANTIATION

We instantiate AP on CoInteract (Luo et al., 2026), using HOI tokens for enhancement and product-reference tokens for preservation. The student jointly predicts a clean RGB video latent and an auxiliary HOI latent. We detach the HOI prediction and re-noise it at the RGB score timestep using independent noise. This supplies the teacher's HOI input.

As shown in Figure 3(b), *Base Attention* (H−) allows target-video queries to attend to all tokens except HOI tokens. Relative to this configuration, *HOI Perturbation* (H+) additionally allows access to HOI tokens, whereas *Product Perturbation* (P−) blocks access to product-reference tokens. All other attention connections remain unchanged. The teacher passes use identical RGB and HOI states; their clean target-video predictions define

$$h = \mu_{\text{real}}^{\text{H+}} - \mu_{\text{real}}^{\text{H-}}, \qquad r = \mu_{\text{real}} - \mu_{\text{real}}^{\text{P-}}, \tag{11}$$

where $\mu_{\text{real}}^{\text{H-}} = \mu_{\text{real}}$. We coordinate h with r using the projection in Sec. 4.3 (Figure 3(c)) and add the resulting response to the DMD teacher target. Details are provided in Appendix C.4.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTING

We train the four-step student on our curated InteractVerse dataset (Sec. 3). We set the teacher CFG scale to $\gamma = 5$ and the HOI AP weight to $\lambda = 1$, applying target-query AP and one-sided projection as described in Sec. 4.4. The qualitative coefficient probe in Appendix D.2 shows that increasing the HOI AP coefficient on Vanilla DMD improves interaction quality, while larger coefficients introduce visible artifacts and product consistency decreases across the sweep.

Baselines. We evaluate on HOIVG-Bench from OmniShow (Zhou et al., 2026). Many-step baselines are Phantom (Liu et al., 2025d), VACE (Jiang et al., 2025), HuMo (Chen et al., 2025), and CoInteract (40–50 steps); few-step baselines are LiveAvatar (Huang et al., 2025a), FlashTalk (Shen et al., 2025), SkyReels-V3 (Li et al., 2026), and LongCat-Video-Avatar 1.5 (Meituan LongCat Team et al., 2026) (4–8 steps). SkyReels-V3 uses reference-to-video generation. For LiveAvatar, FlashTalk, and LongCat, Qwen-Image-Edit combines the separate references into their required single-image input.

Evaluation Metrics. We report VBench (Huang et al., 2024) aesthetics (AES), imaging quality (IQ), and motion smoothness (Smooth); DINOv2 (Oquab et al., 2023) person/product consistency (DINO_{id}/DINO_{obj}); ArcFace (Deng et al., 2019) facial similarity (FaceSim); and SyncNet (Chung & Zisserman, 2016) lip-sync confidence ($\text{Sync}_{\text{conf}}$). Our 50-question VLM-QA protocol assesses action completion, contact, coordinated motion, and structural plausibility on a 0–100 scale. Higher is better for all quality metrics; details are in Appendix D.1.

5.2 MAIN RESULTS

Table 2 shows that our four-step student leads the few-step baselines in HOI quality, reference consistency, and video quality. The improvements in HOI are accompanied by stronger preservation of both the presenter's identity and product appearance. This combination is important for product demonstrations: the generated video must depict convincing manipulation while keeping the presenter recognizable and the product faithful to its reference throughout the interaction. LiveAvatar, FlashTalk, and LongCat-Video-Avatar 1.5 do not accept separate person and product references. We use Qwen-Image-Edit to combine them into a single conditioning image. This adaptation provides both person and product appearance information in a format compatible with these baselines.

Qualitative Results. Figure 4 compares representative frames across person–product pairs with different shapes and packaging. LiveAvatar's highlighted frames show distorted finger configurations, while SkyReels-V3 exhibits hand–product boundary artifacts. HuMo and our teacher, CoInteract, exhibit weaker product-reference consistency in the displayed examples, with noticeable deviations from the reference appearance. Our results retain recognizable person and product appearances, depicting coherent hand–object contact using four-step sampling. Product contours remain consistent across displayed frames despite changes in hand pose. Space limits the main paper to four representative baselines; Appendix D.3 provides additional qualitative comparisons.

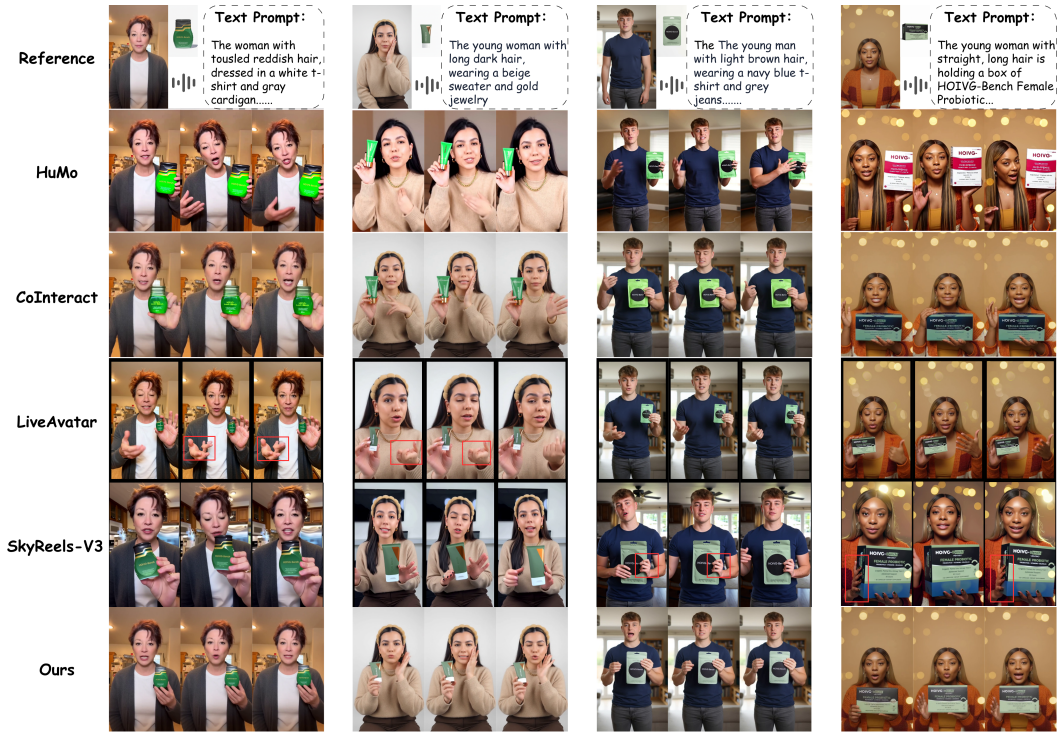

Figure 4: Qualitative comparison with four representative baselines. Each column group corresponds to one person–product pair, with three representative frames per method. The top row provides the person and product references together with the text prompt. Our four-step results appear in the bottom row. We use red boxes to highlight the hand and contact artifacts.

Table 2: Quantitative comparison with baseline methods. Our 4-step model leads on seven of eight metrics among few-step methods. Best and second-best scores are **bold** and underlined, respectively. "—" indicates unsupported. CoInteract is our teacher, and Ours is the distilled 4-step student.

Model	NFE	*Video Quality*			*HOI*	*Reference*			*Audio*
		AES↑	IQ↑	Smooth↑	VLM-QA↑	DINO_{obj}↑	DINO_{id}↑	FaceSim↑	$\text{Sync}_{\text{conf}}$↑
Many NFE									
Phantom	50	0.446	0.738	0.9914	75.612	0.682	0.408	**0.830**	—
VACE	50	**0.455**	0.731	0.9837	78.272	**0.686**	**0.411**	0.684	—
HuMo	50	0.434	**0.744**	0.9866	**82.114**	0.600	0.409	0.776	**7.520**
CoInteract (Teacher)	40	0.439	0.719	**0.9930**	82.068	0.597	0.355	0.708	6.797
Few NFE									
LiveAvatar	4	0.445	0.728	0.9884	75.886	0.623	0.407	0.695	7.495
FlashTalk	4	0.439	0.726	0.9917	74.738	0.589	0.403	0.755	7.700
SkyReels-V3	8	0.417	0.723	0.9881	71.979	0.618	0.376	0.736	—
LongCat-Video-Avatar 1.5	8	0.422	0.729	0.9900	75.684	0.594	0.397	0.715	**7.873**
Ours (Student)	4	**0.450**	**0.730**	**0.9940**	**76.957**	**0.628**	**0.4153**	**0.789**	7.724

5.3 Ablation Study

Table 3 isolates the effects of attention visibility and one-sided projection. Full-Visibility AP (FV-AP) enables full mutual attention between HOI tokens and the target-video, person, and product tokens. Target-Query AP (TQ-AP) enables only target-video queries to attend to HOI keys and values, retaining all other attention connections. Figure 5 presents qualitative comparisons for two person–product pairs, with three frames per variant.

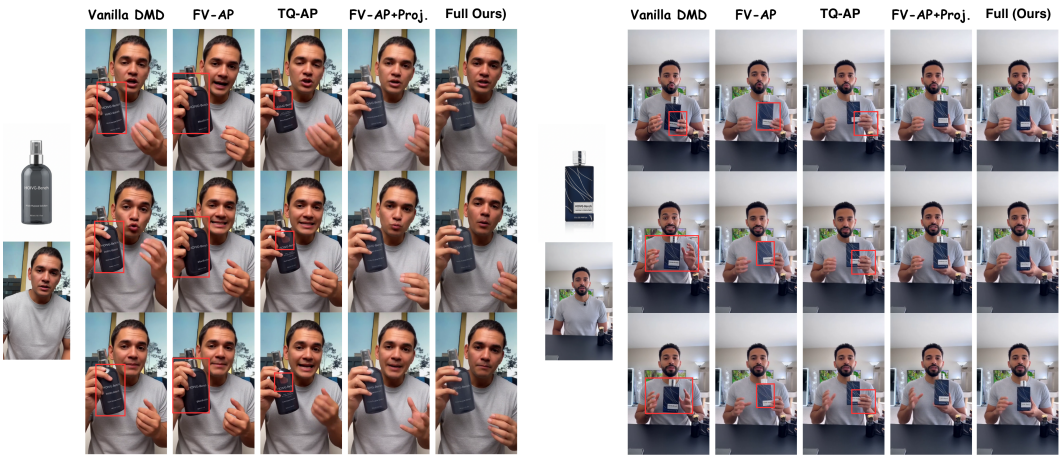

Figure 5: Qualitative ablation of attention perturbation and one-sided projection. Each example shows the product and person references, with five variant columns ordered as in Table 3 and three frames arranged vertically. We use red boxes to highlight the visual artifacts.

Vanilla DMD (Table 3) improves reference consistency over the teacher (Table 2) but reduces interaction plausibility (VLM-QA). FV-AP improves VLM-QA at the cost of product consistency (DINOobj). Restricting the intervention to target-video queries improves both metrics over FV-AP, although DINOobj remains below Vanilla DMD. The one-sided projection improves both metrics under either attention configuration. Combining TQ-AP and the projection yields the strongest HOI and reference-consistency scores among the ablations.

Figure 5 shows the complementary effects of AP and one-sided projection. Vanilla DMD struggles to maintain plausible hand–object contact. FV-AP and TQ-AP improve hand–object interaction, but product consistency remains limited, with noticeable deviations in bottle shape and label appearance from the reference. One-sided projection improves product consistency and hand–object interaction: the left example's rounded bottle contour and label layout are better preserved, while the right example's rectangular bottle is held more plausibly. Our full 4-step model maintains recognizable and consistent product appearance while producing coherent hand–object contact.

Table 3: Ablation study of HOI AP and the one-sided projection (Proj.). FV-AP and TQ-AP denote Full-Visibility AP and Target-Query AP, respectively. ✓ indicates an enabled component; the last row is our full method. Best and second-best scores are **bold** and underlined, respectively.

Components				*Video Quality*			*HOI*	*Reference*			*Audio*
DMD	FV-AP	TQ-AP	Proj.	AES↑	IQ↑	Smooth↑	VLM-QA↑	DINO_{obj}↑	DINO_{id}↑	FaceSim↑	$\text{Sync}_{\text{conf}}$↑
✓				0.4395	0.7294	0.99298	70.159	0.614	0.3773	0.764	7.243
✓	✓			0.4488	0.7297	0.99389	75.013	0.602	0.4065	0.779	7.587
✓		✓		0.4490	**0.7307**	0.99398	76.026	0.609	0.4102	0.780	7.666
✓	✓		✓	0.4457	0.7250	0.99395	76.347	0.611	0.4146	0.781	**7.813**
✓		✓	✓	**0.4496**	0.7303	**0.99402**	**76.957**	**0.628**	**0.4153**	**0.789**	7.724

6 Conclusion and Limitations

We presented FastInteract, a 4-step HOI video generator distilled with Attention Perturbation. Teacher prediction contrasts provide HOI enhancement and product-preservation responses, coordinated through one-sided projection. Experiments show improved interaction plausibility and product-reference consistency over vanilla DMD without additional inference cost. Building on these results, future work will explore extending Attention Perturbation to broader architectures, streaming settings, and long-horizon generation. Detailed limitations are discussed in Appendix A.

AI USE STATEMENT

We used generative AI tools for language polishing, data processing, and evaluation. Specifically, vision-language models were used for reference-image preparation and dataset curation, as described in Section 3 and Appendix B, and for baseline input preparation and VLM-QA evaluation, as described in Section 5.1 and Appendi D.1. The authors take responsibility for the final manuscript, data processing procedures, and reported results.

ETHICS STATEMENT

FastInteract studies efficient human–object interaction video generation, with potential applications in content creation and virtual product demonstrations. All portrait images displayed in this paper are AI-generated and sourced from openly available benchmarks and datasets, rather than photographs of real individuals. Nevertheless, the proposed technology could be misused to impersonate individuals, fabricate endorsements, or generate misleading product demonstrations. Generated videos should not be interpreted as evidence of actual human behavior or product performance. Responsible deployment should include clear disclosure of synthetic content and appropriate authorization for any real-person likeness or voice used as input. The focus on livestream commerce may also limit the representativeness of the data across people, products, and interaction settings.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 and Appendix B describe the construction, filtering, and multimodal annotations of InteractVerse. Section 4 presents the distillation method, while Appendix C provides the supporting mathematical analysis, assumptions, and implementation details, including attention routing, the student trajectory, teacher evaluations, and training losses. Section 5.1 describes the experimental setup and baselines. Appendix D.1 specifies the evaluation metrics, VLM-QA scoring procedure, and baseline input adaptations, while Appendices D.2 and D.3 provide additional qualitative results.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision (ECCV)*, pp. 1–17, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09184.pdf.

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 multimodal conditioning. *arXiv preprint arXiv:2509.08519*, 2025. URL https://arxiv.org/abs/2509.08519.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

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 (CVPR)*, 2019. URL https://arxiv.org/abs/1801.07698v3.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-S2V: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025. URL https://arxiv.org/abs/2508.18621.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7462–7471, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hong_Improving_Sample_Quality_of_Diffusion_Models_Using_Self-Attention_Guidance_ICCV_2023_paper.html.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live Avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025a. URL https://arxiv.org/abs/2512.04677.

Yubo Huang, Sirui Zhao, Xinchen Yao, Zhengye Zhang, Jinyang Huang, Fengqi Cui, Shiwei Wu, and Enhong Chen. DynaForcing: Overcoming dynamic collapse in self-forcing distillation for streaming avatar generation. *arXiv preprint arXiv:2608.17707*, 2026. URL https://arxiv.org/abs/2608.17707.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.17982.

Ziyao Huang, Zixiang Zhou, Juan Cao, Yifeng Ma, Yi Chen, Zejing Rao, Zhiyong Xu, Hongmei Wang, Qin Lin, Yuan Zhou, Qinglin Lu, and Fan Tang. HunyuanVideo-HOMA: Generic human-object interaction in multimodal driven human animation. *arXiv preprint arXiv:2506.08797*, 2025b. URL https://arxiv.org/abs/2506.08797.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jiang_VACE_All-in-One_Video_Creation_and_Editing_ICCV_2025_paper.html.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, Mingshan Chang, Yuqiang Xie, Binjie Mao, Youqiang Zhang, Nuo Pang, Hao Zhang, Yuzhe Jin, Zhiheng Xu, Dixuan Lin, Guibin Chen, and Yahui Zhou. SkyReels-V3 technique report. *arXiv preprint arXiv:2601.17323*, 2026. URL https://arxiv.org/abs/2601.17323.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. OmniHuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025. URL https://arxiv.org/abs/2502.01061.

Bangya Liu, Xinyu Gong, Zelin Zhao, Ziyang Song, Yulei Lu, Suhui Wu, Jun Zhang, Suman Banerjee, and Hao Zhang. ByteLoom: Weaving geometry-consistent human-object interactions through progressive curriculum learning. *arXiv preprint arXiv:2512.22854*, 2025a. URL https://arxiv.org/abs/2512.22854.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled DMD: CFG augmentation as the spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025b. URL https://arxiv.org/abs/2511.22677.

Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *CVPR*, pp. 24001–24010. IEEE, 2025c.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, 2025d. URL https://openaccess.thecvf.com/content/

ICCV2025/html/Liu_Phantom_Subject-Consistent_Video_Generation_via_Cross-Modal_Alignment_ICCV_2025_paper.html.

Xiangyu Liu, Feng Gao, Xiaomei Zhang, Yong Zhang, Xiaoming Wei, Zhen Lei, and Xiangyu Zhu. TurboTalk: Progressive distillation for one-step audio-driven talking avatar generation. *arXiv preprint arXiv:2604.14580*, 2026. URL https://arxiv.org/abs/2604.14580.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/260a14acce2a89dad36adc8eefe7c59e-Abstract-Conference.html.

Xiangyang Luo, Xiaozhe Xin, Tao Feng, Xu Guo, Meiguang Jin, and Junfeng Ma. CoInteract: Physically-consistent human-object interaction video synthesis via spatially-structured co-generation. *arXiv preprint arXiv:2604.19636*, 2026. URL https://arxiv.org/abs/2604.19636.

Meituan LongCat Team, Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, Xiaoming Wei, Tianyu Yang, and Yong Zhang. LongCat-Video-Avatar 1.5 technical report. *arXiv preprint arXiv:2605.26486*, 2026. URL https://arxiv.org/abs/2605.26486.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Qwen Team. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B, 2026.

Le Shen, Qian Qiao, Tan Yu, Ke Zhou, Tianhang Yu, Yu Zhan, Zhenjie Wang, Ming Tao, Shunshun Yin, and Siyuan Liu. SoulX-FlashTalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation. *arXiv preprint arXiv:2512.23379*, 2025. URL https://arxiv.org/abs/2512.23379.

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. URL https://arxiv.org/abs/2503.20314.

Linrui Tian, Qi Wang, and Bang Zhang. StreamChar: Long-horizon streaming character audio-video generation with decoupled orchestration. *arXiv preprint arXiv:2605.25659*, 2026. URL https://arxiv.org/abs/2605.25659.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. FantasyTalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*, 2025. URL https://arxiv.org/abs/2504.04842.

Ruikui Wang, Jinheng Feng, Lang Tian, Huaishao Luo, Chaochao Li, Liangbo Zhou, Huan Zhang, Youzheng Wu, and Xiaodong He. JoyStreamer: Unlocking highly expressive avatars via harmonized text-audio conditioning. *arXiv preprint arXiv:2602.00702*, 2026. URL https://arxiv.org/abs/2602.00702.

Zehao Xia, Yiqun Wang, Zhengda Lu, Kai Liu, Jun Xiao, and Peter Wonka. OMEGA-Avatar: One-shot modeling of 360° Gaussian avatars. *arXiv preprint arXiv:2602.11693*, 2026. URL https://arxiv.org/abs/2602.11693.

Shenbo Xie, Mingrui Cai, Xu Yang, Yifei Liu, and Changxing Ding. Sparsectrl-hoi: Sparse temporal control for human-object interaction video generation. *arXiv preprint arXiv:2607.05994*, 2026.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. URL https://arxiv.org/abs/2406.08801.

Ziyi Xu, Ziyao Huang, Juan Cao, Yong Zhang, Xiaodong Cun, Qing Shuai, Yuchen Wang, Linchao Bao, and Fan Tang. Anchorcrafter: Animate cyber-anchors selling your products via human-object interacting video generation. *IEEE TVCG*, 2026.

Zihui Xue, Mi Luo, Changan Chen, and Kristen Grauman. HOI-Swap: Swapping objects in videos with hand-object interaction awareness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.07754.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. InfiniteTalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025. URL https://arxiv.org/abs/2508.14033.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Hengyuan Zhang, Jingna Sun, Meiguang Jin, and Junfeng Ma. AptAvatar: Fast and vivid long-form audio-driven video generation for production-ready avatars. *arXiv preprint arXiv:2607.24013*, 2026a. URL https://arxiv.org/abs/2607.24013.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2211.12194.

Youliang Zhang, Zhengguang Zhou, Zhentao Yu, Ziyao Huang, Teng Hu, Sen Liang, Guozhen Zhang, Ziqiao Peng, Shunkai Li, Yi Chen, Zixiang Zhou, Yuan Zhou, Qinglin Lu, and Xiu Li. Making avatars interact: Towards text-driven human-object interaction for controllable talking avatars. *arXiv preprint arXiv:2602.01538*, 2026b. URL https://arxiv.org/abs/2602.01538.

Donghao Zhou, Guisheng Liu, Hao Yang, Jiatong Li, Jingyu Lin, Xiaohu Huang, Yichen Liu, Xin Gao, Cunjian Chen, Shilei Wen, Chi-Wing Fu, and Pheng-Ann Heng. OmniShow: Unifying multimodal conditions for human-object interaction video generation. *arXiv preprint arXiv:2604.11804*, 2026. URL https://arxiv.org/abs/2604.11804.

A LIMITATIONS

Although our method achieves promising results, it still has several limitations. While FastInteract reduces sampling to four steps, further acceleration is needed for real-time video generation. Reducing the computational cost of each sampling step and improving inference efficiency are directions for future work. Additionally, our current framework also does not support inputs involving multiple people or multiple objects. Extending it to these settings and improving robustness under challenging conditions remain directions for future research.

B ADDITIONAL DETAILS OF INTERACTVERSE

B.1 DATASET STATISTICS

InteractVerse contains 37,477 curated HOI samples with a total duration of approximately 64.1 hours. All videos are standardized to 25 FPS. Each sample is represented as a temporally continuous interaction segment and is organized into a preceding motion context followed by a target HOI video.

The motion context and target video are extracted from the same continuous sequence rather than independently sampled. This organization preserves the natural temporal transition into the target interaction and provides an explicit observation of the preceding motion state. The two parts are stored as separate modalities while maintaining their original temporal correspondence.

B.2 DATASET CURATION

We construct InteractVerse from e-commerce livestreams using a coarse-to-fine curation pipeline. Lightweight filters are first applied to reduce low-quality and interaction-irrelevant content, after which more expensive visual and multimodal verification is performed on the remaining candidates.

We focus on product-oriented livestream categories in which presenters visibly demonstrate or manipulate commercial products, while excluding categories such as gaming, finance, and education that contain limited observable product interaction. Videos with insufficient resolution or frame rate, corrupted streams, missing audio, or evident audio–visual desynchronization are removed during technical filtering. We then identify interaction-rich segments using face, hand, and product detections on temporally sampled frames. Candidate clips are retained only when a visible presenter, hand activity, and a recognizable product are jointly observed. This criterion reduces clips dominated by static presentation, background products, or incidental person–object co-occurrence. For each retained clip, we select a representative frame and use Qwen-Image-Edit to obtain a person reference image and a product reference image. A VLM then verifies the triplet by checking whether the extracted references preserve the identity and appearance of the corresponding person and product in the source frame, as well as basic visual quality. Samples with missing entities, mismatched references, severe artifacts, or ambiguous correspondence are discarded. Finally, an interaction-aware caption is generated for each verified sample. The caption summarizes the visible participants, product, actions, and interaction progression, providing a semantic description that is consistent with the corresponding video and reference conditions. The verified samples constitute the curated core of InteractVerse, upon which the remaining multimodal annotations are constructed.

B.3 INTERACTION-COMPLETE MULTIMODAL ANNOTATIONS

For each curated HOI sample, we provide a set of aligned and independently accessible annotations corresponding to the four aspects of *interaction completeness* introduced in Sec. 3.1: interaction semantics, participant appearance, temporal dynamics, and interaction-sensitive spatial structure.

As summarized in Table 4, semantic annotations describe the interaction content and action progression. Appearance annotations provide disentangled references for the presenter and product together with the target RGB video. Temporal annotations include the preceding motion context, synchronized audio, and human pose sequence, which provide complementary signals for motion and speech dynamics. Spatial annotations further localize the human body, face, and hand–object interaction regions. A key property of this annotation scheme is that these modalities are factorized without breaking their correspondence. In a conventional RGB video, identity, product appearance, motion,

Table 4: Multimodal annotations provided in InteractVerse. All modalities are aligned with the same underlying HOI event and can be accessed independently.

Group	Field	Format	Description
Semantics	`prompt`	Text	Interaction and action description
Appearance	`input_video`	Video	Target RGB clip
	`person_image`	Image	Disentangled person reference
	`product_image`	Image	Disentangled product reference
Dynamics	`motion_video`	Video	Preceding motion context
	`audio`	Audio	Synchronized speech signal
	`pose_video`	Video	Human pose sequence
Spatial structure	`human_bbox`	BBox sequence	Human body regions
	`face`	BBox sequence	Face regions
	`hand_object`	BBox sequence	Hand–object interaction regions
Metadata	`height`	Integer	Video height after preprocessing
	`width`	Integer	Video width after preprocessing

speech, pose, and contact cues are entangled within the same visual stream. InteractVerse exposes these factors as separate supervision signals while retaining their alignment to the same underlying interaction. This organization enables different subsets of annotations to be used independently or jointly, supporting reference-guided synthesis, motion-conditioned generation, audio-aware generation, pose control, and spatially grounded HOI modeling.

C Score Analysis and Implementation Details

C.1 The Reverse-KL Gradient

Fix the conditioning inputs and noise level t. Let $X_\theta = \alpha_t G_\theta(z, c) + \sigma_t \epsilon$, with (z, ϵ) sampled from a parameter-independent distribution, and let $p_{\text{fake},t}$ be its density, with its dependence on θ left implicit. Assume positive differentiable densities, a teacher target independent of θ, and sufficient regularity to interchange differentiation and integration. For any fixed target density p_t, differentiating $\mathbb{E}[\log p_{\text{fake},t}(X_\theta) - \log p_t(X_\theta)]$ gives

$$\begin{aligned}\nabla_\theta D_{\mathrm{KL}}(p_{\text{fake},t} \| p_t) = \mathbb{E}\big[\alpha_t J_\theta^\top \big(\nabla_x \log p_{\text{fake},t}(X_\theta) - \nabla_x \log p_t(X_\theta)\big)\big] \\ + \mathbb{E}_{x \sim p_{\text{fake},t}}\left[\nabla_\theta \log p_{\text{fake},t}(x)|_{x \text{ fixed}}\right].\end{aligned} \tag{12}$$

The last expectation is zero because it equals $\int \nabla_\theta p_{\text{fake},t}(x)\, dx = \nabla_\theta \int p_{\text{fake},t}(x)\, dx = 0$. Averaging over t, c with weight w_t yields equation 2. In a full-generator derivation, J_θ is the Jacobian of the entire reparameterized sampler. Our implementation instead uses a detached trajectory prefix and differentiates the selected denoising step, producing a local surrogate.

C.2 The Target Induced by CFG Augmentation and AP Responses

At a fixed noise level and fixed auxiliary context, assume the four teacher fields are the exact scores of positive differentiable densities $p_{\text{real}}, p_{\text{real}}^-, p_{\text{real}}^{k,+}, p_{\text{real}}^{k,-}$; this subsection omits the fixed timestep subscript t. The superscripts $k, +$ and $k, -$ denote teacher configurations where target-video queries can and cannot attend to the keys and values of condition k, respectively. At each noisy video state, this AP pair uses identical conditioning inputs and auxiliary states. All conditioning tokens remain present, and all other attention connections are unchanged. We analyze the unprojected AP augmentation in equation 7. Let $\beta = \gamma - 1$, and suppose β, λ are independent of x and

$$0 < Z = \int p_{\text{real}}(x) \left(\frac{p_{\text{real}}(x)}{p_{\text{real}}^-(x)}\right)^\beta \left(\frac{p_{\text{real}}^{k,+}(x)}{p_{\text{real}}^{k,-}(x)}\right)^\lambda dx < \infty. \tag{13}$$

Define the reweighted density

$$p_{\mathrm{aug}}(x) = \frac{1}{Z}\, p_{\mathrm{real}}(x) \left(\frac{p_{\mathrm{real}}(x)}{p^{-}_{\mathrm{real}}(x)}\right)^{\beta} \left(\frac{p^{k,+}_{\mathrm{real}}(x)}{p^{k,-}_{\mathrm{real}}(x)}\right)^{\lambda}. \tag{14}$$

Its log is

$$\log p_{\mathrm{aug}} = \log p_{\mathrm{real}} + \beta \log \frac{p_{\mathrm{real}}}{p^{-}_{\mathrm{real}}} + \lambda \log \frac{p^{k,+}_{\mathrm{real}}}{p^{k,-}_{\mathrm{real}}} - \log Z. \tag{15}$$

Since Z does not depend on x,

$$\nabla_x \log p_{\mathrm{aug}} = s_{\mathrm{real}} + \beta(s_{\mathrm{real}} - s^{-}_{\mathrm{real}}) + \lambda(s^{k,+}_{\mathrm{real}} - s^{k,-}_{\mathrm{real}}) = s^{\mathrm{aug}}_{\mathrm{real}}. \tag{16}$$

For a student density p_{fake} for which the following expectations are finite, this also gives

$$\begin{aligned} D_{\mathrm{KL}}(p_{\mathrm{fake}} \| p_{\mathrm{aug}}) = D_{\mathrm{KL}}(p_{\mathrm{fake}} \| p_{\mathrm{real}}) &- \beta\, \mathbb{E}_{p_{\mathrm{fake}}} \log \frac{p_{\mathrm{real}}}{p^{-}_{\mathrm{real}}} \\ &- \lambda\, \mathbb{E}_{p_{\mathrm{fake}}} \log \frac{p^{k,+}_{\mathrm{real}}}{p^{k,-}_{\mathrm{real}}} + \log Z. \end{aligned} \tag{17}$$

CFG augmentation (CA) and the augmentation term constructed from attention perturbation (AP) responses therefore enter as parallel log-density-ratio contributions. This identity explains their common role in reshaping a distillation target; it does not identify either ratio with an externally measured semantic reward. If the fake score is exact and the reweighted target is independent of θ, the assumptions of the preceding subsection give the corresponding reverse-KL gradient. The implemented update uses prediction differences and does not evaluate these densities or Z.

For the HOI enhancement response, $p^{k,-}_{\mathrm{real}} = p^{\mathrm{H}-}_{\mathrm{real}} = p_{\mathrm{real}}$ and $p^{k,+}_{\mathrm{real}} = p^{\mathrm{H}+}_{\mathrm{real}}$, where H$-$ denotes base attention and H$+$ makes HOI keys and values accessible to target-video queries. Person, product, and motion queries cannot attend to HOI in either configuration; the pair shares all token inputs and all other attention connections. With $\gamma = 1$ and $0 \le \lambda \le 1$, the target is the normalized geometric interpolation $p_{\mathrm{aug}} \propto p^{1-\lambda}_{\mathrm{real}} (p^{\mathrm{H}+}_{\mathrm{real}})^{\lambda}$. For $\lambda = 1$ and arbitrary γ, the augmented score is

$$s^{\mathrm{aug}}_{\mathrm{real}} = s^{\mathrm{H}+}_{\mathrm{real}} + (\gamma - 1)(s_{\mathrm{real}} - s^{-}_{\mathrm{real}}). \tag{18}$$

Thus, before projection, the teacher score with target-video queries' access to HOI enabled replaces the unguided base score, while the CA difference under base attention is retained. This is distinct from applying the original CFG scale to every pathway contribution.

Scope of the interpretation. A perturbed neural predictor need not be a conservative vector field or an exact score of a normalized density. Moreover, the independently reweighted densities at different noise levels need not be the Gaussian marginals of one common clean distribution. The preceding result is therefore an exact implication of its stated score and integrability assumptions, rather than a guarantee for arbitrary attention interventions.

In our implementation, the HOI state is generated by the current student and is correlated with its RGB output. It is held fixed within each teacher comparison, but its distribution changes with the student. The fake model estimates the RGB marginal given c, not an RGB distribution additionally conditioned on that HOI state. Consequently, the fixed-auxiliary-state density identity does not establish that the actual training loop follows the gradient of a single global HOI-conditional reverse KL. Projection, sample-dependent normalization, and the detached trajectory prefix further distinguish the implemented update from that idealized objective.

C.3 Projection and Numerical Reliability

For fixed enhancement and preservation responses h, r defined in equation 11, the objective $\frac{1}{2}\|v - h\|_2^2$ is strictly convex and its feasible set $\{v : \langle v, r\rangle \ge 0\}$ is a closed half-space. If $\langle h, r\rangle \ge 0$, the unique minimizer is h. Otherwise, the Lagrangian is $\frac{1}{2}\|v - h\|_2^2 - \nu\langle v, r\rangle$, with $\nu \ge 0$. Stationarity gives $v = h + \nu r$, and the active constraint gives $\nu = -\langle h, r\rangle / \|r\|_2^2 > 0$, proving equation 9.

For the exact projection, $\langle \mathcal{P}_r(h), r\rangle \geq 0$ and $\|\mathcal{P}_r(h)\| \leq \|h\|$. For the update without AP $d_{\text{base}} = (\mu_{\text{real}} - \mu_{\text{fake}}) + (\gamma - 1)(\mu_{\text{real}} - \mu^{-}_{\text{real}})$, we have

$$\langle d_{\text{total}} - d_{\text{base}}, r\rangle = \lambda \langle \mathcal{P}_r(h), r\rangle \geq 0. \tag{19}$$

This concerns the additional response in prediction space. It does not imply that the complete update has nonnegative inner product with r, or that an optimizer step necessarily improves a product metric.

The implementation computes the geometry in FP32 and applies the projection only when

$$\|h\|_2^2 > \delta, \qquad \|r\|_2^2 > \delta, \qquad \frac{\|r\|_2}{\max(\|\mu_{\text{real}}\|_2, \delta)} \geq \eta, \quad \delta = 10^{-12}, \quad \eta = 10^{-4}. \tag{20}$$

If any test fails, h is retained. Thus exact non-conflict is asserted only for the valid-projection branch, up to floating-point error. In that branch, the denominator is $\max(\|r\|_2^2, \delta)$, which equals $\|r\|_2^2$ by the validity test; no additive epsilon changes the projection coefficient. Responses are subsequently cast to the original model dtype. The positive normalizer in equation 24 preserves the direction of a given total update, but separately normalized unaugmented and augmented updates need not retain the above relative-inner-product inequality.

C.4 Implementation Used for the HOI Instantiation

Our main HOI configuration uses Target-Query AP (TQ-AP) with one-sided projection, corresponding to Ours in Table 3. We use $\gamma = 5$, $\lambda = 1$, and no additional DM or CA rescaling. Both interventions cover all layers and heads. Adaptive coefficients, the optional RMS cap, and additive product guidance are disabled.

Base model and routing. CoInteract's conditioning inputs include person and product references, text, speech audio, and motion context. Its joint sequence contains target-video, person, product, motion, and auxiliary HOI tokens, with a separate patch embedding for HOI. The *Base Attention* configuration (H$-$) in Figure 3(b) uses asymmetric attention: target-video, person, product, and motion queries cannot attend to HOI keys and values, while target-video queries can attend to product-reference keys and values. HOI queries attend to HOI, target-video, and available product-reference tokens. For *HOI Perturbation* (H$+$), we compute target-video attention separately and add HOI keys and values only to the set accessible to target-video queries. Person, product, and motion queries remain blocked from HOI keys and values. For *Product Perturbation* (P$-$), we remove product-reference keys and values only from the set accessible to target-video queries under base attention. Target-video, person, product, and motion queries remain blocked from HOI. All other connections, including HOI queries' access to product-reference tokens, retain their base attention rules in both interventions. All three configurations use the same token inputs. These operations implement the mask in equation 4 by splitting the queries and selecting their key/value sets before Flash Attention. The shared base prediction $\mu^{\text{H}-}_{\text{real}} = \mu_{\text{real}}$ is the blocked member of the HOI pair and the accessible member of the product pair in equation 11.

Comparison with full visibility. Full-Visibility AP (FV-AP) in Table 3 enables full mutual attention between HOI tokens and the target-video, person, product, and motion tokens. FV-AP and TQ-AP use the same token inputs and base attention configuration for comparison, but change different attention connections. FV-AP changes connections involving several query groups, so its prediction contrast reflects their combined effects. TQ-AP restricts the direct intervention to target-video queries and the selected conditioning keys and values. Hidden states are recomputed in each forward pass, and effects can still propagate through other tokens across layers. This restriction therefore controls the attention connections perturbed without guaranteeing a response confined to one semantic attribute, motivating coordination of enhancement and preservation responses.

Student trajectory and teacher forward passes. The student uses four denoising anchors with implementation indices $[1000, 938, 833, 625]$ and timestep shift 5. RGB and HOI trajectories advance jointly from independent Gaussian noise under asymmetric attention. We simulate the prefix gradient-free, separately re-noising RGB and HOI at subsequent anchors. At the selected training

step, the student predicts both clean latents; only the RGB output receives the DMD gradient. The HOI co-output $\hat{u}_0$ is detached and independently re-noised at the sampled RGB score timestep:

$$u_t = \alpha_t \, \mathrm{sg}(\hat{u}_0) + \sigma_t \epsilon_u, \qquad \epsilon_u \sim \mathcal{N}(0, I), \tag{21}$$

where sg denotes stop-gradient. We use no ground-truth HOI target or separate HOI prediction loss. The online denoiser estimates the student RGB marginal given c without receiving u_t.

The frozen teacher uses evaluation mode, reusing RGB and HOI states across AP comparisons. The CFG reference replaces text with a negative prompt and zeros audio, retaining reference images and motion conditions. The RGB-only online denoiser uses native flow MSE, with five critic updates per generator update. Its replay buffer holds four entries, with maximum age two generator updates.

Flow parameterization and losses. For $\alpha_t = 1 - \sigma_t$ and velocity target $\epsilon - x_0$, clean predictions and their differences obey

$$\mu = x_t - \sigma_t v, \qquad \Delta\mu = -\sigma_t \Delta v. \tag{22}$$

All responses are computed in clean VAE-latent space, with the opposite sign to velocity differences. The online denoiser uses the native flow predictor v_ψ, with $\mu_{\text{fake}} = x_t - \sigma_t v_\psi(x_t, t, c)$, and is trained on detached RGB samples $\bar{x}_0$ with D output coordinates using

$$\mathcal{L}_{\text{fake}} = \frac{1}{D}\mathbb{E}\left\| v_\psi(\alpha_t \bar{x}_0 + \sigma_t \epsilon, t, c) - (\epsilon - \bar{x}_0) \right\|_2^2 . \tag{23}$$

For the target and direction in equation 10, we train the generator with the normalized surrogate

$$\begin{aligned} n &= \frac{1}{D}\|\hat{x}_0 - \mu_{\text{target}}\|_1 + 10^{-8}, \\ \mathcal{L}_G &= \frac{1}{2D}\mathbb{E}\left\| \hat{x}_0 - \mathrm{sg}\left(\hat{x}_0 + \frac{d_{\text{total}}}{n} \right) \right\|_2^2 , \end{aligned} \tag{24}$$

Its local gradient is $-\mathbb{E}[(J_\theta^{\text{local}})^\top d_{\text{total}}/(Dn)]$, where J_θ^{local} differentiates only the selected student denoising step. The prefix, auxiliary HOI state, and target are detached.

Finally, the score-to-clean conversion in equation 6 gives the idealized gradient factor

$$-\mathbb{E}\left[w_t \frac{\alpha_t^2}{\sigma_t^2} J_\theta^\top d_{\text{aug}}^x \right] \tag{25}$$

where d_{aug}^x is the unprojected clean-prediction difference. The implemented surrogate equation 24 instead uses detached, sample-dependent normalization and a per-coordinate mean, accounting for its $1/D$ factor.

D Additional Experiments

D.1 Evaluation Details

Video quality and reference consistency. VBench (Huang et al., 2024) computes AES using LAION's aesthetic predictor, IQ using MUSIQ, and Smooth using frame interpolation. $\mathrm{DINO}_{\mathrm{id}}$ and $\mathrm{DINO}_{\mathrm{obj}}$ measure DINOv2 (Oquab et al., 2023) feature cosine similarity between generated person/product regions and their original references. FaceSim uses ArcFace (Deng et al., 2019); $\mathrm{Sync}_{\mathrm{conf}}$ measures SyncNet (Chung & Zisserman, 2016) lip-sync confidence.

Interaction plausibility. Our VLM-QA protocol uses a locally deployed Qwen3.8 (Qwen Team, 2026) and a fixed 50-question bank covering action correctness and completion, contact and spatial support, hand–object motion synchronization, and structural plausibility. Question applicability is determined from the prompt and original product reference before inspecting generated videos. Within each category, we compute the fraction of clean judgments among applicable questions, assigning no credit to uncertain answers. The four categories contain 10, 16, 12, and 12 questions, respectively. We apply category-level critical-failure caps, combine the available category scores with a weighted harmonic mean (weights 0.25, 0.35, 0.25, and 0.15 in the order above, renormalized over available categories), and then apply overall critical-failure caps. An available category with zero score makes the harmonic mean zero. The resulting video scores are on a 0–100 scale and averaged across videos.

Conditioning compatibility. LiveAvatar, FlashTalk, and LongCat-Video-Avatar 1.5 require a single image rather than separate person and product references. Qwen-Image-Edit combines the reference pair into one image for these baselines, as shown in Fig. 8. Reference-consistency metrics use the original separate references. This adaptation enables evaluation on the same benchmark, but the methods still differ in their conditioning interfaces; the component ablations provide the more direct comparison of our distillation choices.

D.2 Qualitative AP Coefficient Probes

We qualitatively probe the two AP responses by varying their coefficients with conditioning inputs fixed within each row. First, we apply the HOI response to Vanilla DMD without one-sided projection. As shown in Figure 6, increasing the coefficient produces progressively stronger and more plausible human–object interaction, but also causes increasing deviation from the product reference. A coefficient near 1 provides a favorable interaction result, whereas larger values introduce visible image artifacts. This sweep motivates the HOI AP weight used for distillation in Sec. 5.1.

Second, we apply the product response to the frozen teacher. In Figure 7, increasing the coefficient progressively improves the generated product's agreement with its reference. This result indicates that the product response identifies a direction that can enhance product consistency. In our full method, we use this response as the direction for one-sided projection; the sweep is diagnostic, and the response is not added separately to the distillation target (Sec. 4.3).

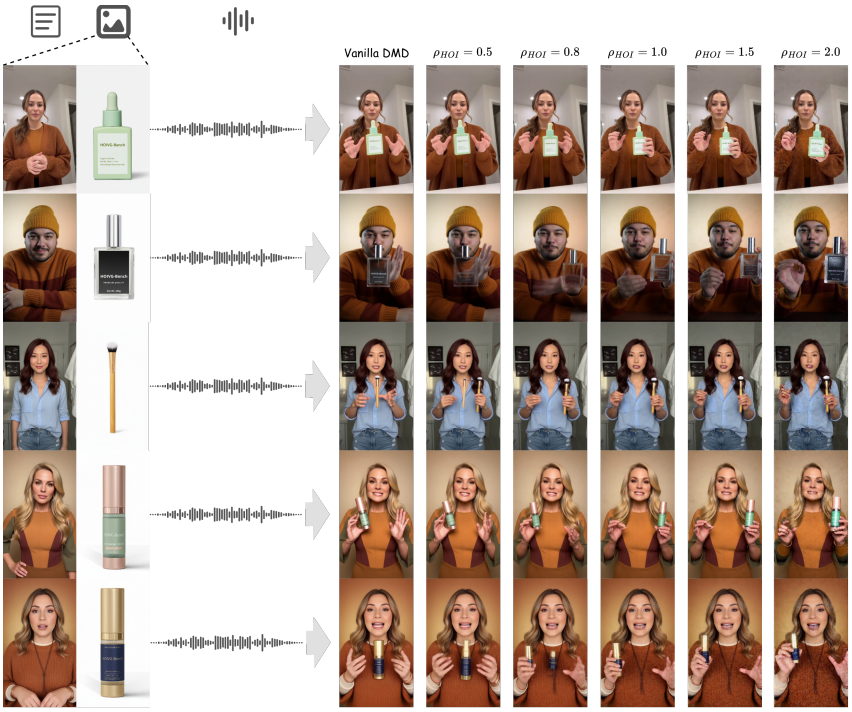

Figure 6: Qualitative HOI-AP probe on Vanilla DMD. ρ_{HOI} scales the HOI-AP response, with Vanilla DMD corresponding to $\rho_{\text{HOI}} = 0$. Increasing it strengthens human–object interaction but gradually reduces product fidelity and introduces artifacts. This confirms that HOI-AP provides an interaction-enhancement direction and motivates $\rho_{\text{HOI}} = 1.0$ as a balanced setting.

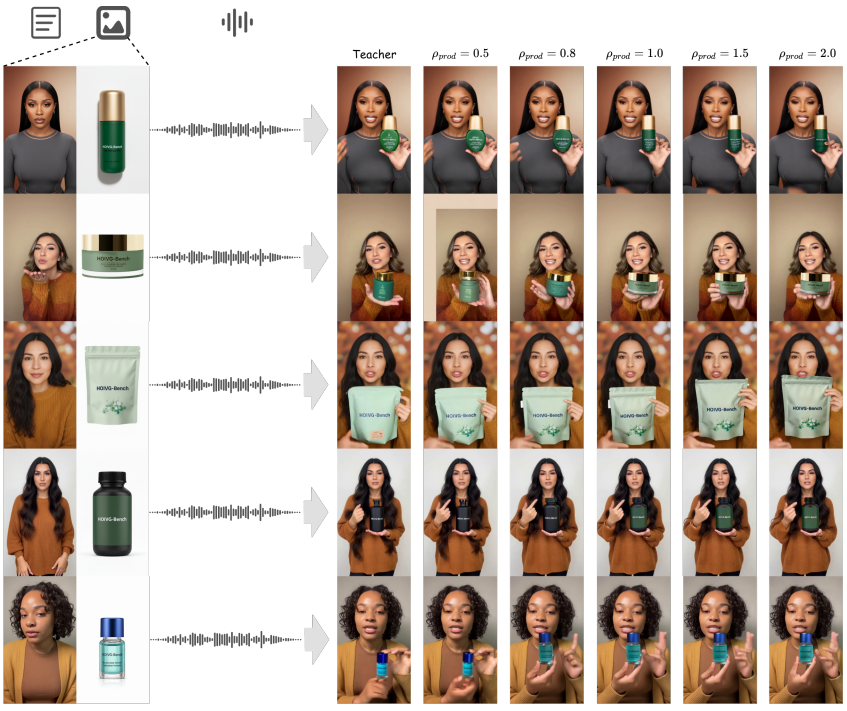

Figure 7: Qualitative Product-AP probe on the frozen teacher. ρ_{prod} scales the Product-AP response, with the unmodified teacher corresponding to $\rho_{\text{prod}} = 0$. Increasing it improves agreement with the product reference in shape, color, and packaging details, indicating that the Product-AP response identifies a direction that can enhance product consistency.

D.3 Additional Qualitative Results

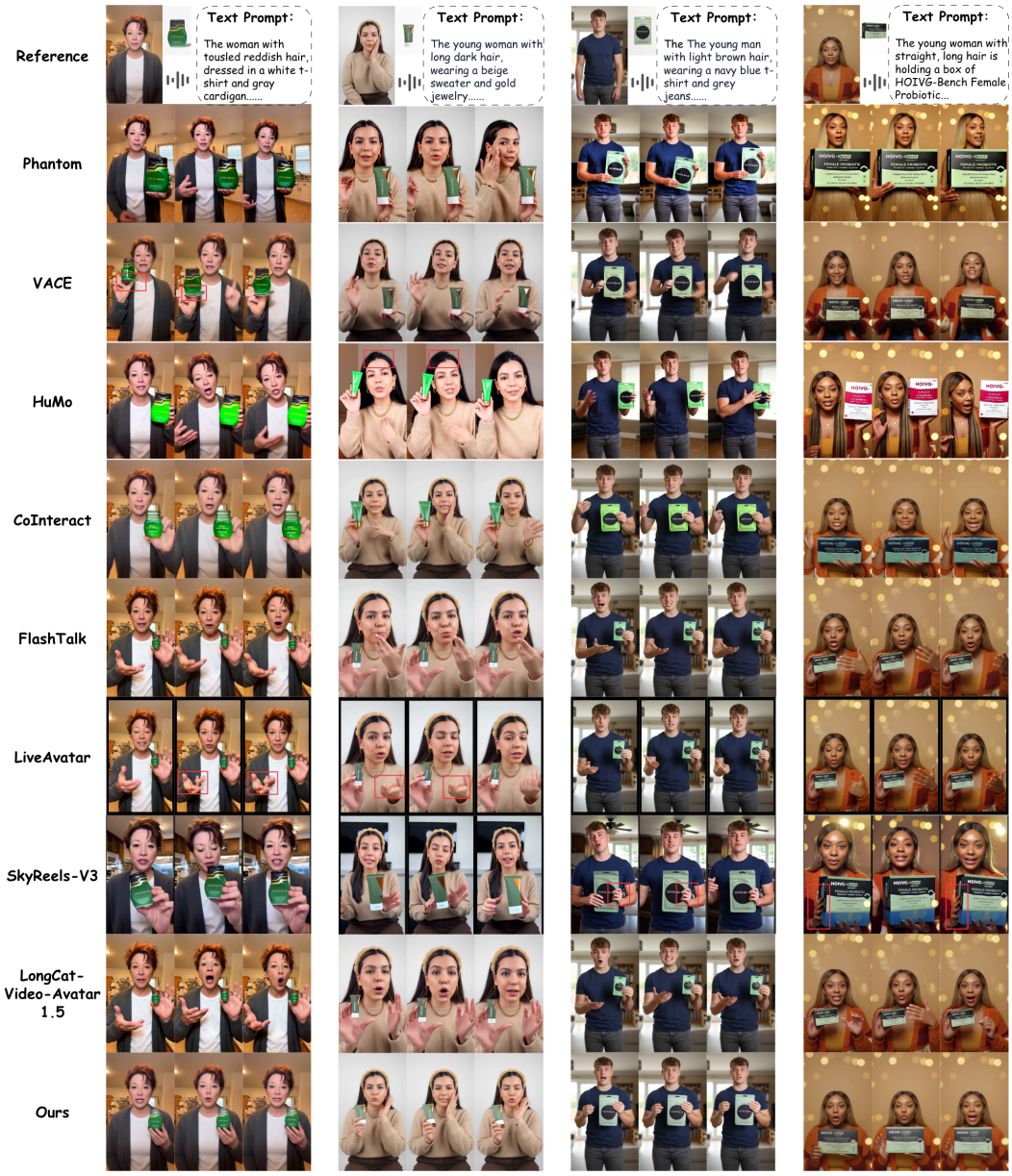

Figure 8: Full qualitative comparison with eight baselines across four person–product pairs. Each column group shows the reference inputs and text prompt, followed by three representative frames per method. Our four-step results appear in the bottom row. For LiveAvatar, FlashTalk, and LongCat-Video-Avatar 1.5, Qwen-Image-Edit combines the person and product references into one image used as the first-frame input. Red boxes highlight visible artifacts.