# IMAGIN-4D: IMAGE-GUIDED CONTROLLABLE INTERACTION GENERATION

**Anonymous authors**
Paper under double-blind review

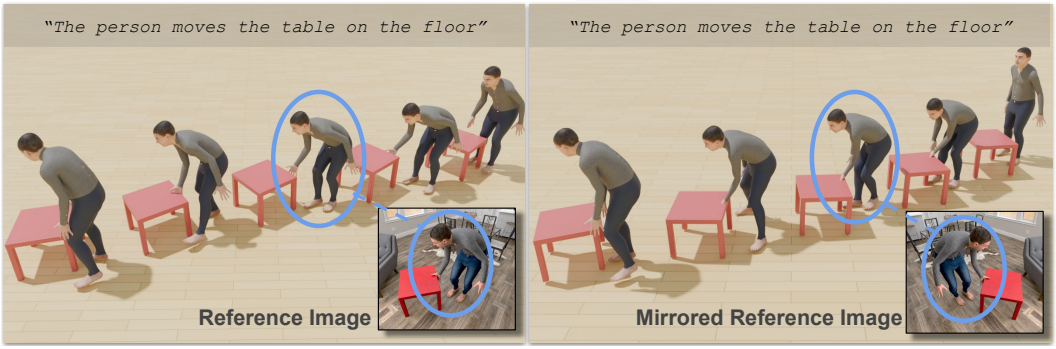



Figure 1: **Image-guided 4D HOI generation.** IMAGIN-4D synthesizes human-object interaction sequences from text, object geometry, waypoints, and a reference image. Text and waypoints specify the action and object trajectory, while the image resolves ambiguous pose, contact, and body-object layout. Mirroring only the reference image, while keeping all other inputs fixed, changes the generated body pose, object pose, contact, and layout to match the mirrored interaction snapshot.

## ABSTRACT

Generating human-object interactions (HOI) is central to character animation, robotics, AR/VR, and embodied AI. Existing methods use text, object geometry, and sparse waypoints to control actions and object trajectories, but leave grasps, approach directions, body and object poses, contacts, and relative layouts ambiguous. We use a reference image to specify the desired interaction snapshot. However, a single global image representation conflates distinct cues and uses the same visual evidence to condition every frame. Thus, we introduce IMAGIN-4D, a diffusion-based HOI generator with *spatio-temporal image conditioning*. For *spatial conditioning*, supervised interaction-state tokens encode body pose, object pose, body-object contact, and spatial relationships at the depicted frame. For *temporal conditioning*, frame-aware tokens query image patches per generated frame to retrieve visual cues. Role-aware conditioning balances these signals: text, waypoints, and interaction-state tokens use separate AdaLN streams, while motion tokens cross-attend to frame-aware visual tokens. Since most HOI MoCap datasets, including FullBodyManipulation (FBM), lack paired images, we render synthetic FBM references; for BEHAVE, we use registered RGB frames directly. Experiments on FBM and BEHAVE show that IMAGIN-4D improves fine-grained interaction control over single-token and uniformly image-conditioned baselines while preserving waypoint-following and motion quality.

## 1 INTRODUCTION

Generating 4D human-object interactions (HOI) is central to character animation, AR/VR, robotics, and embodied AI. This is challenging, because interactions need to look natural both spatially and temporally; that is, bodies and objects need to be posed with realistic contacts and proximal relationships in 3D space, and also move realistically over time. Moreover, it is challenging to computationally describe and exploit such spatial and temporal cues for effectively controlling the generation



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.

process. Recent diffusion-based methods (Xu et al., 2023; Li et al., 2024) synthesize plausible HOI motion from text, object shape, and sparse waypoints, providing control over action semantics and object trajectories. However, these signals underspecify interactions; the same prompt and trajectory can correspond to different grasps, approaching directions, body poses, object poses, and contact or spatial configurations. These details matter for instruction-following agents, e.g., grasping a mug by its handle or opening a drawer correctly.

To reduce ambiguities, one can manually specify 3D HOI keyposes. However, handcrafting these is cumbersome, and capturing them via MoCap is expensive. Instead, we use a reference image to intuitively represent a desired interaction "snapshot"; this may be a photo, a rendering, a sketch, or AI-generated. Thus, our goal is to synthesize HOI motion conditioned on a text prompt, object geometry, sparse waypoints, and a reference image of an HOI snapshot; see Fig. 1. Similar conditioning has recently been explored by ViHOI (Cai et al., 2026). ViHOI extracts a visual token from reference images to guide motion generation. However, this design has two limitations: First, extracting a single global token abstracts distinct, fine-grained spatial cues. Second, these cues influence not only the timeframe of the reference snapshot, but also the motion frames before and after this, because motion is continuous and coherent. Consequently, as we show in our experiments, extracting a single global token compromises fine-grained control necessary for motion generation.

Thus, we introduce IMAGIN-4D, a diffusion-based HOI generator with *spatio-temporal image conditioning*. Our key idea is that effective image conditioning needs to have both spatial and temporal influence on motion generation. For *spatial influence*, IMAGIN-4D extracts multiple, supervised spatial tokens that capture body pose, object pose, and body-object contact and spatial relationships. To obtain these tokens, we introduce a "Spatially-Factorized Image Encoder" (SFIE) that applies separate Q-Former heads (Li et al., 2023b) to image patches. Therefore, the motion generator uses multiple, fine-grained tokens for the reference snapshot instead of just a single one. However, this informs only the snapshot timeframe.

For *temporal influence* we need to inform all other frames as well. Thus, we introduce *frame-aware tokens*, by querying image patches separately for each motion frame, conditioned on the frame index and text prompt. Unlike spatial tokens, frame-aware tokens are not supervised with explicit targets. Instead, they are learned end-to-end through the denoising objective, allowing the model to learn image cues that are informative for each motion frame. The spatial and frame-aware cues encode low-level details of interactions. However, motion generators typically use high-level cues, i.e., text for action semantics and waypoints for the object trajectory. To prevent these cues from dominating over each other, IMAGIN-4D does not concatenate conditioning tokens, but conditions on them differently, via a novel "*role-aware conditioning*". Specifically, spatial tokens, waypoints, and text use separate *AdaLN* streams (Peebles & Xie, 2023), preserving their roles rather than mixing them. In contrast, the frame-aware tokens are *cross-attended* with the learnable motion tokens; this selectively adapts the influence of frame-dependent cues on each motion token.

Training image-conditioned HOI synthesis requires images paired with motion data. However, most HOI MoCap datasets, including FBM, provide no paired images; BEHAVE provides RGB frames registered to its motion, which we use directly. We thus render images from FullBodyManipulation (FBM) sequences (Li et al., 2023a) using body textures (Black et al., 2023), objects, and indoor scenes (Straub et al., 2019). For evaluation, standard metrics such as FID, R-precision, and contact F1 measure motion quality, semantic alignment, or contact accuracy, but do not measure the quality of image conditioning. Thus, we introduce the *image-adherence metric* that evaluates to what extent the generated motion matches the reference snapshot.

We evaluate IMAGIN-4D on the challenging FBM and BEHAVE (Bhatnagar et al., 2022), with held-out object categories and cross-domain image inputs. Results show that the global image conditioning with a single token as in prior work improves motion metrics such as FID and R-Precision, but fails to control the depicted interaction state. In contrast, our spatially factorized tokens substantially improve image adherence; adding frame-aware tokens further improves motion quality, and our full training recipe additionally improves contact accuracy. As a downstream application, we retrain the image branch on line drawings to obtain a sketch-to-motion variant, showing that the same conditioning mechanism extends beyond RGB references to user-editable visual inputs.

In summary, our contributions are:

- We introduce a HOI generator with a novel *spatio-temporal image conditioning*, using a reference interaction snapshot for controllable motion generation.
- We introduce *role-aware conditioning*: spatial tokens, waypoints, and text use separate AdaLN streams, and frame-aware tokens use cross-attention, improving adherence and motion quality.
- We also introduce a synthetic *motion-to-image rendering* pipeline and an *image-adherence metric* for evaluating to what extent the generated HOI motions match the reference interaction snapshot.

## 2 Related Work

**Data for dynamic HOI.** Whole-body HOI datasets (Taheri et al., 2020; Bhatnagar et al., 2022; Huang et al., 2022; Jiang et al., 2023b; Li et al., 2023a; Zhao et al., 2024; Lv et al., 2024; Kim et al., 2025; Zhang et al., 2024b; Lu et al., 2025) capture paired human-object motion, while hand-object benchmarks (Fan et al., 2023; Liu et al., 2022; Zhan et al., 2024; Liu et al., 2024) focus on manipulation and scene-aware corpora. Some datasets (Hassan et al., 2019; Wang et al., 2022; Jiang et al., 2024b) place motion in 3D environments. InterAct (Xu et al., 2025a) unifies several sources under a representation. These datasets support learning motion, contact, and scene constraints, but they are not designed for image-conditioned HOI generation. Most lack images; BEHAVE (Bhatnagar et al., 2022) provides registered RGB frames, which we use directly as references. For FullBodyManipulation (Li et al., 2023a), which has no images, we render conditioning images at known interaction frames using textured bodies (Black et al., 2023), objects, and indoor scenes (Straub et al., 2019).

**Controllable human motion generation.** Text-to-motion diffusion models (Ho et al., 2020; Tevet et al., 2023; Zhang et al., 2024c; Chen et al., 2023) and tokenized generators (Zhang et al., 2023a; Jiang et al., 2023a; Guo et al., 2024; Zhang et al., 2023c;d) synthesize human motion from language. Additional control is introduced through guidance (Karunratanakul et al., 2023), joint or keyframe constraints (Xie et al., 2024; Cohan et al., 2024), sparse trackers (Barquero et al., 2025), trajectory conditioning (Shafir et al., 2024; Wan et al., 2024; Karunratanakul et al., 2024), or masked control (Tessler et al., 2024). These interfaces work when the desired motion is expressible as body trajectories, masks, or key poses. They become cumbersome for HOI because the target state couples human pose, object pose, body-object layout, and contact. A reference image specifies this coupled state directly.

**4D Human-object interaction synthesis.** HOI synthesis models human motion, object motion, and their coupling through contact. Prior work studies hand-object manipulation (Taheri et al., 2022; 2024; Zhou et al., 2022; Zhang et al., 2024a; Christen et al., 2024), whole-body object interaction (Zhang et al., 2022; Ghosh et al., 2023; Li et al., 2023a; Xu et al., 2023), and language- or trajectory-conditioned 4D HOI generation (Diller & Dai, 2024; Peng et al., 2025; Cha et al., 2024; Song et al., 2024; Li et al., 2024; Ron et al., 2026). CHOIS (Li et al., 2024) is the closest non-image-based baseline, since it controls long-horizon interactions with text, object geometry, and object waypoints. Other methods incorporate scene affordances, interaction fields, simulation, or priors distilled from image, video, and VLM models (Wang et al., 2024; Jiang et al., 2024b;a; Yi et al., 2024; Kulkarni et al., 2024; Yuan et al., 2023; Wang et al., 2023; Xu et al., 2025b; Li & Dai, 2024; Xu et al., 2024a; Li et al., 2026; Zhang et al., 2025). These methods improve realism and controllability, but most expose control through text, trajectories, contact maps, key poses, scene geometry, or priors. Our work instead uses a single reference image as the specification of the desired interaction state.

**Image-conditioned HOI and motion synthesis.** Image conditioning is used in 2D generation to constrain appearance, layout, identity, pose, or motion beyond text (Ye et al., 2023; Mou et al., 2024; Li et al., 2023c; Huang et al., 2023; Hu et al., 2024; Xu et al., 2024b; Zhu et al., 2024). Its use for 4D HOI generation is recent. ViHOI (Cai et al., 2026) is closest to our setting: it extends a CHOIS-style generator with a frozen VLM and Q-Former adapter, then injects image-text tokens into the motion model. MP-HOI (Wang et al., 2026) fuses text, image, and pose priors, whereas SIGHT (Gavryushin et al., 2025) and IKMo (Zhao et al., 2025) use image and text conditioning to guide the motion. While these methods show that images can guide motion, they typically treat the image as a global condition, sparse anchor, or external prior. In HOI, this discards structure: the same image contains contact, body pose, object pose, body-object offset, and layout, and these cues matter at different times. Our method preserves this structure through supervised interaction-state tokens and frame-aware visual retrieval.

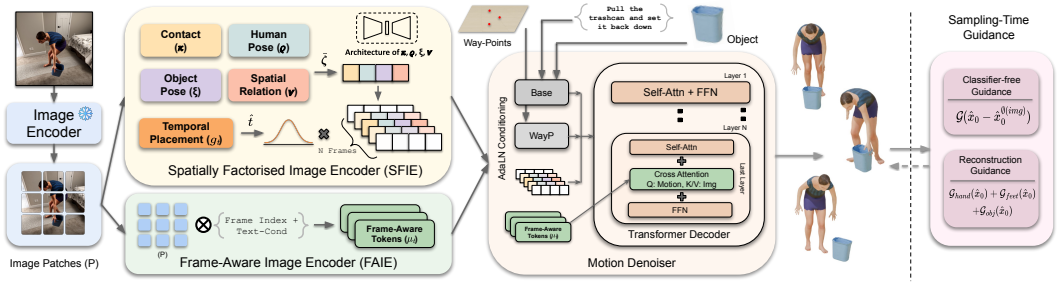



Figure 2: **IMAGIN-4D overview (Sec. 3).** Given a reference image $\mathcal{I}$, text prompt $y$, object geometry $\mathcal{O}$, and sparse waypoints $\mathcal{W}$, IMAGIN-4D generates a 4D human-object motion sequence. A frozen image encoder extracts patch tokens $\mathbf{P}$ from $\mathcal{I}$. The Spatially Factorized Image Encoder (SFIE) reads these patches with role-specific queries and produces supervised latent tokens for contact $\boldsymbol{\kappa}$, human pose $\boldsymbol{\rho}$, object pose $\boldsymbol{\xi}$, and body-object spatial relation $\boldsymbol{\nu}$. These tokens are trained to match role-autoencoder latents derived from the paired motion sequence. Their concatenated summary $\bar{\boldsymbol{\zeta}}$ predicts the reference frame $\hat{t}$ depicted by the image. In parallel, the Frame-Aware Image Encoder re-queries $\mathbf{P}$ with frame- and text-conditioned queries to produce per-frame visual tokens $\boldsymbol{\mu}_t$. The motion denoiser routes conditions by role: base conditioning, waypoint features, and window-gated spatial image evidence modulate transformer layers through separate AdaLN streams, while $\boldsymbol{\mu}_t$ enters through late cross-attention. Sampling-time guidance improves image adherence

## 3 METHOD

IMAGIN-4D generates a human-object motion from a text prompt $\mathbf{Y}$, object shape $\mathcal{O}$, sparse trajectory waypoints $\mathcal{W}$, and a reference image $\mathcal{I}$. The output is a $T$-frame motion sequence $\mathbf{x}_0 \in \mathbb{R}^{T\times D}$, where $D$ denotes the motion dimensionality of one frame. The reference image depicts a desired interaction snapshot, corresponding to frame $t^\star$ in the target sequence. During training, $t^\star$ is known from the paired image-motion data; at test time, it is predicted.

The reference image resolves interaction ambiguities left by text and sparse waypoints, including human pose, object pose, and body-object contact and layout. IMAGIN-4D uses two image representations. First, a *Spatially Factorized Image Encoder* (SFIE) extracts supervised role-specific spatial tokens for the reference snapshot. Second, *frame-aware tokens* re-query the same image patches for each motion frame. The denoiser routes text, sparse waypoints, and window-gated spatial image evidence through separate AdaLN (Peebles & Xie, 2023) streams, and injects frame-aware visual tokens through late cross-attention. Fig. 2 summarizes the architecture.

### 3.1 CONDITIONAL MOTION FORMULATION

Let $\mathbf{x}_0 \in \mathbb{R}^{T\times D}$ denote a HOI motion sequence. Following CHOIS (Li et al., 2024), sparse controls are represented by a binary mask $\mathbf{m} \in \{0,1\}^{T\times D}$, where $m_{t,d} = 0$ denotes a input-specified value and $m_{t,d} = 1$ denotes a value to synthesize. The sparse conditioning tensor is $\mathbf{x}_{\text{cond}} = (1-\mathbf{m}) \odot \mathbf{x}_0$. The sparse tensor and mask tell a diffusion denoiser which motion values are user-specified and which must be generated. In our setting, $\mathbf{x}_{\text{cond}}$ contains the initial human and object pose, sparse object waypoints, and final object translation. Object shape is encoded with a BPS encoder (Prokudin et al., 2019), yielding an object descriptor $\mathcal{O}_{BPS}$.

We use conditional diffusion for motion generation. At diffusion step $n$, the denoiser, $\mathcal{D}_\theta$, receives a noisy motion sequence $\mathbf{x}_n$ and predicts the clean motion, $\hat{\mathbf{x}}_0 = \mathcal{D}_\theta(\mathbf{x}_n, n, \mathbf{c})$, where $\mathbf{c} = \{\mathbf{Y}, \mathcal{O}_{BPS}, \mathbf{x}_{\text{cond}}\}$ contains text, object shape, sparse controls.

### 3.2 SPATIO-TEMPORAL IMAGE CONDITIONING

**Spatially Factorized Image Encoder (SFIE).** A pooled image embedding entangles pose, contact, and layout, which constrain the motion. Thus, we encode the reference image into role-specific spatial tokens. A frozen DINOv2 image encoder (Oquab et al., 2024) extracts patch tokens $\mathbf{P} \in \mathbb{R}^{S\times C}$ from $\mathcal{I}$, where $S$ denotes image patches and $C$ the feature dimension. For each role $h \in \mathcal{H}$, a lightweight Q-Former (Li et al., 2023b) reads $\mathbf{P}$ with learned queries $\mathbf{Q}_h$:

$$\mathbf{F}_h = \mathrm{QFormer}_h(\mathbf{P}; \mathbf{Q}_h), \qquad h \in \mathcal{H}. \tag{1}$$

Here, $\mathbf{F}_h$ denotes the token output for role $h$. For human pose, object pose, contact, and layout, these outputs are $\boldsymbol{\rho}$, $\boldsymbol{\xi}$, $\boldsymbol{\kappa}$, and $\boldsymbol{\nu}$. Each token is supervised to match the latent code of a lightweight role autoencoder trained on the corresponding ground-truth interaction quantity. We concatenate and project the tokens into $\bar{\boldsymbol{\zeta}}$, used by the reference-frame localizer and image-AdaLN stream.

**Role supervision.** Each role is supervised at the ground-truth reference frame $t^\star$ using targets derived from the paired motion sequence. The contacts decoders reconstruct hand contact positions and binary contact flags. The human-pose decoder predicts main body-joint pose. The object-pose decoder predicts object translation and orientation. The spatial-relation decoder predicts per-joint unit vectors from body joints to the object center. The SFIE loss is

$$\begin{aligned}\mathcal{L}_{\mathrm{SFIE}} = \lambda_{\mathrm{con,pos}}\mathcal{L}^{\mathrm{mse}}_{\mathrm{con,pos}} + \lambda_{\mathrm{con,flag}}\mathcal{L}^{\mathrm{bce}}_{\mathrm{con,flag}} + \lambda_{\mathrm{hum}}\mathcal{L}^{\mathrm{mse}}_{\mathrm{hum}} \\ + \lambda_{\mathrm{obj}}\mathcal{L}^{\mathrm{mse}}_{\mathrm{obj}} + \lambda_{\mathrm{spa}}\mathcal{L}^{\mathrm{cos}}_{\mathrm{spa}},\end{aligned} \tag{2}$$

where $\lambda$ denotes loss weights. The contact-position, contact-flag, human-pose, and object-pose losses supervise their corresponding decoder outputs, using MSE for continuous quantities and BCE for binary contact flags. The spatial loss $\mathcal{L}^{\mathrm{cos}}_{\mathrm{spa}}$ uses cosine distance between predicted and ground-truth joint-to-object unit vectors, making the spatial token encode body-object layout rather than absolute distance.

**Reference-frame localizer.** The reference image depicts one moment of the target sequence, but the corresponding frame is unknown at test time. We predict it from the spatial-token summary:

$$\boldsymbol{\pi} = \mathrm{softmax}\big(W_2\,\mathrm{GELU}(W_1\bar{\boldsymbol{\zeta}})\big) \in \mathbb{R}^T. \tag{3}$$

Here, $W_1, W_2$ are MLP weights, and $\pi_k$ is the $k$-th entry of $\boldsymbol{\pi}$, giving the probability that frame $k$ matches the reference image. During training, we supervise this distribution with a Gaussian-smoothed target centered at $t^\star$:

$$\mathcal{L}_{\mathrm{fp}} = -\sum_{k=0}^{T-1} q_k \log \pi_k, \qquad q_k = \frac{\exp\left(-(k-t^\star)^2/(2\sigma_q^2)\right)}{\sum_{r=0}^{T-1}\exp\left(-(r-t^\star)^2/(2\sigma_q^2)\right)}. \tag{4}$$

The predicted frame is $\hat{t} = \arg\max_k \pi_k$. All image-window operations use $\hat{t}$ during both training and inference, so the denoiser is trained under the same localization errors it sees at test time. Gradients from the denoising loss are stopped through $\hat{t}$; the localizer is trained only with Eq. (4).

**Window-gated spatial image conditioning.** The spatial tokens describe the reference snapshot and constrain motion most strongly near $\hat{t}$. Applying them uniformly can falsely constrain several frames before and after $\hat{t}$ toward the depicted pose and contact state. We therefore gate the spatial-token summary with a temporal window:

$$\tilde{\boldsymbol{\zeta}}_t = r_t\, w_{\sigma_g}(t-\hat{t})\,\bar{\boldsymbol{\zeta}}. \tag{5}$$

Here, $w_{\sigma_g}$ is a smooth unit-peak temporal window centered at $\hat{t}$. The gate $r_t$ is set to 0 at frames with specified waypoint constraints and to 1 otherwise. This prevents image modulation from competing with explicit trajectory constraints.

**Frame-Aware Image Encoder (FAIE).** The supervised spatial tokens summarize the reference snapshot, but do not provide frame-specific evidence for the rest of the sequence. Frames near $\hat{t}$ require precise contact and pose cues, while distant frames rely more on object identity, coarse layout, or approach direction. We therefore re-query image patches for each motion frame:

$$\boldsymbol{\mu}_t = \mathrm{QFormer}_{\mathrm{fvt}}\left(\mathbf{P}; \mathbf{Q}_{\mathrm{fvt}}(t, \mathbf{e}_y)\right), \qquad \mathbf{M} = \{\boldsymbol{\mu}_t\}_{t=1}^{T}. \tag{6}$$

Here, $\mathbf{e}_y$ is the text embedding, $\mathbf{Q}_{\mathrm{fvt}}(t, \mathbf{e}_y)$ denotes frame- and text-conditioned queries, and $\mathbf{M}$ is the set of frame-aware image tokens used for cross-attention. The frame-aware token module is trained only through the denoising objective.

### 3.3 Role-Aware Motion Denoiser

The denoiser receives conditioning signals with distinct roles. Text provides sequence-level action semantics. Sparse waypoints provide trajectory control. Window-gated spatial image evidence constrains frames around the reference snapshot. Frame-aware tokens provide frame-dependent image evidence. A shared conditioning path can let dense image evidence perturb frames where waypoints already specify the object trajectory. We therefore route each signal by role.

Let $\mathbf{c}_{\text{text}}$ denote the text-conditioning vector combined with the diffusion-step embedding. Let $\mathbf{z}_t^{\text{wpt}}$ denote the sparse-constraint vector from $\mathbf{x}_{\text{cond}}$ and $\mathbf{m}$ at frame $t$. AdaLN converts each conditioning signal into per-layer modulation parameters for transformer. At transformer layer $\ell$, the AdaLN parameters for frame $t$ are

$$\boldsymbol{\eta}_t^{(\ell)} = \text{AdaLN}_{\text{text}}^{(\ell)}(\mathbf{c}_{\text{text}}) + \text{AdaLN}_{\text{wpt}}^{(\ell)}(\mathbf{z}_t^{\text{wpt}}) + \text{AdaLN}_{\text{img}}^{(\ell)}(\tilde{\boldsymbol{\zeta}}_t). \tag{7}$$

Here, $\boldsymbol{\eta}_t^{(\ell)}$ contains shift, scale, and residual-gate parameters used by the transformer block. Separate streams let the model adjust image modulation without changing parameters that encode text semantics or waypoint constraints. The image AdaLN stream is zero-initialized, so training starts from the text-and-waypoint model. Frame-aware tokens are routed differently. Instead of modulating every transformer layer, the denoiser reads visual memory $\mathbf{M}$ through cross-attention in the final decoder layer. Each motion token accesses frame-dependent image evidence after self-attention integrates temporal context and sparse trajectory constraints.

### 3.4 Synthetic Image Generation

Image-guided motion generation requires human-object motion *paired* with reference images of a chosen interaction frame. BEHAVE (Bhatnagar et al., 2022) provides RGB frames registered to the body and object, which we use directly as references. FullBodyManipulation (FBM) (Li et al., 2023a) provides MoCap without images, so we render images using its SMPL-X (Pavlakos et al., 2019) body meshes, inserting the posed object, and assigning sequence-level body and object appearances. We create three image domains. *MeshImg* renders the body and object on a white background. *SceneImg* places the same posed body-object pair in Replica indoor scenes (Straub et al., 2019) and applies BEDLAM body textures (Black et al., 2023). *EditImg* applies FLUX.2-dev (Black Forest Labs, 2025) to SceneImg renders to produce more photorealistic test-time references. EditImg is used only for evaluation. For FBM, SceneImg is used for the main experiments and architecture ablations, and MeshImg for the cross-domain analysis. For BEHAVE, training and evaluation use the dataset's real RGB frame at the reference frame; no rendering is involved.

For each 120-frame window, we render a temporal grid and contact-centered frame $t^\star$, defined as the contact-weighted temporal centroid of the left- and right-hand contact flags; see Appendix J for definition and rendering details. During training, one rendered image is sampled per sequence per epoch; at evaluation, the contact-centered frame is the canonical reference for image adherence.

### 3.5 Training and Sampling

**Denoising loss.** The denoiser predicts $\mathbf{x}_0$ directly. Since image conditioning should matter most near the reference snapshot, we upweight reconstruction errors around the predicted reference frame:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(1 + \lambda_{\text{cf}}\exp\left(-\frac{(t-\hat{t})^2}{2\sigma_{\text{cf}}^2}\right)\right)\|\mathbf{x}_{0,t} - \hat{\mathbf{x}}_{0,t}\|_1\right]. \tag{8}$$

Both this loss and the image-conditioning gate use $\hat{t}$, matching training and inference behavior.

**Auxiliary losses.** Following prior HOI generation work (Li et al., 2024), we add a forward-kinematics loss $\mathcal{L}_{\text{FK}}$, object-point loss $\mathcal{L}_{\text{objpts}}$, and foot-contact loss $\mathcal{L}_{\text{feet}}$. We additionally use a hand-object coupling loss $\mathcal{L}_{\text{cpl}}$ that ties object-centroid displacement to that of the contacting hand at ground-truth contact frames, and an image-consistency loss $\mathcal{L}_{\text{img}}$ at reference frame $t^\star$. It compares interaction quantities decoded from the generated frame with reference-image targets, including human pose, object pose, body-object contact, and layout. Its weight increases at lower-noise diffusion steps, where $\hat{\mathbf{x}}_0$ is reliable for frame-level geometric losses. The full objective is:

| Method | Img Enc | Image Adherence (cm) ↓ | | | FID ↓ | R-Precision ↑ | | | Contact ↑ | | | Interaction ↓ | | WPErr ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_{\mathbf{GT}}$ | $A_{\mathbf{W10}}$ | $A_{\mathbf{Any}}$ | | R@1 | R@2 | R@3 | $C_{\text{prec}}$ | $C_{\text{rec}}$ | $C_{F_1}$ | FootS | HandP | |
| InterDiff (Xu et al., 2023) | – | – | – | – | 20.80 | – | – | 0.08 | 0.63 | 0.28 | 0.33 | 0.42 | 0.55 | 72.72 |
| MDM (Tevet et al., 2023) | – | – | – | – | 6.16 | – | – | 0.51 | 0.72 | 0.47 | 0.53 | 0.48 | 0.66 | 19.42 |
| CHOIS (Li et al., 2024) | – | – | – | – | 0.69 | 0.322 | 0.534 | 0.64 | 0.80 | 0.64 | 0.67 | **0.35** | 0.59 | **2.87** |
| CHOIS+Img | CLIP | 19.90 | 16.83 | 13.08 | 0.47 | 0.333 | 0.544 | 0.688 | 0.796 | 0.625 | 0.655 | 0.41 | 0.58 | 3.31 |
| CHOIS+Img | Qwen2.5-VL | 20.64 | 17.40 | 13.11 | 0.75 | 0.317 | 0.510 | 0.653 | 0.798 | 0.534 | 0.640 | 0.40 | 0.62 | 3.93 |
| ViHOI$^\star$ (Cai et al., 2026) | Qwen2.5-VL | 20.55 | 17.61 | 13.33 | 0.71 | 0.313 | 0.517 | 0.670 | 0.800 | 0.611 | 0.648 | 0.42 | 0.59 | 3.43 |
| CHOIS+Img | DINOv2 | 19.61 | 16.48 | 13.23 | 0.63 | 0.330 | 0.543 | 0.687 | 0.805 | 0.587 | 0.636 | 0.45 | 0.60 | 3.70 |
| **Ours (Temporal)** | DINOv2 | 19.07 | 16.23 | 12.56 | 0.52 | 0.340 | 0.540 | 0.701 | 0.805 | 0.669 | 0.697 | 0.39 | 0.62 | 6.37 |
| **Ours (Spatial)** | DINOv2 | 10.11 | 8.91 | 7.95 | 0.30 | 0.345 | 0.556 | 0.706 | 0.811 | 0.621 | 0.663 | 0.44 | 0.56 | 6.28 |
| **Ours (Spatial+Temporal)** | DINOv2 | **7.75** | **7.11** | **6.98** | **0.25** | **0.352** | **0.569** | **0.715** | **0.815** | **0.766** | **0.768** | 0.41 | **0.53** | 5.10 |

Table 1: **Evaluation on the FullBodyManipulation dataset (Sec. 4.1).** All methods are evaluated on the FullBodyManipulation (Li et al., 2023a) test set using text, object geometry, and sparse object waypoints as conditions; image-conditioned methods additionally use one rendered *SceneImg* reference at the contact-centered frame. We report image adherence at the ground-truth frame ($A_{\text{GT}}$), within a $\pm$10-frame window ($A_{\text{W10}}$), and over the full sequence ($A_{\text{Any}}$), along with motion quality (FID), text-motion alignment (R-Precision), hand-object contact, interaction artifacts (foot sliding, hand penetration), and waypoint error. The image-adherence score is not applicable for image-free methods. ViHOI$^\star$ denotes our single-image adaptation of ViHOI (Cai et al., 2026).

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{SFIE}} + \lambda_{\text{fp}}\mathcal{L}_{\text{fp}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} \\ + \lambda_{\text{FK}}\mathcal{L}_{\text{FK}} + \lambda_{\text{objpts}}\mathcal{L}_{\text{objpts}} + \lambda_{\text{feet}}\mathcal{L}_{\text{feet}} + \lambda_{\text{cpl}}\mathcal{L}_{\text{cpl}}.\end{aligned} \tag{9}$$

All loss weights, optimizer settings, learning-rate schedule, data augmentation, and remaining hyperparameters are reported in Appendix O, and architecture details in Appendix N.

**Image classifier-free guidance.** We use classifier-free guidance on image condition (Ho & Salimans, 2022). During training, with probability $p_{\text{drop}}$, we replace all image inputs by their null form: spatial tokens $\boldsymbol{\zeta}$ and frame-aware tokens $\mathbf{M}$ are zeroed. Text, object shape, and sparse controls remain unchanged. At sampling, we run two denoiser forwards: one with full condition and one with dropped image condition. Guided output is

$$\hat{\mathbf{x}}_0^{\text{cfg}} = \hat{\mathbf{x}}_0 + (s_{\text{img}} - 1)\left(\hat{\mathbf{x}}_0 - \hat{\mathbf{x}}_0^{\setminus\text{img}}\right), \tag{10}$$

where $\hat{\mathbf{x}}_0$ is the full-conditioning estimate, $\hat{\mathbf{x}}_0^{\setminus\text{img}}$ uses null image inputs, and $s_{\text{img}}$ is the image-guidance scale. Setting $s_{\text{img}} = 1$ recovers standard conditional prediction.

**Mesh-level hand and foot guidance.** Following CHOIS (Li et al., 2024), we apply reconstruction guidance during the last $K = 20$ denoising steps to reduce hand-object penetration and foot-floor artifacts. At each guided step, we compute a mesh-level penalty on the predicted clean motion and take a variance-scaled gradient step before sampling $\mathbf{x}_{n-1}$. The penalty combines a hand term, penalizing hand-object penetration and attracting the hand to the object at predicted contact frames, and a foot term, penalizing support-foot deviation from the floor. We restrict this guidance to the final denoising steps as the mesh penalties assume that $\hat{\mathbf{x}}_0$ is already close to a valid clean motion.

## 4 Experiments

**Datasets & image protocol.** We follow the CHOIS protocol (Li et al., 2024) on FullBodyManipulation (FBM) (Li et al., 2023a) and BEHAVE (Bhatnagar et al., 2022; Xu et al., 2025a). FBM uses the CHOIS splits and text prompts, with nine training and four held-out object categories; BEHAVE sequences use 120-frame windows. FBM comparisons and ablations use rendered *SceneImg* references; BEHAVE uses its real RGB frames at the reference frame (Appendix M). Rendering, reference-frame selection, and the additional image domains are detailed in Sec. 3.4 and Appendix J.

**Metrics.** Our image-adherence metric measures agreement with the reference interaction state at $t^\star$ ($A_{\text{GT}}$), within $\pm$10 frames ($A_{\text{W10}}$), and across the sequence ($A_{\text{Any}}$), in cm. We also report FID and R-Precision using the frozen CHOIS evaluator, hand-object contact precision/recall/$F_1$, foot sliding, hand-object penetration, and object-trajectory error at sparse waypoints (definitions in Appendix K).

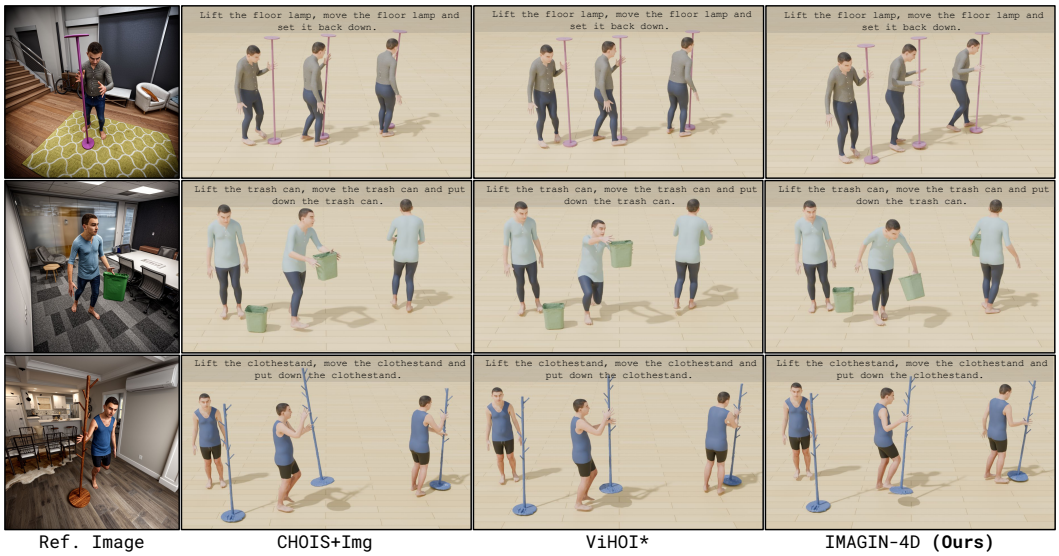



Figure 3: **Qualitative comparison on FullBodyManipulation (Sec. 4.1).** Each row shows a reference image (*EditImage*) and generated motion from CHOIS+Img, ViHOI⋆, and IMAGIN-4D, all conditioned on the same text prompt, object shape, waypoints, and reference image. Single-token image conditioning often misses fine-grained details, such as contact, hand placement, or body-object layout, while IMAGIN-4D better matches the depicted interaction state.

| **Method** | $A_{GT}\downarrow$ | $A_{W10}\downarrow$ | $A_{Any}\downarrow$ | FID$\downarrow$ | $\mathbf{C_p}\uparrow$ | $\mathbf{C_r}\uparrow$ | $\mathbf{C_{F_1}}\uparrow$ | WP$\downarrow$ |
|---|---|---|---|---|---|---|---|---|
| CHOIS (Li et al., 2024) | – | – | – | 1.39 | 0.526 | 0.351 | 0.375 | **3.05** |
| CHOIS+Img | 22.02 | 18.81 | 13.71 | 1.57 | 0.521 | 0.361 | 0.386 | 4.07 |
| ViHOI* (Cai et al., 2026) | 22.97 | 20.34 | 14.55 | 1.44 | 0.518 | 0.348 | 0.378 | 4.49 |
| **IMAGIN-4D** | **12.40** | **12.23** | **11.18** | **0.76** | **0.538** | **0.381** | **0.397** | 4.93 |

Table 2: **Comparison on BEHAVE (Sec. 4.1).** All image-conditioned methods are trained and tested on BEHAVE using single *MeshImg* reference at the contact-centered frame. We report image adherence at the ground-truth frame ($A_{GT}$), within a $\pm$10-frame window ($A_{W10}$), and over the full sequence ($A_{Any}$), plus FID, contact precision/recall/$F_1$, and waypoint error (WP). ViHOI⋆ denotes our single-image adaptation of ViHOI. IMAGIN-4D improves adherence, FID, and contact over CHOIS+Img and ViHOI⋆.

**Baselines and implementation.** Image-free baselines are CHOIS (Li et al., 2024), InterDiff (Xu et al., 2023), and MDM (Tevet et al., 2023). We retrain CHOIS with a pooled image token from CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), or Qwen-VL (Cai et al., 2026). We re-implement ViHOI (Cai et al., 2026), adapting its three-image input to one and denote it ViHOI⋆ (Appendix L). Appendix F.1 compares patch encoders for our factorized heads. All image-conditioned FBM rows, including the baselines, are sampled with image guidance scale $s_{\text{img}}$=2; Tab. 3 (B) reports the corresponding ablation with guidance off.

### 4.1 Main Comparison on FBM and BEHAVE

Table 1 evaluates whether image conditioning improves fine-grained interaction control on FBM without degrading motion quality or waypoint following. CHOIS remains the strongest waypoint follower, reaching WPErr $=$ 2.87 cm because it has no image condition "competing" with other condition signals. Image-free methods do not receive a reference image, so image-adherence metrics are reported only for image-conditioned methods.

Single-global-token baselines, CHOIS+Img and ViHOI⋆, improve global motion metrics such as FID but show limited image adherence. Across CLIP, DINOv2, and Qwen2.5-VL encoders, their $A$_Any remains 13.08 and 13.33 cm. With Qwen2.5-VL, ViHOI⋆ improves FID over CHOIS+Img

but has comparable adherence. This suggests a representation bottleneck: one image token compresses spatial evidence needed to recover body pose, object pose, contact, and layout.

Our spatial encoder addresses this bottleneck with supervised interaction tokens localized at the conditioning frame. **Ours (Spatial)** reduces $A_{\text{Any}}$ from $12.56$ to $7.95$ cm relative to **Ours (Temporal)** that uses global image representation from DINOv2 and improves FID from $0.52$ to $0.30$. Spatial tokens give strong reference-frame control, but they are fixed across the sequence. **Ours (Spatial+Temporal)** adds frame-aware tokens that query the image patches for each motion frame, testing whether image evidence can also improve the surrounding motion. The full model achieves the lowest image-adherence errors and FID ($0.25$), and the highest contact precision, recall, and $F_1$ in Tab. 1; its R-Precision is within $0.001$ of Ours (Spatial). Waypoint error increases relative to CHOIS because image and waypoint conditions compete for control; in the controlled ablations of Tab. 3, merging the image AdaLN stream into the base stream reduces waypoint error from $5.69$ to $3.8$ cm but increases $A_{\text{GT}}$ from $8.43$ to $13.4$ cm (Appendix F).

Figure 3 shows the qualitative failure mode under photorealistic EditImg conditioning; Fig. 4 in Appendix A shows further comparisons with SceneImg references. Baselines using global image representation with a single token often preserve action category but miss the depicted contact, object orientation, or hand-object layout.

Table 2 repeats the comparison on BEHAVE using the dataset's real RGB frames as references, testing whether spatial image factorization transfers to a different motion distribution and to real images. The same trend holds: spatial factorization improves image adherence over global image conditioning, showing that the gain is not specific to FBM. This isolates the effect of the image representation under real-image conditioning.

**Architecture ablations.** In the controlled ablations, removing any SFIE head increases image-adherence error; removing the layout head increases $A_{\text{Any}}$ most, from $7.45$ to $9.0$ cm. Replacing cross-attention with AdaLN for frame-aware tokens increases FID from $0.28$ to $0.60$ (Appendix F).

### 4.2 Qualitative Results

The image-flip probe (Fig. 1) flips only the reference image, keeping text, waypoints, object geometry, conditioning frame, and random seed fixed. The generated grasp side and contact region change with the image, while motion is not mirrored since the non-image conditions are fixed (Appendix C).

**Additional experiments.** The appendix reports cross-domain image transfer (Appendix G), sketch-to-motion (Appendix D), physical quality (Appendix H), conditioning-frame sensitivity (Appendix I), and perceptual studies of motion plausibility (Appendix B) and adherence to real images (Appendix A, Fig. 5). Alternative control interfaces are discussed in Appendix P.

## 5 Conclusion

We present IMAGIN-4D, an image-guided HOI generator that uses a reference interaction snapshot to control 4D human-object motion. Our experiments show that conditioning motion generation on a single pooled image token improves motion-level metrics such as FID and R-Precision, but provides weak control over the depicted contact, body pose, object pose, and body-object layout. In contrast, IMAGIN-4D's Spatially-Factorized Image Encoder (SFIE) extracts separate interaction-state tokens, substantially improving image adherence, while frame-aware visual tokens improve motion quality by adapting image conditioning to each generated frame. Role-aware conditioning further integrates these visual cues with text and object waypoints through separate conditioning pathways rather than concatenating them into a single stream. We validate this control on FBM and BEHAVE using rendered FBM references and real BEHAVE images, and further show that the same conditioning extends to sketches. Across these settings, IMAGIN-4D provides substantially finer control over the generated interaction than text and sparse waypoints alone. Together, these results show that spatio-temporal image conditioning provides fine-grained control over human-object interactions while improving motion quality at a moderate cost in waypoint precision. Code and trained models will be publicly released.

## AI USE STATEMENT

We used generative AI tools to assist with writing code for training, evaluation, and figure generation. All AI-assisted code was manually reviewed and approved by the authors before use. AI tools were not used for research ideation or experimental design. We also used FLUX.2-dev to edit synthetic SceneImg renders into EditImg evaluation references (Appendix J). We take responsibility for the final content of this work.

## REFERENCES


German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, Ferran Rigual, Nicky Sijia He, Cristina Palmero, Sergio Escalera, Yuting Ye, and Robin Kips. From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

Black Forest Labs. FLUX.2. `https://bfl.ai/announcements/flux-2`, 2025.

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. ViHOI: Human-object interaction synthesis with visual priors. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2026.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3D motion generation for hand-object interaction. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. DiffH2O: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*. ACM, 2024.

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. ACM, 2024.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3d human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

Alexey Gavryushin, Alexandros Delitzas, Luc Van Gool, Marc Pollefeys, Kaichun Mo, and Xi Wang. SIGHT: Synthesizing image-text conditioned and geometry-guided 3D hand-object trajectories. *arXiv preprint arXiv:2503.22869*, 2025.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. *Computer Graphics Forum (CGF)*, 2023.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*. IEEE, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, 2020.

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. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*. Springer, 2022.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, 2023a.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023b.

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*. ACM, 2024a.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024b.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. NIFTY: Neural object interaction fields for guided human motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. Zerohsi: Zero-shot 4d human-scene interaction by video generation. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *Transactions on Graphics (TOG)*, 42(6), 2023a.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. CHOIS: Controllable human-object interaction synthesis. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*. PMLR, 2023b.

Lei Li and Angela Dai. GenZI: Zero-shot 3D human-scene interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023c.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4d dataset of mocap human object interactions. In *International Conference on Computer Vision (ICCV)*. IEEE, 2025.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. HIMO: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. HOI-Diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*. IEEE, 2025.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *International Conference on Computer Vision (ICCV)*. IEEE, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 2021.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. HOIDiNi: Human-object interaction through diffusion noise optimization. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*. IEEE, 2026.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2024.

Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. HOIAnimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soeren Pirk, and Michael J. Black. GRIP: Generating interaction poses using spatial cues and latent consistency. In *International Conference on 3D Vision (3DV)*. IEEE, 2024.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. MaskedMimic: Unified physics-based character control through masked motion inpainting. *Transactions on Graphics (TOG)*, 43(6), 2024.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2023.

Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. TLControl: Trajectory and language control for human motion synthesis. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Yin Wang, Ziyao Zhang, Zhiying Leng, Haitian Liu, Frederick W. B. Li, Mu Li, and Xiaohui Liang. Multimodal prior-augmented text-driven 3D human-object interaction generation. *Science China Information Sciences*, 2026.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, 2022.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2024.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-shot text to 3D dynamic human-object interaction. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, 2024a.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3d human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025a.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025b.

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 *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024b.

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.

Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OakInk2: A dataset of bimanual hands-object manipulation in complex task completion. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision (ECCV)*. Springer, 2024a.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023a.

Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. InteractAnything: Zero-shot human object interaction synthesis via LLM feedback and object affordance parsing. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M3: Capture multiple humans and objects interaction within contextual environment. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023b.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023c.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-grained spatio-temporal motion generation and editing. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, 2023d.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024c.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'M HOI: Inertia-aware monocular capture of 3D human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Yang Zhao, Yan Zhang, and Xubo Yang. IKMo: Image-keyframed motion generation with trajectory-pose conditioned motion diffusion model. *arXiv preprint arXiv:2505.21146*, 2025.

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pp. 145–162. Springer, 2024.

# APPENDIX

## A QUALITATIVE RESULTS

The supplementary video shows reference images, generated interaction frames, and full 120-frame motion sequences across FBM and BEHAVE. We include held-out object categories, different reference-frame positions, and cross-domain image inputs.

The examples include both successes and failures. Successful cases preserve the reference contact region, body-object layout, and object pose near the conditioning frame while maintaining plausible motion before and after the interaction. Failure cases show wrong contact side, hand-object penetration, temporal drift, and sensitivity to out-of-domain edited images. Fig. 4 shows additional qualitative comparisons on FullBodyManipulation with SceneImg references.

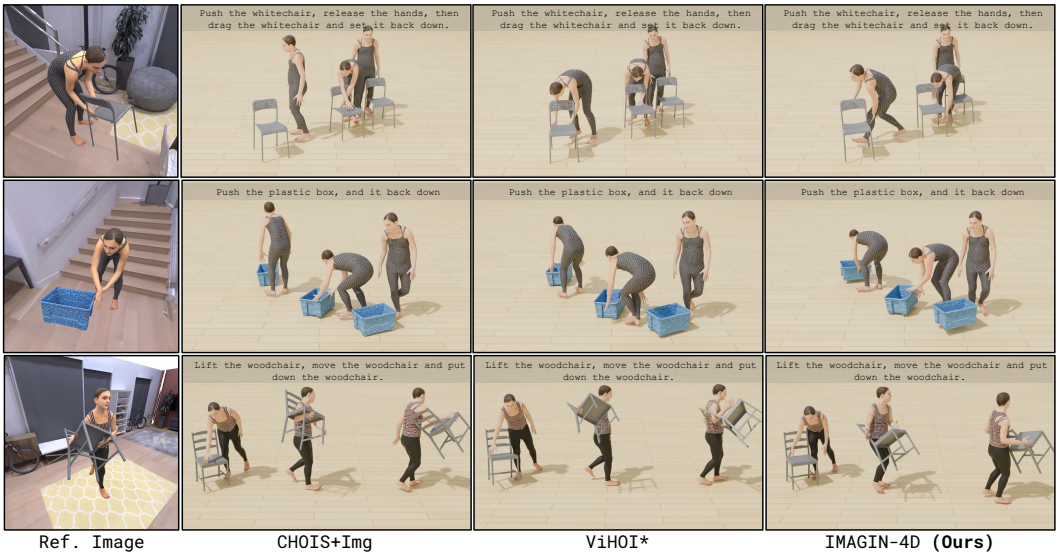



Figure 4: **Qualitative comparison on FullBodyManipulation (SceneImg) (Appendix A).** Each row shows the *SceneImg* reference and generated motion from CHOIS+Img, ViHOI⋆, and IMAGIN-4D under the same text prompt, object shape, waypoints, and reference image. Single-token image conditioning often misses contact, hand placement, object orientation, or body-object layout. IMAGIN-4D better matches the depicted interaction state.

## B PERCEPTUAL STUDIES

**Protocol.** We conducted two blind studies with 22 participants and randomized left-right or method placement. Each participant made one choice per example. The motion-plausibility study used ten FBM test examples and compared IMAGIN-4D, CHOIS+Img, and ViHOI⋆. The real-image study used six images as shown in Fig. 5 and compared IMAGIN-4D with ViHOI⋆, with text, object geometry, waypoints, and conditioning frame fixed across methods.

**Motion plausibility.** For each of ten examples, participants selected the most plausible motion. IMAGIN-4D received 73% of votes, ViHOI⋆ 18%, and CHOIS+Img 9%.

**Conditioning on real images.** Participants compared IMAGIN-4D and ViHOI⋆ on six real images as shown in Fig. 5: the box, trash can, white chair, floor lamp, and small table. They selected IMAGIN-4D as the closer image match in 78% of comparisons.

IMAGIN-4D preserves image-consistent object configurations, keeping the chair and table upright and the trash can low at the side. ViHOI⋆ flips the table and raises the trash can to chest height. Both methods tilt the lamp despite its upright pose in the image, indicating disagreement between the image condition and the supplied pulling trajectory.

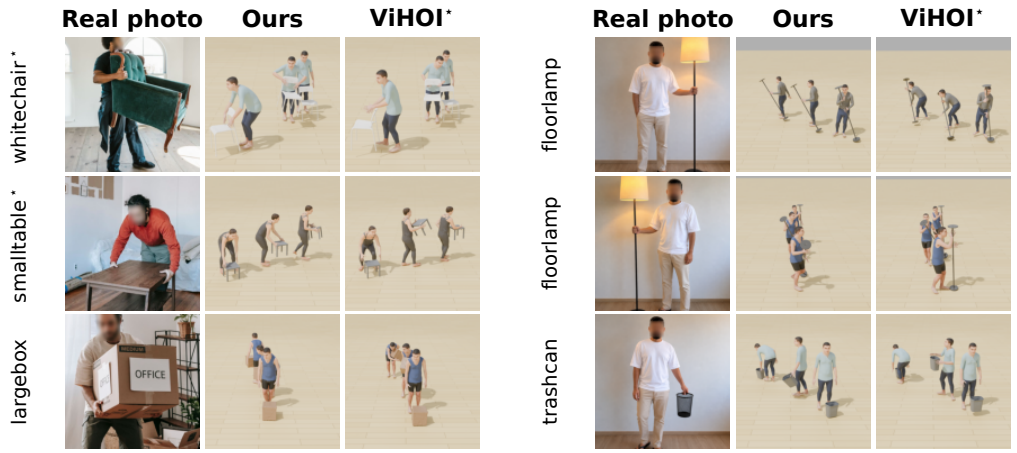



Figure 5: **Conditioning on real images (Appendix A).** Each row compares IMAGIN-4D and ViHOI* conditioned on the same image. Both methods receive text, object geometry, waypoints, and a conditioning frame from an FBM test window matched by action and object category. Renderings overlay frames near the start, middle, and end of each 120-frame sequence, offset along the motion direction. * marks object categories held out from training. The floor-lamp images differ only in which hand holds the lamp.

## C Image-Flip Probe

We test whether the model causally uses the reference image by horizontally flipping only the image at inference time while keeping the text, waypoints, object geometry, conditioning frame, and sampling seed fixed. The expected behavior is that grasp side, contact region, and approach direction change relative to the unflipped image, but the generated motion should not become a perfect mirror because the non-image conditions remain unchanged; see Fig. 6.

This probe isolates the effect of image conditioning. In contrast, training-time flip augmentation mirrors the image, body joints, object pose, rotations, contact labels, and waypoints jointly so that augmented samples remain physically consistent.

## D Sketch-to-Motion

We evaluate whether IMAGIN-4D can condition on line drawings instead of RGB references. Sketch images are produced from the same rendered reference frames using off-the-shelf line extractors. The motion data, reference-frame selection, camera, and scene placement remain unchanged, isolating the input image domain.

The architecture is unchanged; only the conditioning image changes. At 200k training steps, the sketch-conditioned model reaches $A_{\mathrm{W10}} = 9.66\,\mathrm{cm}$ and $A_{\mathrm{Any}} = 9.00\,\mathrm{cm}$. For context, the matched SceneImg RGB run in Tab. 5 reaches $7.65$ and $7.45\,\mathrm{cm}$, respectively, with FID $0.25$ versus $0.33$ for sketches. These results show that sketches provide usable control for controllable generation of motion; see Fig. 7.

## E Failure Cases

We observe three recurring failure modes. First, FBM contains imperfect hand poses and contact annotations, and the model inherits these artifacts from the training data. This is most visible near contact frames, where hands can look unnatural even when body-object layout and object motion are plausible. Second, the model can produce hand-object interpenetration, especially for small or thin objects. Image adherence and contact prediction improve interaction control, but they do not




Figure 6: **Flip-image consistency test (Appendix C).** We horizontally flip only the reference image at inference time while keeping the text prompt, object geometry, and waypoints fixed. The generated contact side and body-object layout change with the flipped image, showing that IMAGIN-4D uses visual evidence rather than ignoring the image condition. The motion is not an exact mirror because the unchanged non-image conditions must still be satisfied.

guarantee mesh-level physical validity. Third, a single reference image specifies one interaction snapshot, not the full approach or release motion. The generated sequence can therefore realize the depicted contact slightly before or after the annotated frame, reflected by gaps between $A_{\text{GT}}$, $A_{\text{W10}}$, and $A_{\text{Any}}$.

We retain the CHOIS backbone and motion representation to isolate the effect of image conditioning. The representation lacks articulated fingers, limiting grasp fidelity even when hand position and contact side match the reference. Foot sliding, object drift, and penetration remain (Appendix H); articulated hands and improved contact supervision may help.

## F Ablations

**Spatial decomposition (A).** The controlled reference for all ablations in Tab. 3 is the SceneImg→SceneImg configuration of Tab. 5; the full model (top row) is shown separately for reference. The four supervised Q-Former heads encode contact, human pose, object pose, and body-object layout from the same patch grid. Relative to the controlled reference, removing the layout head causes the largest increase in $A_{\text{Any}}$ (7.45 to 9.0 cm), while removing the human-pose head increases FID the most (0.25 to 0.40). Removing the contact head reduces $C_{F_1}$ from 0.677 to 0.60, with little change in FID or waypoint error. Removing the object-pose head changes $A_{\text{Any}}$ least among the four removals (7.45 to 7.7 cm) but increases waypoint error to 6.4 cm.

**Image classifier-free guidance (B).** Disabling image CFG ($s_{\text{img}}$=1) increases $A_{\text{Any}}$ from 7.45 to 8.46 cm and FID from 0.25 to 0.34, while waypoint error remains similar (5.69 vs. 5.66 cm). This supports image CFG for stronger reference adherence.

**Role-aware conditioning (C).** Removing the image AdaLN stream merges spatial image tokens into the base stream with text and time, while retaining waypoint modulation and frame-aware cross-attention. Relative to the controlled reference, $A_{\text{GT}}$ increases from 8.43 to 13.4 cm and $A_{\text{W10}}$ from

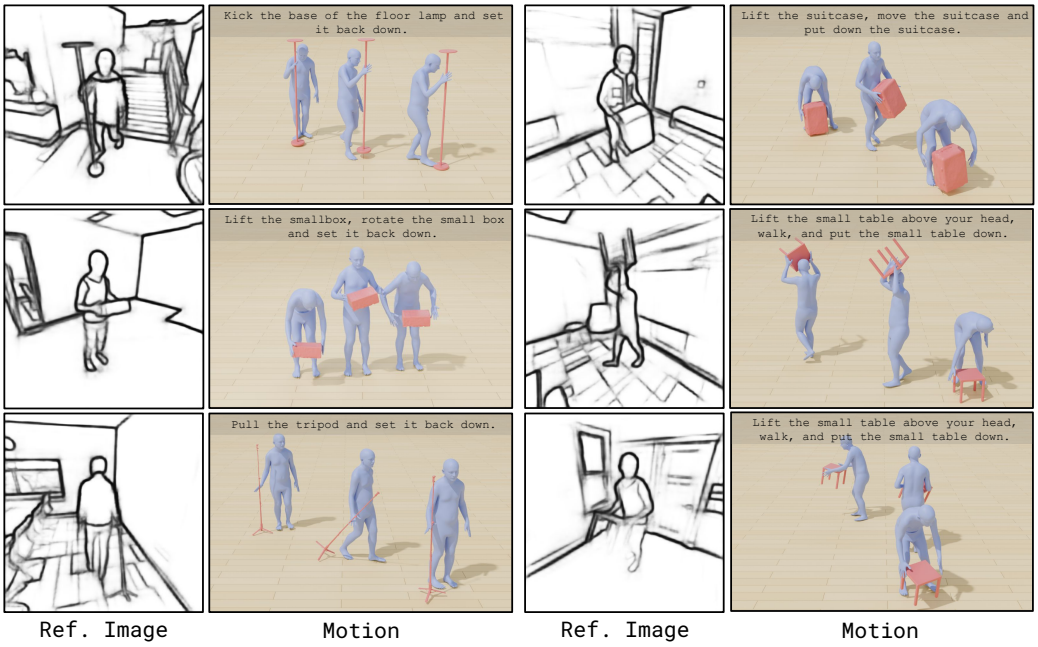



Figure 7: **Sketch-to-motion (Appendix D).** We replace the RGB reference image with a line drawing and retrain the model. Despite removing texture, color, and scene appearance, the model preserves the depicted interaction layout and generates a complete motion sequence. This shows that IMAGIN-4D can also support sketch-based conditioning, where users specify the desired contact and body-object arrangement with a simple drawing.

7.65 to 10.4 cm, while waypoint error decreases from 5.69 to 3.8 cm. This exposes a trade-off between temporally aligned image adherence and waypoint precision.

Routing the frame-aware tokens through per-layer AdaLN instead of final-layer cross-attention increases $A_{\text{GT}}$ from 8.43 to 9.2 cm and FID from 0.25 to 0.60, while $C_{F_1}$ changes from 0.677 to 0.66. These results favor final-layer cross-attention for frame-aware conditioning.

**Visual input modality (D).** Replacing RGB references with line drawings during training and evaluation increases $A_{\text{Any}}$ from 7.45 to 9.00 cm and FID from 0.25 to 0.33, with waypoint error changing from 5.69 to 6.01 cm. Sketches retain silhouette and object outlines but omit color and texture. See Appendix D for details.

### F.1 Qwen vs. DINO Patch Features

ViHOI uses Qwen-VL features, so we test whether replacing DINOv2 with Qwen2.5-VL improves our image-conditioning interface. This comparison is architecture-specific: our SFIE uses small role-specific Q-Former heads that read patch tokens for contact, human pose, object pose, and body-object layout. Such heads require patch features that remain spatially discriminative across image regions.

Tab. 4 shows that DINOv2 patch tokens have lower within-image and across-image cosine similarity than most tested Qwen2.5-VL layers, indicating greater spatial and image-specific variation for our role-specific Q-Former heads. This suggests that DINOv2 is better suited to our factorized patch-level conditioning interface, without implying that it is a stronger visual encoder in general.

## G Cross-Domain Image Transfer

Table 5 evaluates image-domain transfer while keeping the motion sequences and target interaction states fixed.

| Variant | Image Adh.,(cm)↓ | | | FID↓ | $C_{F_1}$ ↑ | WP |
|---|---|---|---|---|---|---|
| | GT | W10 | Any | | | (cm)↓ |
| **IMAGIN-4D** | **7.75** | **7.11** | **6.98** | **0.25** | **0.768** | 5.10 |
| *(A) Spatially Factorized Image Encoder (per-role).* | | | | | | |
| − contact ($\boldsymbol{\kappa}$) | 10.0 | 8.6 | 8.0 | 0.30 | 0.60 | 5.7 |
| − human-pose ($\boldsymbol{\rho}$) | 10.5 | 9.0 | 8.0 | 0.40 | 0.66 | 5.7 |
| − object-pose ($\boldsymbol{\xi}$) | 9.3 | 8.1 | 7.7 | 0.30 | 0.68 | 6.4 |
| − body-object layout ($\boldsymbol{\nu}$) | 11.0 | 9.5 | 9.0 | 0.34 | 0.64 | 5.7 |
| *(B) Image classifier-free guidance.* | | | | | | |
| $s_{\text{img}}$=1 (off) | 9.7 | 8.78 | 8.46 | 0.34 | 0.674 | 5.66 |
| *(C) Role-aware conditioning.* | | | | | | |
| − image AdaLN stream | 13.4 | 10.4 | 7.1 | 0.57 | 0.72 | 3.8 |
| frame-aware via AdaLN | 9.2 | 8.3 | 7.9 | 0.60 | 0.66 | 6.0 |
| *(D) Visual input modality.* | | | | | | |
| RGB → sketch | 10.9 | 9.66 | 9.00 | 0.33 | 0.654 | 6.01 |

Table 3: **Component and conditioning ablations on FullBodyManipulation (FBM)** (482-window full eval, step 200k). The reported model (top row) matches the "Ours (Spatial+Temporal)" row of Tab. 1 and is shown for reference. The ablations use the SceneImg→SceneImg configuration of Tab. 5 as the controlled reference; each row changes only the stated component, routing choice, or visual input while the remaining setup is fixed. The full model additionally uses logit-normal diffusion-step sampling and hand-object coupling (Appendix O). In (C), "− image AdaLN stream" merges image features into the base AdaLN alongside text and time (waypoint AdaLN stream and frame-aware cross-attention retained); "frame-aware via AdaLN" routes the per-frame frame-aware tokens through a per-layer AdaLN stream instead of the last-layer cross-attention (all other routes retained). In (D), both training and evaluation use line-drawing sketches in place of RGB references.

| Encoder | intra-cos↓ | cross-cos↓ | eff-dim↑ | rank99↑ |
|---|---|---|---|---|
| DINOv2 ViT-B/14 + reg | **0.27** | **0.63** | 8.34 | 123 |
| Qwen2.5-VL layer 3 | 0.44 | 0.93 | 5.27 | 170 |
| Qwen2.5-VL layer 6 | 0.59 | 0.97 | 3.01 | 160 |
| Qwen2.5-VL layer 12 | 0.56 | 0.97 | 3.22 | 171 |
| Qwen2.5-VL layer 18 | 0.58 | 0.98 | 2.92 | 182 |
| Qwen2.5-VL layer 24 | 0.59 | 0.98 | 2.73 | 182 |
| Qwen2.5-VL layer 28 | 0.30 | 0.92 | **8.54** | **194** |

Table 4: Patch-token redundancy on 100 FBM reference images. DINOv2 gives lower within-image and across-image cosine similarity, indicating more spatially diverse and image-discriminative patch features for our role-specific Q-Former heads.

MeshImg yields slightly lower adherence errors than SceneImg under matched train-test domains. However, models trained on MeshImg transfer poorly to SceneImg and EditImg: $A_{\text{GT}}$ increases by roughly 23–24 cm, together with a large increase in FID.

SceneImg transfers substantially better to EditImg. Under the matched cross-domain setup, $A_{\text{GT}}$ increases from $8.43$ to $12.68$ cm and FID from $0.25$ to $0.37$, while contact $F_1$ remains similar ($0.677$ to $0.683$). This motivates training on SceneImg, which introduces indoor backgrounds, textures, and lighting (Fig. 8).

EditImg contains FLUX.2-dev edits of synthetic SceneImg renders. Appendix A provides real-image comparisons and perceptual evaluation.

## H Physical Quality

| **Train → Test** | $A_{\text{GT}} \downarrow$ | $A_{\text{W10}} \downarrow$ | $A_{\text{Any}} \downarrow$ | FID↓ | $\mathbf{C_p} \uparrow$ | $\mathbf{C_r} \uparrow$ | $\mathbf{C_{F_1}} \uparrow$ | WP↓ |
|---|---|---|---|---|---|---|---|---|
| MeshImg → MeshImg | 7.85 | 7.04 | 6.87 | 0.26 | 0.825 | 0.653 | 0.695 | 5.64 |
| SceneImg → SceneImg | 8.43 | 7.65 | 7.45 | 0.28 | 0.823 | 0.633 | 0.677 | 5.69 |
| MeshImg → SceneImg | 31.66 | 23.40 | 15.53 | 2.68 | 0.822 | 0.576 | 0.635 | 6.90 |
| MeshImg → EditImg | 30.80 | 23.27 | 15.33 | 1.58 | 0.815 | 0.573 | 0.633 | 5.90 |
| SceneImg → EditImg | 12.68 | 10.70 | 9.50 | 0.37 | 0.819 | 0.645 | 0.683 | 5.77 |

Table 5: **Cross-domain image generalization (Appendix G).** We train and test IMAGIN-4D on FullBodyManipulation dataset across various image-conditioning domains while keeping motion data, text, object shape, and sparse waypoints fixed. *MeshImg* uses white-background body-object renders, *SceneImg* adds indoor scenes, textures, and lighting, and *EditImg* applies FLUX.2-dev image editing to SceneImg to increase photorealism. Models trained with MeshImg transfer poorly to SceneImg and EditImg, whereas the SceneImg-trained model transfers substantially better to EditImg, motivating SceneImg as the training domain for more photorealistic edited references.

We further evaluate physical artifacts not captured by our primary motion and image-adherence metrics, focusing on object drift, foot sliding, and human-object penetration.

**Off-contact object motion.** We measure object drift as the mean per-frame centroid displacement when neither hand is in contact. Adding the off-contact penalty reduces drift from $8.5$ to $6.1$ mm/frame, compared with $3.4$ mm/frame for ground-truth FBM motion.

**Foot sliding.** We test an $L_1$ penalty on the planar displacement of foot joints near the estimated floor. It reduces foot sliding from $0.41$ to $0.29$ cm (weight 10) and $0.34$ cm (weight 2), compared with $0.30$ cm for ground truth, but increases FID from $0.25$ to $0.53$ and $0.49$. We therefore omit it from the final model.

**Whole-body penetration.** Using the rest-frame object SDF, we obtain a mean zero-clipped penetration depth of $0.57$ cm for IMAGIN-4D and $0.61$ cm for ground-truth FBM motion.

**Hand-object penetration.** IMAGIN-4D achieves HandP (hand penetration metric) $0.53$, compared with $0.59$ for CHOIS and ViHOI$^\star$ (Tab. 1). Relative to ViHOI$^\star$, IMAGIN-4D reduces $A_{\text{Any}}$ from $13.33$ to $6.98$ cm without increasing HandP.

## I Image Prediction Errors and Timing Sensitivity

**Image-to-state accuracy.** On the full test set, reference-frame predictions achieve contact-flag $F_1 = 0.96$, body MPJPE $4.7$ cm, layout-orientation error $6.9°$, and object translation/rotation errors of $14.1$ cm/$21.5°$. We evaluate one prediction per test window at $t^\star$ (482 total), without additional alignment. Body MPJPE uses the 18 predicted joints; layout error measures angular error of joint-to-object directions; and rotation error is the geodesic angle between predicted and ground-truth rotations. These metrics evaluate the image heads rather than generated-motion quality.

**Timing sensitivity.** We shift the conditioning frame predicted by the model by $k \in \{\pm1, \pm3, \pm5, \pm15\}$ frames, while keeping the reference image fixed, and average results over $+k$ and $-k$. $A_{\text{Any}}$ remains 7.4–7.6 cm across all offsets, with FID $0.25$–$0.37$ and contact $F_1$ between $0.67$ and $0.69$, indicating that the depicted state is still realized within the sequence. $A_{\text{GT}}$ at the original $t^\star$ is approximately unchanged within $\pm3$ frames, but increases by $1.0$ cm at $\pm5$ and $5.0$ cm at $\pm15$. Thus, within the tested offsets, conditioning-frame errors primarily shift when the depicted interaction occurs rather than preventing it from occurring. This experiment measures sensitivity to imposed timing errors.

## J Data Generation Details

FullBodyManipulation (FBM) (Li et al., 2023a) provides human-object motion without images, so we render reference images from its ground-truth motion sequences. BEHAVE (Bhatnagar et al., 2022) provides RGB frames registered to the motion, which we use directly (Appendix M). The rendering details below apply to FBM.




Figure 8: **Conditioning image domains (Appendix J).** We render conditioning images from each ground-truth sequence from the FullBodyManipulation dataset and use the contact-centered frame for evaluation. *MeshImg* is a clean body-object render, *SceneImg* adds Replica scenes, body textures, and posed objects, and *EditImg* applies image editing for more photorealistic references. SceneImg is used for evaluation, while MeshImg and EditImg analyze image-domain transfer.

**Reference-frame selection.** For each 120-frame window, we render frames from a uniform temporal grid and one contact-centered frame. The contact-centered frame is the temporal centroid of left- and right-hand contact flags:

$$t^\star = \text{round}\left[\frac{\sum_t t(c_t^{\text{L}} + c_t^{\text{R}})}{\sum_t (c_t^{\text{L}} + c_t^{\text{R}})}\right]. \tag{11}$$

If no hand contact occurs, we use the window center.

**Rendering domains.** We use three image domains. *MeshImg* renders the posed body and object on a white background. *SceneImg* places the same human-object pair in Replica indoor scenes (Straub et al., 2019) with BEDLAM body textures (Black et al., 2023), object materials, and Blender Cycles shading. *EditImg* applies a directive-prompt FLUX.2-dev (Black Forest Labs, 2025) image-editing pass to SceneImg, with prompts designed to preserve body pose, hand placement, object pose, and contact while changing appearance; see Fig. 8.

SceneImg is used for the main FBM training and evaluation and for the architecture ablations of Tab. 3 because it contains scene context, lighting, shadows, clothing texture, and object appearance. MeshImg is used for the cross-domain analysis. EditImg is used only at test time to evaluate robustness to more photorealistic references.

**Scene placement and camera.** For SceneImg, we sample valid navigable scene locations, place the body-object pair on the local floor, and select a camera that maximizes human-object visibility while avoiding severe scene occlusion. Images are rendered at $512 \times 512$ using Blender Cycles. We render each conditioning image once and reuse it across methods, so all methods receive identical visual inputs.

**Rendering and editing settings.** SceneImg is rendered with Blender Cycles at 64 samples per pixel with denoising. For each 120-frame window, we render eight uniformly spaced frames and the contact-centered frame $t^\star$. Cameras use a $60°$ field of view, widened to at most $90°$ when needed, and are placed 1.5–3.5 m from the subject at 1.6 m height. We select an unoccluded viewpoint maximizing human and object joint visibility. MeshImg uses PyRender with a frontal camera, $60°$ field of view, white background, and three directional lights.

EditImg applies FLUX.2-dev to the SceneImg render at $t^\star$ using 28 inference steps, guidance scale 4.0, seed 0, and 1024-pixel maximum resolution. All 482 edits are retained without manual fil-




Figure 9: **Why EditImg is derived from SceneImg.** Columns 1–2 show direct MeshImg→EditImg results. Columns 3–5 show the same interaction as MeshImg, SceneImg, and SceneImg-derived EditImg. Direct MeshImg editing can preserve the interaction in some cases, but can also change human pose, object pose, contact, or body-object layout. Since image adherence assumes that the reference preserves the target interaction geometry, we use the more reliable SceneImg→EditImg protocol for evaluation.

tering. The prompt is: *"Photorealistic photograph of a person interacting with a [object]. The person is wearing [outfit]. Realistic skin texture, fabric weave, and material rendering. Sharp focus, 50mm lens, magazine quality. Preserve exactly as in the input image: the body pose, hand placement, the [object] position and orientation, the contact between hands and [object], the surrounding room geometry and layout, the floor and wall surfaces, and the direction and color of the existing scene lighting. Only upgrade the realism of materials, textures, skin, fabric, and the person's clothing/shoes/hair as described above."* Here, [object] is the object category and [outfit] is selected deterministically from ten predefined descriptions.

**Main-image protocol.** Unless otherwise stated, the FBM test image is SceneImg sampled at $t^\star$. Main FBM comparisons and architecture ablations are trained and evaluated on SceneImg. MeshImg and EditImg are reserved for cross-domain probes so that the domain-shift experiment changes the image distribution rather than the motion sequence.

**Why EditImg is derived from SceneImg.** We apply FLUX.2-dev editing to SceneImg rather than MeshImg to keep the evaluation geometry controlled. MeshImg editing sometimes produces plausible photographic images, but the edit can also change human pose, object pose, hand-object contact, or body-object layout. This breaks the image-adherence protocol, which assumes that the reference image preserves the target interaction geometry from the ground-truth motion. SceneImg already contains indoor context, shadows, body texture, and object appearance, so editing mainly changes visual realism while better preserving the interaction state. We therefore use SceneImg→EditImg as the photorealistic transfer setting. Fig. 9 shows two failure cases and one partially successful MeshImg edit.

## K METRICS

We report five metric groups. Distances are reported in centimeters unless stated otherwise.

**Image adherence.** Image adherence measures whether the generated sequence realizes the interaction state depicted in the reference image. At the annotated conditioning frame $t^\star$, the reference image defines a target human-object configuration. We report three variants: $A_{\text{GT}}$, evaluated exactly at $t^\star$; $A_{\text{W10}}$, the best score within a $\pm 10$-frame window around $t^\star$; and $A_{\text{Any}}$, the best score over the full 120-frame sequence. Lower is better. For details, see Appendix K.1.

**Motion quality and text alignment.** FID and R-Precision are computed using the frozen CHOIS text-to-motion evaluator (Li et al., 2024). This keeps the protocol directly comparable to prior text-and-waypoint HOI generation methods.

**Contact.** Contact is measured by precision, recall, and $F_1$ between generated left/right hand contact flags and the ground-truth hand contact annotations of the respective dataset (FBM or BEHAVE), aggregated over all frames and both hands of all test windows.

**Interaction artifacts.** We report foot sliding and hand-object penetration. Foot sliding measures the motion of foot joints while they are predicted to be in floor contact. Hand-object penetration measures mesh-level hand penetration into the object using the rest-frame object SDF. For BEHAVE categories without rest-frame SDFs, hand penetration is not reported.

**Waypoint following.** Waypoint error measures the planar distance between the generated object trajectory and the prescribed sparse waypoints at waypoint timestamps. Since all controllable baselines receive waypoints, this metric measures whether image conditioning degrades trajectory control.

### K.1 Image Adherence Metric

Let the generated sequence be $\{(\mathbf{j}_t, \mathbf{o}_t)\}_{t=1}^{T}$, where $\mathbf{j}_t \in \mathbb{R}^{24\times 3}$ are body-joint positions and $\mathbf{o}_t \in \mathbb{R}^3$ is the object centroid. The reference image corresponds to the ground-truth interaction state $(\mathbf{j}^\star, \mathbf{o}^\star)$ at frame $t^\star$. Predicted and ground-truth poses are mapped to the same canonical interaction frame before evaluation.

For each generated frame $t$, we compute

$$d(t) = 0.8\, d_{\text{hum}}(t) + 0.2\, \|\mathbf{o}_t - \mathbf{o}^\star\|_2, \tag{12}$$

where

$$d_{\text{hum}}(t) = \sum_{k=1}^{24} w_k \|\mathbf{j}_{t,k} - \mathbf{j}^\star_k\|_2, \qquad w_k = \frac{\exp(-\|\mathbf{j}^\star_k - \mathbf{o}^\star\|_2/\sigma)}{\sum_{r=1}^{24} \exp(-\|\mathbf{j}^\star_r - \mathbf{o}^\star\|_2/\sigma)}, \tag{13}$$

with $\sigma = 0.3$m. The weights emphasize joints close to the object in the reference interaction state, making the metric sensitive to contact-relevant body parts while still accounting for full-body configuration. We report

$$A_{\text{GT}} = d(t^\star), \tag{14}$$

$$A_{\text{W10}} = \min_{|t-t^\star|\le 10} d(t), \tag{15}$$

$$A_{\text{Any}} = \min_{1\le t\le T} d(t). \tag{16}$$

$A_{\text{GT}}$ measures exact temporal alignment. $A_{\text{W10}}$ allows small timing shifts while still requiring the interaction to occur near the intended frame. $A_{\text{Any}}$ diagnoses whether the desired interaction is realized somewhere in the sequence even when temporal placement is imperfect.

## L Baseline Construction

**Image-free baselines.** We report the published FBM numbers from CHOIS (Li et al., 2024). These include CHOIS, MDM (Tevet et al., 2023), and InterDiff (Xu et al., 2023), with MDM and InterDiff retrained for the text-and-waypoint HOI generation task by CHOIS.

**Pooled-image CHOIS baselines.** To test whether image conditioning helps under a simple interface, we retrain CHOIS with one pooled visual token. We use three encoders: CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), and Qwen-VL (Cai et al., 2026). These rows isolate the effect of encoder choice when the conditioning interface is fixed to a single pooled token.

**ViHOI re-implementation.** The closest image-conditioned competitor is ViHOI (Cai et al., 2026), which compresses reference images into a single visual token using a Qwen-VL backbone and Q-Former adapter. We implement ViHOI following its published formulation and adapt its three-image input to our single-image setting for a matched comparison. We mark this adapted row as ViHOI$^\star$.

**Controlled encoder-interface comparison.** The Qwen-VL pooled baseline and ViHOI$^\star$ hold the visual backbone family fixed while changing the image interface from pooling to a Q-Former token. The CLIP and DINOv2 pooled baselines hold the single-token interface fixed while changing the encoder. These controlled rows separate encoder strength from the representational bottleneck of single-token conditioning.

## M BEHAVE PROTOCOL

BEHAVE (Bhatnagar et al., 2022) is used as a second-benchmark replication with a different motion distribution from FBM. We train and evaluate a separate model on BEHAVE using the dataset's real RGB frames as reference images: for each 120-frame window, the RGB frame captured at the contact-centered reference frame $t^\star$ is the conditioning image, so no rendering is involved. Ground-truth body and object poses come from the dataset registrations (Bhatnagar et al., 2022; Xu et al., 2025a). Unlike the FBM experiments, this benchmark therefore conditions on real images.

For categories without rest-frame object SDFs, we do not report hand-object penetration. All other metrics follow Appendix K.

## N ARCHITECTURE DETAILS

**Spatially factorized image encoder.** The image encoder extracts $16 \times 16$ DINOv2 patch tokens from the reference image and projects them to the denoiser width. Four Q-Former heads read the same patch grid with separate learnable queries: contact, human pose, object pose, and body-object layout. Their outputs are concatenated and projected into a spatial image representation.

**Frame-aware visual tokens.** A separate frame-aware Q-Former re-queries the same DINOv2 patch grid for each frame. Its queries are conditioned on the frame index and text embedding, producing a frame-dependent visual memory. This memory is used by the final decoder layer through cross-attention.

**Reference-frame localizer.** A two-layer MLP predicts a distribution over the 120 motion frames from the spatial image representation. During training, it is supervised with a Gaussian-smoothed label centered at the ground-truth reference frame. The predicted frame is used for temporal image gating during both training and inference.

**Denoiser.** Each transformer decoder layer uses AdaLN-Zero modulation. Text, waypoint, and spatial-image conditions are routed through separate AdaLN streams. Frame-aware visual tokens are routed through late cross-attention instead of global modulation.

## O TRAINING AND OPTIMIZATION DETAILS

**Backbone.** The denoiser is a 4-layer transformer with width $512$ and $4$ attention heads, trained as an $x_0$-prediction DDPM on 120-frame motion windows. Each frame contains object translation, object rotation, body joint positions, body joint rotations, and semantic contact channels. Object geometry is represented with a 1024-point BPS encoding (Prokudin et al., 2019).

**Conditioning.** The model uses three AdaLN streams: a base stream for diffusion step and text conditioning, a waypoint stream for sparse trajectory constraints, and an image stream for window-gated spatial image tokens. The waypoint and image streams are zero-initialized so training starts close to the text-and-waypoint model. Frame-aware visual tokens are injected through a cross-attention block in the final decoder layer.

**Image heads.** The image branch uses frozen DINOv2 ViT-B/14 patch features (Oquab et al., 2024). Separate Q-Former heads (Li et al., 2023b) predict contact, human pose, object pose, and

body-object layout at the reference frame. These supervised heads form spatial image tokens used by the denoiser and reference-frame localizer.

**Losses.** The training objective combines the denoising loss, SFIE supervision, reference-frame localization loss, image-consistency loss at the reference frame, forward-kinematics loss, object-point loss, and foot-contact loss. The denoising loss is upweighted near the predicted reference frame so that image conditioning receives stronger supervision where the reference image is most informative.

The model replaces uniform sampling of the diffusion step with a logit-normal distribution, $n = \lfloor \sigma(z)\, N \rfloor$ with $z \sim \mathcal{N}(0,1)$ and $N{=}1000$ diffusion steps, concentrating samples at intermediate noise levels ($n$ is the diffusion step of Sec. 3.1).

Symbols below match Sec. 3. The SFIE supervision (Eq. (2)) uses $\lambda_{\mathrm{con,pos}}{=}\lambda_{\mathrm{con,flag}}{=}2.0$, $\lambda_{\mathrm{hum}}{=}1.5$, $\lambda_{\mathrm{obj}}{=}1.0$, and $\lambda_{\mathrm{spa}}{=}2.5$, with the contact-position MSE and contact-flag BCE sharing a single scalar weight applied to their sum. The reference-frame localizer (Eq. (4)) uses $\lambda_{\mathrm{fp}}{=}0.1$ with label-Gaussian $\sigma_q{=}3$ frames. The window-gated image conditioning (Eq. (5)) uses a Gaussian temporal window $w_{\sigma_g}$ with $\sigma_g{=}5$ frames, and the denoising-loss frame emphasis (Eq. (8)) uses $\lambda_{\mathrm{cf}}{=}5.0$ with spread $\sigma_{\mathrm{cf}}{=}5$ frames. The auxiliary losses (Eq. (9)) use $\lambda_{\mathrm{img}}{=}2.0$, $\lambda_{\mathrm{FK}}{=}0.5$, $\lambda_{\mathrm{objpts}}{=}1.0$, and $\lambda_{\mathrm{feet}}{=}1.0$. Image classifier-free guidance uses drop probability $p_{\mathrm{drop}}{=}0.1$.

**Optimization.** We train with AdamW in bf16 at learning rate $10^{-4}$ on four GPUs with per-GPU batch size 32 (effective batch 128). The schedule warms up over a single step and is then held flat by an adaptive scheduler that halves the learning rate on detected NaN and recovers on subsequent stable windows. Diffusion uses 1000 DDPM steps with a cosine noise schedule and predicts $\mathbf{x}_0$. We track an exponential moving average of the denoiser with decay 0.995 updated every 10 steps; all reported numbers use the EMA weights. We apply left-right flip augmentation with probability 0.5 jointly to image features, body joints, object pose, rotations, contact labels, and waypoint constraints. Evaluation uses no flip augmentation.

## P Alternative Conditioning Interfaces

**Structured visual control.** ControlNet (Zhang et al., 2023b), IKMo (Zhao et al., 2025), and SIGHT (Gavryushin et al., 2025) use spatial, keyframe, or image-based controls. IMAGIN-4D instead controls whole-body HOI through supervised interaction-state tokens and frame-aware visual retrieval.

**Detailed text.** Text can describe grasp, pose, contact, and layout, but an image specifies these spatial cues jointly. In Fig. 6, mirroring only the reference image changes grasp side and contact while all other conditions remain fixed.

**Reconstruction-based control.** Another option is to reconstruct a 3D interaction state and use it as a keyframe. This additionally requires estimating the state and its timing. IMAGIN-4D learns both within its conditioning pipeline; Appendix I analyzes their errors.