

VHOI: Controllable Video Generation of Human-Object Interactions from Sparse Trajectories via Motion Densification

Supplementary Material

The supplementary material is organized as follows. We begin with FiLM preliminaries and a more detailed diagram of the fuser placement inside the DiT block in Sec. 8, followed by a perceptual user study in Sec. 8. We provide additional HOI mask generation quality analysis in Sec. 11. We then detail the comparison to a UNet-based model in Sec. 10. Next, we visualize the HOI mask color palette in Sec. 13, and provide the prompt template in Sec. 14, which is used to process the text prompt for training the augmentor.

7. FiLM preliminaries

Feature-wise Linear Modulation (FiLM) [64] augments a network backbone with a learned feature-wise affine operator that conditions on an auxiliary signal without altering the backbone topology. Let $\mathbf{x} \in \mathbb{R}^{N \times D}$ denote an intermediate feature matrix (e.g., a sequence of N tokens with feature dimension D) and $\mathbf{z} \in \mathbb{R}^d$ a conditioning embedding (e.g., motion features in our case). FiLM realizes a mapping $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^{2D}$ that yields feature-wise parameters $(\gamma, \beta) = \psi(\mathbf{z})$, and applies them as

$$\mathbf{x}' = \mathbf{x} \odot \gamma + \beta, \quad (8)$$

FiLM implements a conditional reparameterization that aligns feature responses with the semantics encoded in the conditioning stream, leading to tight cross-modal coupling. In addition, it is also lightweight and architecture-agnostic, making it a popular choice for conditioning large pretrained backbones in a parameter-efficient manner.

In VHOI, FiLM modulation is applied in both stages with stage-specific motion conditioning. Specifically, in the trajectory augmentor \mathcal{A} , the fuser \mathcal{E}_a predicts motion-dependent scale/shift fields from F_a , applies visibility gating to obtain $\tilde{\gamma}_{\text{aug}}, \tilde{\beta}_{\text{aug}}$, and injects them into visual tokens via normalized residual modulation. In the dense control model \mathcal{D} , \mathcal{E}_m and the confidence branch analogously produce gated HOI conditioning from \tilde{F}_{hoi} for the same pre-attention modulation pathway. Therefore, a shared FiLM-style design is used to inject sparse or dense HOI motion cues through lightweight conditional affine modulation.

8. Inside the DiT Block

We provide a detailed diagram in Fig. 7 illustrating where the fuser is inserted into the DiT block for the augmentor. The DiT tokens are first modulated by the AdaLN scale and shift operation, with fusion occurring before the attention layer.

9. Perceptual User Study

Despite the efforts in benchmarking video generation quality, existing quantitative metrics still miss nuances of interaction realism. To complement them, we conduct a user study with 46 participants. The survey consists of 20 questions in total, and is split into two groups of 10 questions. Each comparison presents our video alongside either TORA* or Go-With-the-Flow in random order. Group 1 asks: *Which video exhibits more natural human-object interaction?*. Group 2 shows videos with overlaid trajectories, and asks *Which video follows the trajectories more accurately?* Aggregated preferences are shown in Fig. 8. Participants favor our method in 62.2% of interaction comparisons vs TORA* and 86.1% vs. Go-With-the-Flow, and in 60.9% and 75.2% of trajectory comparisons, respectively. Two-sided binomial z -tests with $\alpha = 0.05$ confirm that all improvements are statistically significant. Qualitatively, Go-With-the-Flow often produces rigid shifts and incoherent motion, whereas TORA* is competitive yet frequently lacks precise instance awareness during contact-rich interactions.

10. Comparison with MotionI2V

We additionally compare our method against MotionI2V [70], a two-stage controllable video generation framework based on a UNet backbone with optical-flow based densification. Because MotionI2V operates at a lower resolution ($16 \times 512 \times 312$) than ours ($49 \times 720 \times 490$), we downsample our generated videos and compute the trajectory error (TE), and contact accuracy (CA) at their native resolution for a fair comparison, while the rest of the metrics are reported with the respective native resolution. Quantitative evaluations are presented in Tab. 4. The qualitative results of MotionI2V exhibit restricted motion and reduced coherence; further qualitative comparisons are included in the supplementary video.

Method	FVD↓	TE↓	CA↑	CLIPSIM↑	VBench					
					SC↑	BC↑	DD↑	MS↑	AQ↑	IQ↑
MotionI2V [70]	1629	19.06	0.745	0.3054	0.95	0.96	0.26	0.99	0.47	0.65
VHOI (Ours)	915	14.82	0.830	0.3036	0.93	0.94	0.58	0.99	0.51	0.68

Table 4. **Quantitative Comparison with MotionI2V.** As evidenced by the significantly higher FVD and CA and lower TE, MotionI2V produces less realistic interactions compared with VHOI.

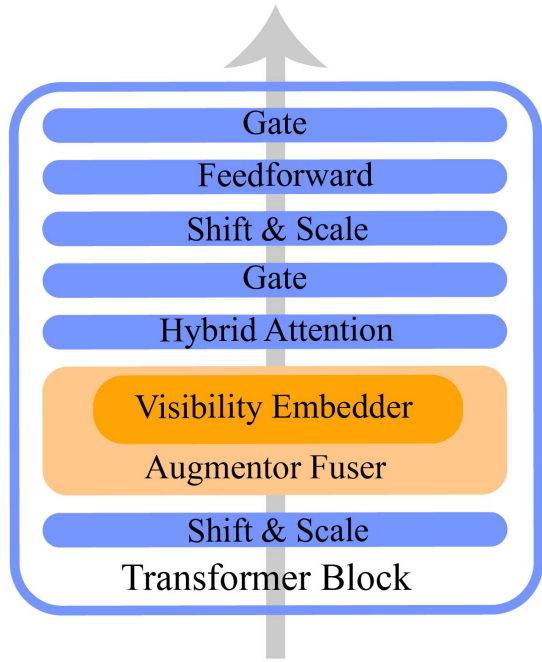


Figure 7. The fuser is inserted between the AdaLN scale-and-shift and the hybrid attention. The same placement is used in the dense model. The fuser is fine-tuned, while all other layers remain frozen.



Figure 8. **User Study Results.** VHOI is preferred over Tora* (finetuned) and Go-With-the-Flow in terms of both human-object interaction quality and trajectory adherence.

11. HOI Mask Quality Evaluation

We evaluate how well the generated HOI masks match the ground truth on the HOI-Gen-1M validation set. Following our controllable setting, for each frame we first identify

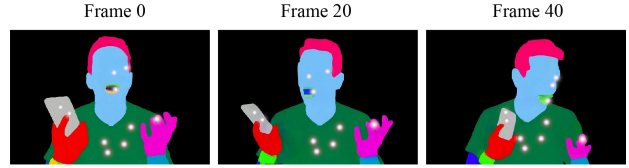


Figure 9. Overlaying the trajectory on the HOI masks improves trajectory adherence during generation, as the dense model leverages both the sparse trajectory input and the generated HOI masks. The figure shows three frames of inferred HOI masks during validation, with trajectories indicated by white dots.

HOI classes touched by the control trajectories and compute class-wise IoU on those classes only. We then report a dataset-level weighted metric, denoted as $mIoU_w$, where each class contribution is weighted by its total ground-truth area: $mIoU_w = \sum_{c \in \mathcal{C}_{ctrl}} \omega_c IoU_c$, $\omega_c = \frac{A_c^{gt}}{\sum_{k \in \mathcal{C}_{ctrl}} A_k^{gt}}$, where \mathcal{C}_{ctrl} denotes trajectory-controlled classes and A_c^{gt} is the ground-truth pixel count of class c . Our model obtains an $mIoU_w$ of 0.71. Despite temporal and spatial inconsistencies in both pseudo ground-truth masks and augmentor predictions, this result indicates that the predicted HOI masks remain well aligned with the target mask distribution.

12. Additional Improvement

While VHOI significantly outperforms existing methods in terms of video quality, its comparable performance on TE and CA metrics may stem from the dense model’s lack of explicit trajectory awareness when structured HOI masks are used as the sole motion guidance. To address this, we introduce a simple modification: overlaying trajectories onto object masks during training and evaluate on overlaid HOI masks for validation, as illustrated in Fig. 9. This results in consistent improvements in both TE and CA, as reported in Tab. 5. We also tried to additionally overlay the trajectories on the human segmentation masks during training, but we found that it did not lead to additional performance gain.

Method	FVD↓	TE↓	CA↑	CLIPSIM↑	VBench					
					SC↑	BC↑	DD↑	MS↑	AQ↑	IQ↑
VHOI	915	10.64	0.827	0.3036	0.93	0.94	0.58	0.99	0.51	0.68
VHOI-Traj	933	8.74	0.833	0.3042	0.93	0.94	0.62	0.98	0.51	0.67

Table 5. Overlaying trajectories onto object masks during dense model training further improves TE and CA on HOI-Gen dataset.

13. HOI mask color palette

We visualize the HOI-mask color palette in Fig. 10. This palette introduces both part and interaction awareness into the augmentor, enabling richer motion guidance for the dense model.

Background	Right Lower Arm
Apparel	Right Lower Leg
Face Neck	Right Shoe
Hair	Right Sock
Left Foot	Right Upper Arm
Left Hand	Right Upper Leg
Left Lower Arm	Torso
Left Lower Leg	Upper Clothing
Left Shoe	Lower Lip
Left Sock	Upper Lip
Left Upper Arm	Lower Teeth
Left Upper Leg	Upper Teeth
Lower Clothing	Tongue
Right Foot	Object
Right Hand	

Figure 10. **HOI Masks Color Palette.** We visualize the color encoding of the 29 classes, where each color corresponds to a distinct part. The color scheme for human parts follows SAPIEN [39], and light gray is used for the object.

```
Task: Rewrite the caption as a concise
description of only the human and
foreground object motion.
Do not describe background, scene,
appearance, clothing, colors, lighting,
camera movement, or objects not
interacted with.
Focus only on the main action and use up
to four short sentences. Output as one
line without line breaks.
Caption: ``Original video caption``
```

Figure 11. Prompt template used to process text prompts for augmentor training. We prompt Qwen3 [95] to generate motion-centric captions aligned with the Sapien HOI-mask palette. We append the mask color scheme to the output: “Use Sapien human mask colors (fixed palette): background black; object gray; face neck baby blue; hair hot pink; torso green; left hand magenta, right hand red; left lower arm teal, right lower arm neon green; left upper arm orange, right upper arm dark blue.”

14. Augmentor prompt processing

For training the augmentor, the text prompt needs to get rid of the appearance, background and lighting detail, to facilitate HOI mask generation. We use Qwen3 [95] to distill a motion-centric prompt based on the original video caption. Lastly, we append the fixed Sapien palette legend so that the generated prompt always refers to the canonical mask colors used by the augmentor.