IMAGIN-4D: Image-Guided Controllable Interaction Generation

ANONYMOUS AUTHORS

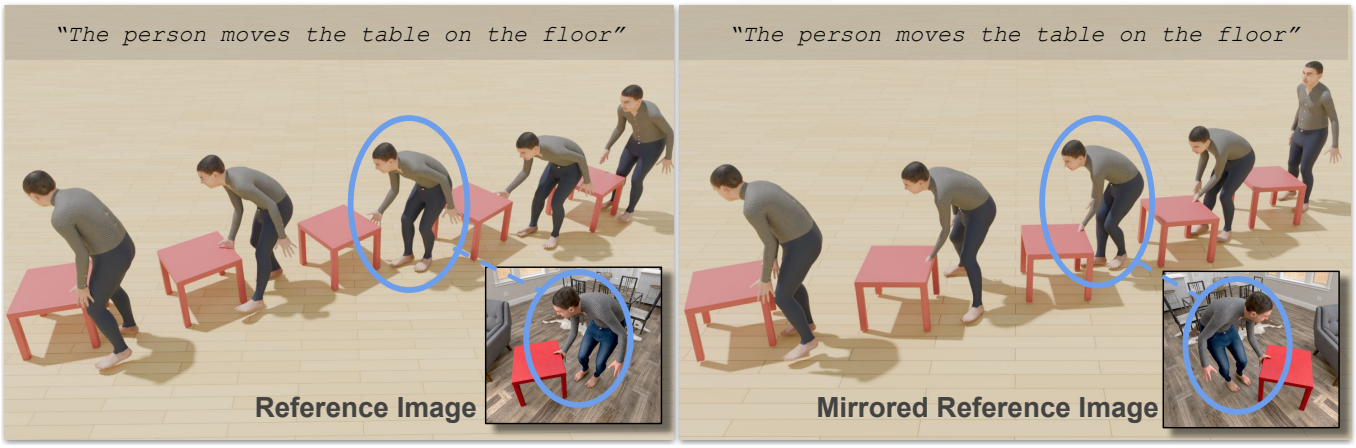

Fig. 1. **Image-conditioned 4D HOI generation.** Given a text prompt, object geometry, object waypoints, and a reference image, IMAGIN-4D synthesizes a 4D human-object interaction sequence. Text and waypoints specify the action and object trajectory, but leave fine-grained interaction details such as pose, contact, and layout ambiguous. We resolve this ambiguity with a reference image that specifies the interaction snapshot. To test whether IMAGIN-4D follows this visual evidence, we keep the text prompt, object geometry, and waypoints fixed, and mirror only the reference image. IMAGIN-4D generates different motions that satisfy the corresponding snapshot: body pose, object pose, contact, and body-object layout change consistently with the mirrored reference. This is achieved through spatio-temporal image conditioning, which separates spatial cues for the depicted interaction state from frame-aware cues for the surrounding motion. Unlike single-token image conditioning, this preserves fine-grained visual evidence while generating the HOI sequence.

Generating human-object interactions (HOI) is central to character animation, robotics, AR/VR, and embodied AI. Controlled HOI generation has several practical implications and recent methods synthesize motion from language, object geometry, and sparse geometric constraints, providing control over action semantics and object trajectories. However, these signals underspecify the interaction: the same prompt and trajectory can correspond to different grasps, approaching directions, body poses, object poses, contacts, and body-object layouts. We address this ambiguity by using a reference image as a visual specification of the desired interaction snapshot. However, a single global image representation conflates distinct interaction cues and conditions all motion frames on the same visual evidence. Thus, we introduce IMAGIN-4D, a diffusion-based HOI generator that decomposes the image conditioning signal *spatio-temporally*. For *spatial conditioning*, IMAGIN-4D extracts supervised interaction-state tokens that capture body pose, object pose, and body-object contact and spatial relationships at the depicted frame. For *temporal conditioning*, it computes frame-aware tokens by querying image patches for each generated frame, allowing different parts of the sequence to attend to different visual cues from the same image. To balance image, text, and waypoint cues, IMAGIN-4D uses role-aware conditioning: text, waypoints, and interaction-state tokens are conditioned through separate AdaLN streams, while frame-aware visual tokens are cross-attended with the learnable motion tokens. Since existing HOI motion datasets lack paired images, we build a synthetic motion-to-image rendering pipeline from FullBodyManipulation (FBM) dataset sequences and introduce an image-adherence metric that evaluates whether generated motions match the reference snapshot. 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. Code and models will be released at https://imagin4d.github.io.

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 process.

Recent diffusion-based methods [Li et al. 2024a; Xu et al. 2023] 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 been explored by ViHOI [Cai et al. 2026] concurrently to us. Specifically, 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 the 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. 2023a] 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 and 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, existing HOI datasets contain MoCap data without paired images. We thus render images from FullBody-Manipulation (FBM) sequences [Li et al. 2023c] 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 (concurrent) 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, and adding frame-aware tokens further improves motion quality and 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 [Bhatnagar et al. 2022; Huang et al. 2022; Jiang et al. 2023b; Kim et al. 2025; Li et al. 2023c; Lu et al. 2025; Lv et al. 2024; Taheri et al. 2020; Zhang et al. 2024c; Zhao et al. 2024] capture paired human-object motion, while hand-object benchmarks [Fan et al. 2023; Liu et al. 2022, 2024; Zhan et al. 2024] focus on manipulation and scene-aware corpora. Some datasets [Hassan et al. 2019; Jiang et al. 2024b; Wang et al. 2022] 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; they do not provide motion-frame-aligned images that specify a target interaction state. We render conditioning images from FullBodyManipulation [Li et al. 2023c] 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 [Chen et al. 2023; Ho et al. 2020; Tevet et al. 2023; Zhang et al. 2024a] and tokenized generators [Guo et al. 2024; Jiang et al. 2023a; Zhang et al. 2023c,a,b] synthesize human motion from language. Additional control is introduced through guidance [Karunratanakul et al. 2023], joint or keyframe constraints [Cohan et al. 2024; Xie et al. 2024], sparse trackers [Barquero et al. 2025], trajectory conditioning [Karunratanakul et al. 2024; Shafir et al. 2024; Wan 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 [Christen et al. 2024; Taheri et al. 2022, 2024; Zhang et al. 2024b; Zhou et al. 2022], whole-body object interaction [Ghosh et al. 2023; Li et al. 2023c; Xu et al. 2023; Zhang et al. 2022], and language- or trajectory-conditioned 4D HOI generation [Cha et al. 2024; Diller and Dai 2024; Li et al. 2024a; Peng et al. 2025; Ron et al. 2025; Song et al. 2024]. CHOIS [Li et al. 2024a] 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 [Jiang et al. 2024a,b; Kulkarni et al. 2024; Li et al. 2024b; Li and Dai 2024; Wang et al. 2023, 2024; Xu et al. 2025b, 2024a; Yi et al. 2024; Yuan et al. 2023; 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 [Hu et al. 2024; Huang et al. 2023; Li et al. 2023b; Mou et al. 2024; Xu et al. 2024b; Ye et al. 2023; 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.

3 Method

IMAGIN-4D generates a human-object motion from a text prompt $\mathbf{Y}$, object shape 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 and 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. 2024a], 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 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 $\mathbf{c} = \{\mathbf{Y}, O_{BPS}, \mathbf{x}_{\text{cond}}\}$ 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. 2023a] reads $\mathbf{P}$ with learned queries $\mathbf{Q}_h$:

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

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 $\tilde{\boldsymbol{\zeta}}$, 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

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

where λ 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 $\mathcal{L}^{\text{cos}}_{\text{spa}}$ 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.

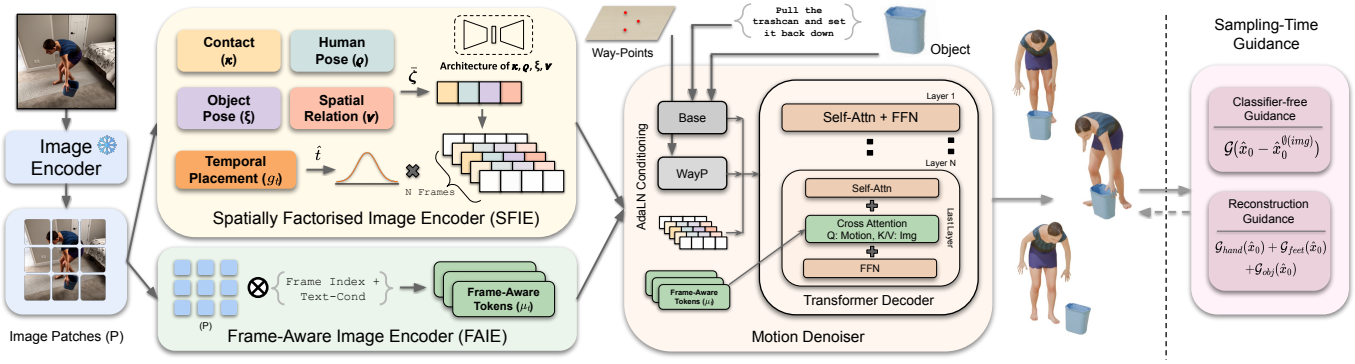

Fig. 2. **IMAGIN-4D overview (Sec. 3).** Given a reference image $\mathcal{I}$, text prompt y, object geometry O, and sparse waypoints $\mathcal{W}$, IMAGIN-4D generates a 4D human-object motion sequence. A frozen image encoder extracts patch tokens $\mathbf{P}$ from $\mathcal{I}$. The Spatially Factorized 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

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 π_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_{σ_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 earlier or later frames rely more on object identity, coarse layout, or approach direction. We therefore re-query image patches separately 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 different roles. Text provides sequence-level action semantics. Sparse waypoint constraints provide trajectory control. Window-gated spatial image evidence provides local constraints 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 according to its role.

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

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

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

No existing dataset provides human-object motion *paired* with images, which is required for image-guided motion generation. The most related datasts are FullBodyManipulation (FBM) [Li et al. 2023c] and BEHAVE [Bhatnagar et al. 2022], which only provide motions. Thus, we render images using meshes of these datasets (FBM has SMPL-X [Pavlakos et al. 2019] and BEHAVE has SMPL-H [Romero et al. 2017] bodies), 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. Since BEDLAM textures are baked to the SMPL-X UV layout, SceneImg is used only for OMOMO, which provides SMPL-X bodies [Pavlakos et al. 2019]. BEHAVE provides SMPL-H bodies [Romero et al. 2017], so we use MeshImg for BEHAVE training and evaluation. For FBM, SceneImg is used for the main experiments and MeshImg for ablations and cross-domain analysis.

For each 120-frame window, we render a uniform temporal grid and one contact-centered frame $t^\star$. We define $t^\star$ as the contact-weighted temporal centroid of the left- and right-hand contact flags; see Sup. Mat.or details. At training time, one rendered image is sampled per sequence per epoch; at evaluation time, 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:

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

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. 2024a], we add a forward-kinematics loss $\mathcal{L}_{\text{FK}}$, an object-point loss $\mathcal{L}_{\text{objpts}}$, and a foot-contact loss $\mathcal{L}_{\text{feet}}$. We also use an image-consistency loss $\mathcal{L}_{\text{img}}$ at the ground-truth reference frame $t^\star$. This loss compares interaction quantities decoded from the generated frame with the corresponding reference-image targets, including human pose, object pose, and body-object contact and layout. Its weight increases at lower-noise diffusion steps, where $\hat{\mathbf{x}}_0$ is reliable enough for frame-level geometric supervision. The full objective is:

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

All loss weights, optimizer settings, learning-rate schedule, data augmentation, and remaining hyperparameters are reported in Sup. Mat.

Image classifier-free guidance. We use classifier-free guidance on image condition [Ho and Salimans 2022]. During training, with probability p_{drop}, we replace all image inputs by their null form: spatial tokens ζ 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

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

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

Mesh-level hand and foot guidance. Following CHOIS [Li et al. 2024a], 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, which penalizes hand-object penetration and attracts the hand to the object at frames predicted to be in contact, and a foot term, which penalizes support-foot deviation from the floor. We restrict this guidance to the final denoising steps because 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. 2024a] on FullBodyManipulation (FBM) [Li et al. 2023c] and BEHAVE [Bhatnagar et al. 2022; Xu et al. 2025a]. FBM contains human-object MoCap sequences with nine training and four held-out object categories; we use the same train/test split, and text prompts as CHOIS. BEHAVE provides RGB-D human-object sequences, which we convert to the same 120-frame windowing convention. Since these datasets do not provide reference images for our image-conditioned generation protocol, we use the rendered conditioning images described in Sec. 3.4. For the main FBM comparison and ablations, both training and testing use *SceneImg*. *MeshImg* and *EditImg* are used for cross-domain image analysis; see Fig. 5. Applying FLUX on MeshImg often changes pose or object layout, so we generate EditImg from SceneImg; see Sup. Mat. BEHAVE is trained and evaluated with *MeshImg* because BEHAVE uses SMPL-H bodies, which are incompatible with the SMPL-X UV layout used by BEDLAM textures. Rendering details and reference-frame selection are given in Sec. 3.4 and Sup. Mat.

Metrics. We report image adherence, motion quality, text alignment, contact, interaction artifacts, and waypoint following. Image adherence is a new metric that we define to measure whether the generated sequence realizes the reference interaction state; we report A_{GT} at $t^\star$, A_{W10} within a ±10-frame window, and A_{Any} over the full sequence, all in cm. Motion quality and text alignment are measured by FID and R-Precision using the frozen CHOIS evaluator. Contact is measured by hand-object contact precision, recall, and F_1. Interaction artifacts are measured by foot sliding and hand-object penetration. Waypoint error measures object-trajectory error at sparse waypoint frames. For formal definitions, see Sup. Mat.

Baselines and implementation. We compare against image-free HOI baselines: CHOIS [Li et al. 2024a], InterDiff [Xu et al. 2023], and MDM [Tevet et al. 2023]. For image-conditioned baselines, we retrain CHOIS with a single pooled image token from CLIP [Radford et al. 2021], DINOv2 [Oquab et al. 2024], and Qwen-VL [Cai et al.

Method	Img Enc	Image Adherence (cm) ↓			FID ↓	R-Precision ↑			Contact ↑			Interaction ↓		WPErr ↓
		A_{GT}	A_{W10}	A_{Any}		R@1	R@2	R@3	C_{prec}	C_{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. 2024a]	–	–	–	–	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★ [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	0.565	0.714	0.805	0.669	0.697	0.39	0.62	6.37
Ours (Spatial)	DINOv2	10.11	8.91	7.95	0.30	0.352	0.569	0.716	0.811	0.621	0.663	0.44	0.56	6.28
Ours (Spatial+Temporal)	DINOv2	**8.43**	**7.65**	**7.45**	**0.28**	**0.365**	**0.582**	**0.726**	**0.823**	**0.633**	**0.677**	0.43	**0.53**	5.69

Table 1. **Evaluation on the FullBodyManipulation dataset (Sec. 4.1).** All methods are evaluated on the FullBodyManipulation [Li et al. 2023c] 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_{GT}), within a ±10-frame window (A_{W10}), and over the full sequence (A_{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★ denotes our single-image adaptation of ViHOI [Cai et al. 2026].

Method	A_{GT}↓	A_{W10}↓	A_{Any}↓	FID↓	C_p↑	C_r↑	C_{F_1}↑	WP↓
CHOIS [Li et al. 2024a]	–	–	–	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 ±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★.

Train → Test	A_{GT}↓	A_{W10}↓	A_{Any}↓	FID↓	C_p↑	C_r↑	C_{F_1}↑	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 3. **Cross-domain image generalization (Sec. 4.2).** 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 methods with SceneImg transfer much better to EditImg, supporting SceneImg for robust conditioning on photorealistic references.

2026]. We also re-implement ViHOI [Cai et al. 2026] (since code is not yet public), adapting its three-image input to our single-image setting and marking this row as ViHOI★; see Sup. Mat. for details.

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.

Baselines using a single global image token, namely CHOIS+Img and ViHOI★, improve some global motion metrics such as FID but provide limited image adherence. For CHOIS+Img and ViHOI★, A_{Any} remains between 13.08 and 13.33 cm across CLIP, DINOv2, and Qwen2.5-VL image encoders. ViHOI★ improves FID relative to CHOIS+Img with the same Qwen2.5-VL encoder, but its adherence remains comparable. This indicates a representation bottleneck: a single image token compresses spatial evidence that is needed to recover body pose, object pose, and body-object contact and layout.

Our spatial encoder addresses this bottleneck with supervised interaction tokens localized at the conditioning frame. **Ours (Spatial)** reduces A_{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. Waypoint error increases relative to CHOIS because image and waypoint conditions compete for control; Sup. Mat. analyzes this tradeoff.

Figure 3 shows the qualitative failure mode under photorealistic EditImg conditioning. 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 MeshImg, testing whether spatial image factorization transfers to a different motion distribution. 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 the BEHAVE rendering protocol.

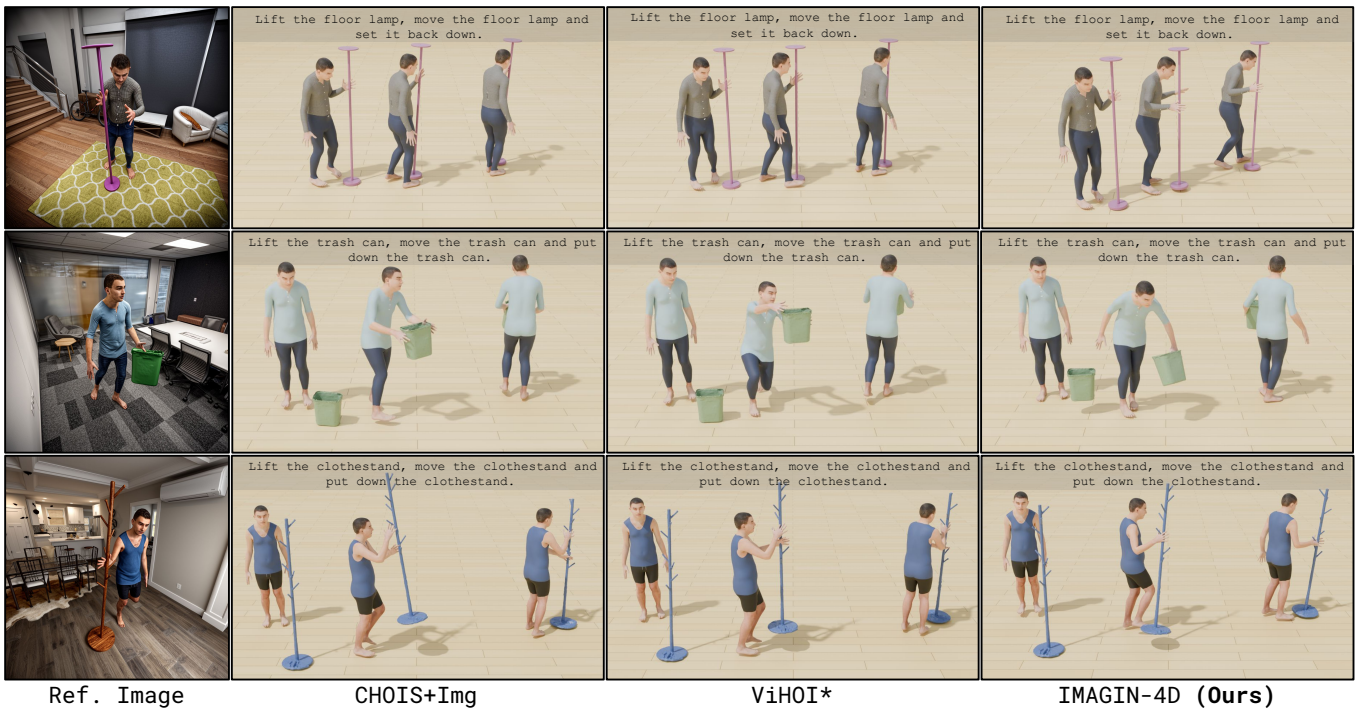

Fig. 3. **Qualitative comparison on FullBodyManipulation (Sec. 4.1).** Each row shows a reference image (*EditImage*) and generated motion from CHOIS+Img, ViHOI★, and IMAGIN-4D, all conditioned on the same text prompt, object shape, waypoints, and reference image. Single-token image conditioning often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

4.2 Cross-Domain Image Generalization

Table 3 tests whether image conditioning transfers across rendering domains. The motion ground truth is fixed; only the conditioning image changes. Training on MeshImg transfers poorly to SceneImg or EditImg, with A_{GT} degrading by roughly 23–24 cm and FID increasing sharply. Training on SceneImg transfers much better to EditImg: A_{GT} increases by only 4.25 cm relative to same-domain SceneImg. Together with Fig. 5, this supports using SceneImg as the main training domain: it contains enough scene, lighting, texture, and material variation to transfer to more photographic references. Detailed cross-domain analysis is provided in Sup. Mat.

4.3 Qualitative Results

Figure 4 tests whether the generated motion causally depends on the image. We horizontally flip only the conditioning image at inference time, keeping text, waypoints, object geometry, conditioning frame, and random seed fixed. The generated grasp side and contact region change, showing that the model actually uses the image. The generated motion is not fully mirrored, because the non-image conditions remain unchanged and still need to be met; this is the desired behavior for a conditional generator that must reconcile image evidence with text, object geometry, and waypoints.

Finally, Fig. 6 shows sketch-to-motion as a downstream authoring application. We retrain the image branch using line drawings from the same rendered reference frames. Image adherence is lower for sketch conditioning than for RGB conditioning because line drawings remove appearance cues that help localize contact and pose; quantitative results are provided in Sup. Mat. This shows that the method is not limited to photometric RGB inputs, although sketches remove appearance cues useful for contact and pose localization.

5 Conclusion

We present IMAGIN-4D, an image-guided HOI generator that uses a reference interaction snapshot to control 4D human-object motion. Our results show that reference images are useful for HOI generation only when their spatial cues are preserved and propagated across time. A single pooled visual token can improve motion metrics like FID and R-Precision, but it collapses the fine-grained cues needed to recover the depicted contact, object pose, and body-object layout. In contrast, spatially factorized tokens recover the reference interaction state, and frame-aware tokens extend this visual influence to the surrounding motion frames. Role-aware conditioning then preserves these visual cues by routing image, text, and waypoint signals through separate streams rather than merging them into one conditioning signal. Our rendering pipeline and image-adherence metric enable controlled evaluation across image domains, including sketches, and across datasets like FBM and BEHAVE. Together, these results establish spatio-temporal image conditioning as a practical way to specify fine-grained human-object interactions beyond text and sparse waypoints, without sacrificing motion quality. Code and trained models will be publicly released.

Fig. 4. **Flip-image consistency test (Sec.4.3).** 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.

Fig. 5. **Conditioning image domains (Sec. 4.2).** 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.

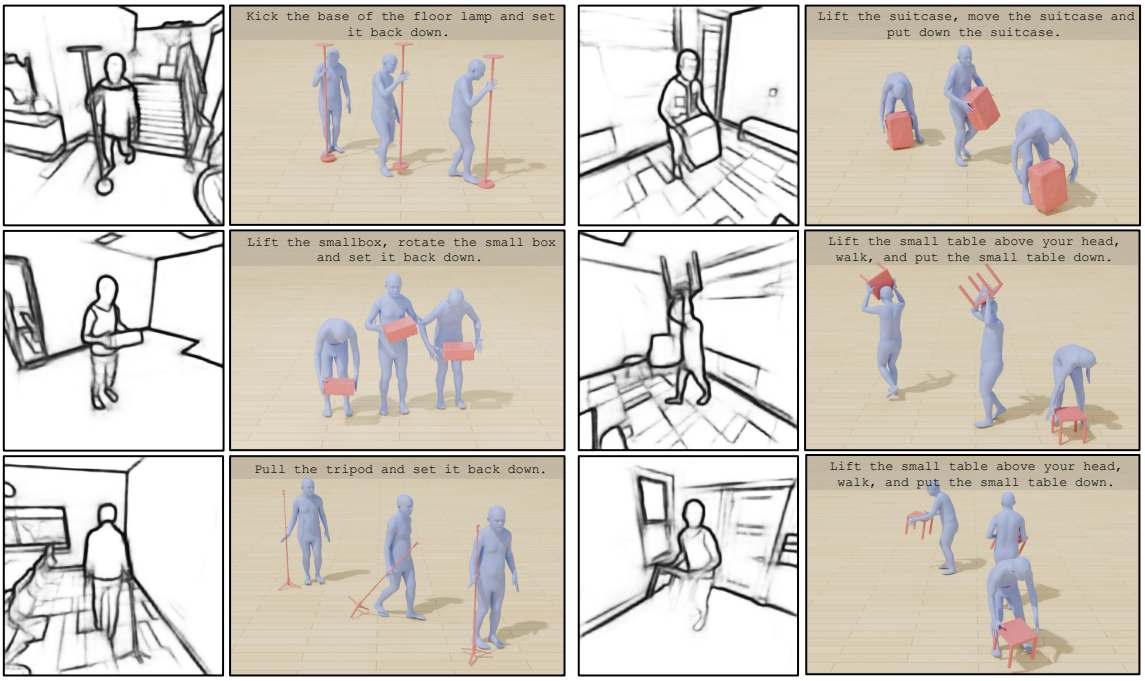

Fig. 6. **Sketch-to-motion (Sec.4.3).** We replace the RGB reference image with a line drawing and retrain the model. Despite removing texture, color, and scene appearance, the model preserves the depicted interaction layout and generates a complete motion sequence. This shows that IMAGIN-4D can also support sketch-based conditioning, where users specify the desired contact and body-object arrangement with a simple drawing.

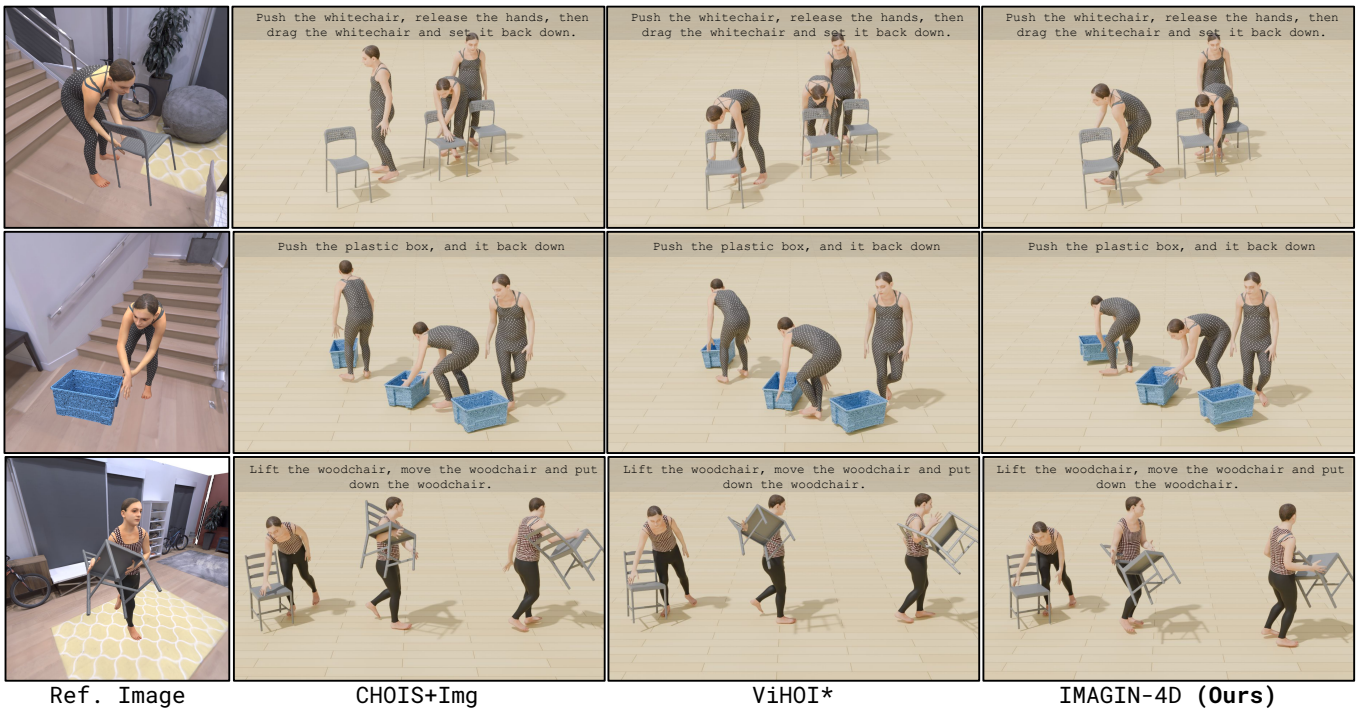

Fig. 7. **Qualitative comparison on FullBodyManipulation (SceneImg) (Sec. 4.1).** Each row shows the *SceneImg* reference and generated motion from CHOIS+Img, ViHOI★, and IMAGIN-4D under the same text prompt, object shape, waypoints, and reference image. Single-token image conditioning often misses contact, hand placement, object orientation, or body-object layout. IMAGIN-4D better matches the depicted interaction state.

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. 2025. From Sparse Signal to Smooth Motion: Real-Time Motion Generation with Rolling Prediction Models. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2022. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. 2023. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. 2026. ViHOI: Human-Object Interaction Synthesis with Visual Priors. *arXiv preprint arXiv:2603.24383* (2026).

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. 2024. Text2HOI: Text-Guided 3D Motion Generation for Hand-Object Interaction. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. 2024. DiffH2O: Diffusion-Based Synthesis of Hand-Object Interactions from Textual Descriptions. In *SIGGRAPH Asia 2024 Conference Papers*. ACM.

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

Christian Diller and Angela Dai. 2024. CG-HOI: Contact-Guided 3D Human-Object Interaction Generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Alexey Gavryushin, Alexandros Delitzas, Luc Van Gool, Marc Pollefeys, Kaichun Mo, and Xi Wang. 2025. 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. 2023. 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. 2024. MoMask: Generative Masked Modeling of 3D Human Motions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. 2019. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints. In *International Conference on Computer Vision (ICCV)*. IEEE.

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

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

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2024. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and Controllable Image Synthesis with Composable Conditions. In *International Conference on Machine Learning (ICML)*. PMLR.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 2022. InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction. In *German Conference on Pattern Recognition (GCPR)*. Springer.

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

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. 2024a. Autonomous Character-Scene Interaction Synthesis from Text Instruction. In *SIGGRAPH Asia 2024 Conference Papers*. ACM.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. 2023b. Full-Body Articulated Human-Object Interaction. In *International Conference on Computer Vision (ICCV)*. IEEE.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. 2024b. Scaling Up Dynamic Human-Scene Interaction Modeling. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. 2024. Optimizing Diffusion Noise Can Serve As Universal Motion Priors. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *International Conference on Computer Vision (ICCV)*. IEEE.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. 2025. ParaHome: Parameterizing Everyday Home Activities Towards 3D Generative Modeling of Human-Object Interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. 2024. NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. 2024b. ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation. *arXiv preprint arXiv:2412.18600* (2024).

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. 2024a. CHOIS: Controllable Human-Object Interaction Synthesis. In *European Conference on Computer Vision (ECCV)*. Springer.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*. PMLR.

Jiaman Li, Jiajun Wu, and C. Karen Liu. 2023c. Object Motion Guided Human Motion Synthesis. *Transactions on Graphics (TOG)* 42, 6 (2023).

Lei Li and Angela Dai. 2024. GenZI: Zero-Shot 3D Human-Scene Interaction Generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 2022. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yun Liu, Haolin Yang, Xu Si, Lei Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. 2024. TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. 2025. HUMOTO: A 4D Dataset of Mocap Human Object Interactions. In *International Conference on Computer Vision (ICCV)*. IEEE.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. 2024. HIMO: A New Benchmark for Full-Body Human Interacting with Multiple Objects. In *European Conference on Computer Vision (ECCV)*. Springer.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2024. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. In *AAAI Conference on Artificial Intelligence*. AAAI Press.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. 2024. 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. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *International Conference on Computer Vision (ICCV)*. IEEE.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2025. HOI-Diff: Text-Driven Synthesis of 3D Human-Object Interactions using Diffusion Models. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*. IEEE.

Sergey Prokudin, Christoph Lassner, and Javier Romero. 2019. Efficient Learning on Point Clouds with Basis Point Sets. In *International Conference on Computer Vision (ICCV)*. IEEE.

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. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning (ICML)*. PMLR.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. 2025. HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization. *arXiv preprint arXiv:2506.15625* (2025).

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. 2024. Human Motion Diffusion as a Generative Prior. In *International Conference on Learning Representations (ICLR)*. OpenReview.

Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. 2024. HOIAnimator: Generating Text-Prompt Human-Object Animations Using Novel Perceptive Diffusion Models. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797* (2019).

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. 2022. GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision (ECCV)*. Springer.

Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soeren Pirk, and Michael J. Black. 2024. GRIP: Generating Interaction Poses Using Spatial Cues and Latent Consistency. In *International Conference on 3D Vision (3DV)*. IEEE.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. 2024. 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. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations (ICLR)*. OpenReview.

Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. 2024. TLControl: Trajectory and Language Control for Human Motion Synthesis. In *European Conference on Computer Vision (ECCV)*. Springer.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. 2023. PhysHOI: Physics-Based Imitation of Dynamic Human-Object Interaction. *arXiv preprint arXiv:2312.04393* (2023).

Yin Wang, Ziyao Zhang, Zhiying Leng, Haitian Liu, Frederick W. B. Li, Mu Li, and Xiaohui Liang. 2026. Multimodal Priors-Augmented Text-Driven 3D Human-Object Interaction Generation. *arXiv preprint arXiv:2602.10659* (2026).

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2024. Move as You Say, Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2024. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *International Conference on Learning Representations (ICLR)*. OpenReview.

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. 2025a. InterAct: Advancing Large-Scale Versatile 3D Human-Object Interaction Generation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. 2023. InterDiff: Generating 3D Human-Object Interactions with Physics-Informed Diffusion. In *International Conference on Computer Vision (ICCV)*. IEEE.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. 2025b. InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. 2024a. InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024b. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. 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. 2024. Generating Human Interaction Motions in Scenes with Text Control. In *European Conference on Computer Vision (ECCV)*. Springer.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. PhysDiff: Physics-Guided Human Motion Diffusion Model. In *International Conference on Computer Vision (ICCV)*. IEEE.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. 2024. OakInk2: A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. 2024b. GraspXL: Generating Grasping Motions for Diverse Objects at Scale. In *European Conference on Computer Vision (ECCV)*. Springer.

Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. 2025. InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. 2024c. HOI-M3: Capture Multiple Humans and Objects Interaction within Contextual Environment. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023c. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024a. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2024).

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. 2023a. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. In *International Conference on Computer Vision (ICCV)*. IEEE.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. 2023b. FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. 2022. COUCH: Towards Controllable Human-Chair Interactions. In *European Conference on Computer Vision (ECCV)*. Springer.

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

Yang Zhao, Yan Zhang, and Xubo Yang. 2025. 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. 2022. TOCH: Spatio-Temporal Object-to-Hand Correspondence for Motion Refinement. In *European Conference on Computer Vision (ECCV)*. Springer.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. In *European Conference on Computer Vision (ECCV)*. Springer.

Supplementary Material

IMAGIN-4D: Image-Guided Controllable Interaction Generation

Overview. This supplement provides metric definitions, image-adherence details, data generation, baseline construction, architecture and training details, ablations, cross-domain analysis, BEHAVE protocol notes, sketch-conditioning results, qualitative probes, and failure cases.

A 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_{GT}, evaluated exactly at $t^\star$; A_{W10}, the best score within a ± 10-frame window around $t^\star$; and A_{Any}, the best score over the full 120-frame sequence. Lower is better. For details, see Sec. B.

Motion quality and text alignment. FID and R-Precision are computed using the frozen CHOIS text-to-motion evaluator [Li et al. 2024a]. 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 BEHAVE contact annotations.

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.

B 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

$$d(t) = 0.8\, d_{\mathrm{hum}}(t) + 0.2\, \|\mathbf{o}_t - \mathbf{o}^\star\|_2, \tag{11}$$

where

$$d_{\mathrm{hum}}(t) = \sum_{k=1}^{24} w_k \|\mathbf{j}_{t,k} - \mathbf{j}^\star_k\|_2, \qquad w_k = \frac{\exp\left(-\|\mathbf{j}^\star_k - \mathbf{o}^\star\|_2/\sigma\right)}{\sum_{r=1}^{24} \exp\left(-\|\mathbf{j}^\star_r - \mathbf{o}^\star\|_2/\sigma\right)}, \tag{12}$$

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

$$A_{\mathrm{GT}} = d(t^\star), \tag{13}$$

$$A_{\mathrm{W10}} = \min_{|t-t^\star|\le 10} d(t), \tag{14}$$

$$A_{\mathrm{Any}} = \min_{1\le t\le T} d(t). \tag{15}$$

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

C Data Generation Details

FullBodyManipulation (FBM) [Li et al. 2023c] and BEHAVE [Bhatnagar et al. 2022] provide human-object motion but not reference images for image-conditioned generation. We therefore render reference images from ground-truth motion sequences.

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:

$$t^\star = \mathrm{round}\left[\frac{\sum_t t(c^{\mathrm{L}}_t + c^{\mathrm{R}}_t)}{\sum_t (c^{\mathrm{L}}_t + c^{\mathrm{R}}_t)}\right]. \tag{16}$$

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.

SceneImg is used for the main FBM training and evaluation because it contains scene context, lighting, shadows, clothing texture, and object appearance. MeshImg is used for controlled ablations and for BEHAVE, where SMPL-H bodies do not share the SMPL-X UV layout used by BEDLAM textures. 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 512×512 using Blender Cycles. The Cycles render is deterministic for a fixed sequence, scene, and camera, so all methods receive identical conditioning images.

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.

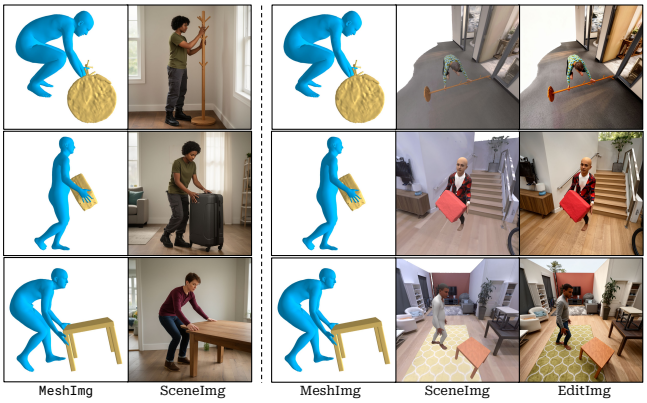

Fig. 8. **Why EditImg is derived from SceneImg.** Columns 1–2 show direct MeshImg→EditImg results. Columns 3–5 show the same interaction as MeshImg, SceneImg, and SceneImg-derived EditImg. Direct MeshImg editing can preserve the interaction in some cases, but can also change human pose, object pose, 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.

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. 8 shows two failure cases and one partially successful MeshImg edit.

D Baseline Construction

Image-free baselines. We report the published FBM numbers from CHOIS [Li et al. 2024a]. 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*.
Controlled encoder-interface comparison. The Qwen-VL pooled baseline and ViHOI* 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.

E 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. 2023a] 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.

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

F Architecture Details

Spatially factorized image encoder. The image encoder extracts 16×16 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.

G 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.

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.

The diagnostic supports the ablation in Tab. 5: replacing the DINOv2 patch grid with Qwen2.5-VL hidden states worsens A_{Any} by +2.1 cm at similar FID and waypoint error. This does not imply that DINOv2 is a stronger visual encoder in general. Rather, for our factorized patch-level conditioning, DINOv2 preserves spatial variation that the contact, pose, object, and layout heads can exploit. Qwen2.5-VL features are semantically stronger but more globally compressed, which is less suited to this interface.

H Ablations

Tab. 5 reports single-axis ablations anchored on full IMAGIN-4D, corresponding to the "Ours (Spatial+Temporal)" row of Tab. 1. Each row replaces one component choice while holding rest fixed.

Variant	Image Adh. (cm)↓			FID↓	C_{F_1} ↑	WP
	GT	W10	Any			(cm)↓
IMAGIN-4D, full	**8.43**	**7.65**	**7.45**	**0.28**	0.677	5.69
(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_{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 5. **Single-axis ablations on FullBodyManipulation (FBM) dataset** (482-window full eval, step 200k). The first row reports the full IMAGIN-4D (matching the "Ours (Spatial+Temporal)" row of Tab. 1). Each subsequent row replaces *one* component or recipe choice with the rest of the system held fixed. 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.

Spatial decomposition (A). The four supervised Q-Former heads encode contact, human pose, object pose, and body-object layout from the same patch grid. Rows A.1–A.4 remove one role at a time. The layout head has the largest effect on A_{Any}: cross-frame approach paths rely on body-to-object directions, and removing it costs 1.5 cm of A_{Any} at otherwise neutral waypoint error. The human-pose head produces the largest FID hit, consistent with its role as the body-realism anchor at $t^\star$. Removing the contact head reduces C_{F_1} from 0.677 to 0.60 while leaving waypoint and motion-quality metrics nearly intact, isolating the contact pathway. The object-pose head is the only one whose removal visibly degrades waypoint error (+0.7 cm); its removal has the smallest effect on adherence, matching its smaller feature dimension and supervision weight.

Image classifier-free guidance (B). Disabling image CFG (s_{img}=1) costs ∼ 1 cm of A_{Any} at neutral FID and waypoint error, confirming the headline number's dependence on this sampling-time lever.

Role-aware conditioning (C). The denoiser routes text, sparse waypoints, and window-gated image evidence through separate AdaLN streams, and frame-aware tokens through final cross-attention (Sec. 3.3). Row C.1 collapses the image route; the placeholder row swaps the frame-aware route from cross-attention to AdaLN.

Removing the image AdaLN stream (C.1) merges the spatial-token summary into the base AdaLN with text and time, while keeping the waypoint stream and frame-aware cross-attention. A_{GT} regresses from 8.4 to 13.4 cm and A_{W10} from 7.65 to 10.4 cm. Waypoint error

improves to 3.8 cm because the shared AdaLN now spends its capacity on text and trajectory rather than reconstructing the image. This isolates the trade-off referenced in Sec. 4.1: image and waypoint conditioning compete for the same modulation channel unless they are routed separately. The bi-stream variant therefore sits on a Pareto front opposite the full method — it pays 5 cm of $\hat{t}$-frame adherence to recover 1.9 cm of waypoint precision. The full IMAGIN-4D selects the adherence-prioritized operating point of this front, in line with the controllable-generation framing of the paper.

The frame-aware-via-AdaLN row swaps the last-layer cross attention for a per-layer AdaLN stream on the same per-frame frame-aware tokens, keeping every other component fixed. $\hat{t}$-frame adherence is comparable to the cross attention route (A_{GT} 8.4 → 9.2 cm) because frame-aware tokens contribute marginally to adherence in both routings, but FID regresses sharply (0.28 → 0.60) and contact F_1 drops by 1.7 pp. Per-layer multiplicative AdaLN amplifies the unsupervised frame-aware token noise across the entire decoder stack, while last-layer cross-attention applies the same evidence once and leaves the earlier-layer signal flow unperturbed. Cross-attention is therefore the preferred routing for tokens that are trained only through the denoising objective.

Visual input modality (D). Replacing the RGB reference with line-drawing sketches at training and test time costs ∼1.5 cm of A_{Any} at near-neutral FID and waypoint error: sketches retain silhouette and object outline but lose the appearance cues that pin down contact side, body texture, and object material. See Sec. K for details.

I Cross-Domain Transfer

We evaluate cross-domain transfer using MeshImg, SceneImg, and EditImg. The motion sequence and target interaction state are fixed; only the conditioning image domain changes. This isolates visual-domain transfer from motion-distribution transfer.

Same-domain rows define the no-shift setting. MeshImg is slightly easier than SceneImg because it removes background clutter, texture variation, and lighting. However, MeshImg training transfers poorly to SceneImg and EditImg because the image heads are trained on white-background body-object renders and then tested on textured scenes or edited photographic images.

SceneImg training transfers better to EditImg. SceneImg already contains indoor context, lighting, shadows, body texture, and object material, so editing mainly changes appearance while preserving the body-object layout. This supports using SceneImg as the main FBM training domain and SceneImg→EditImg as the controlled photorealistic transfer setting.

J 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 MeshImg reference images.

We use MeshImg because BEHAVE provides SMPL-H bodies, while our SceneImg and EditImg pipeline uses SMPL-X-compatible BEDLAM body textures [Black et al. 2023]. This keeps BEHAVE focused on architectural transfer rather than mixing motion-distribution shift with rendering-domain shift.

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

K 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 97.32 mm A_{W10} compared to 76.79 mm A_{W10} for the matched RGB model; see Table. 5. FID remains similar. Image classifier free guidance improves sketch adherence. These results show that sketches provide usable control for controllable generation of motion.

L 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.

M 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.

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.

N 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 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_{GT}, A_{W10}, and A_{Any}.