POSESTAGE: UNIFYING EXTERNAL POSE AND INTERNAL STRUCTURE CONTROL IN MULTI-OBJECT SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable image generation can place objects, specify their orientation, or articulate a human body, but in scenes with several open-category objects these controls are rarely available at the same time. We present PoseStage, a unified framework that controls both the external pose and the internal structure of multiple objects from two 2D geometric maps. PoseStage supports three novel pose-and-structure tasks within one framework: generation, reference-guided generation, and editing. A model trained only on standard conditional generation handles the latter two at inference time, without task-specific training, by extending its input sequence. To prevent the appearance and text of different instances from mixing, we apply a training-free instance-routing mask at inference that removes direct attention between instances and reduces attribute confusion. Training a single model to control both aspects requires images annotated with external pose and internal structure at the same time, which existing datasets do not provide. We therefore build a hybrid dataset of real-world and synthetic images that provides aligned annotations for external pose and internal structure, and a benchmark that evaluates both controls, on which PoseStage outperforms existing methods.

1 INTRODUCTION

Controllable image generation lets users specify not only what appears in a scene but also the geometry of each object. The external pose of an object specifies where it appears, how large it is, and how it is oriented relative to the camera; its internal structure captures the shape and part configuration of the object itself. Most existing methods, however, control only one of these two aspects. In a natural scene, several open-category objects often appear together, and their image location, scale, 3D orientation, and internal structure may all need to be controlled at once.

Prior work provides useful ideas for representing these controls. Dense coordinate maps can project camera-relative translation, scale, and rotation into the image plane, as demonstrated by SceneDesigner (Qin et al., 2025). Category-agnostic curves can describe internal structure without assuming a human-like joint system, as demonstrated by SKDream (Xu et al., 2025). But neither representation alone is sufficient to solve this problem. External pose alone does not describe internal deformation, while an object-centered structural condition cannot handle real multi-object scenes. This raises the question: how can we simultaneously control the external pose and internal structure of multiple objects within a unified framework?

To address this problem, we propose PoseStage, which encodes the external pose and internal structure of multiple objects as 2D images and feeds them as conditioning inputs to the generator. Scenes with multiple objects, however, raise an additional challenge: a generator may assign one object's text or appearance to a different object. To reduce this attribute confusion, we apply an instance-routing mask inside the MMDiT backbone. The mask removes direct attention between different instances, so each instance primarily attends to its own text description or reference image, while shared spatial tokens maintain the overall scene coherence.

Because this routing mask is applied only at inference, it requires no additional training: the model is trained purely on standard conditional generation, without multi-instance appearance supervision or source–target edit pairs. Having learned a reliable mapping from the geometric maps to object pose

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.

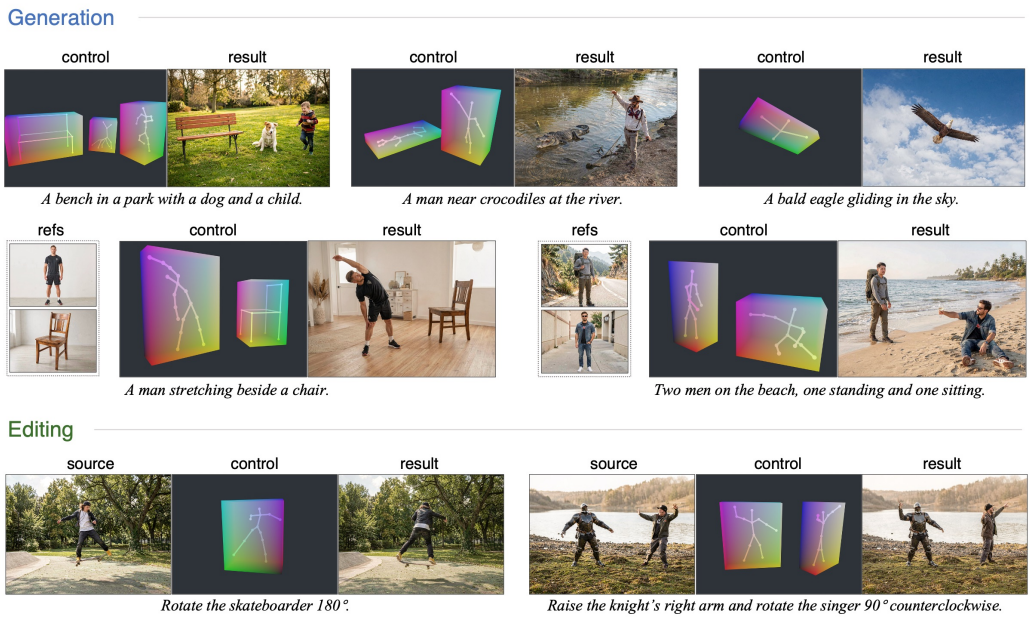

Figure 1: **Applications of PoseStage.** (A) **generation**: PoseStage synthesizes images based on text prompts and unified geometric controls. It supports both single-object and multi-object generation with optional appearance references. (B) **editing**: PoseStage modifies the pose and structure of target instances while seamlessly preserving the original background and unedited elements.

and structure, and assisted by the strong generalization of the pretrained backbone, the same model supports three tasks at inference by extending its input sequence: pose-and-structure generation, reference-guided pose-and-structure generation, and pose-and-structure editing.

Our contributions are threefold. First, we propose PoseStage, a unified framework that controls both external pose and internal structure for multiple objects from two 2D geometric maps, one for each, and demonstrate that it enables the three tasks above, which prior work does not jointly support. Second, to supply the aligned supervision that this requires, we build a hybrid real–synthetic dataset that pairs natural scene diversity with exact synthetic geometry. Third, we construct a benchmark that jointly evaluates external pose and internal structure in both single- and multi-object scenes, on which PoseStage outperforms existing methods.

2 Related Work

Layout and image-conditioned generation. Layout-guided diffusion models associate text entities with boxes, masks, points, or regional tokens. GLIGEN inserts grounded tokens through gated layers, LayoutDiffusion models object-aware regions, and InstanceDiffusion supports several instance representations (Li et al., 2023; Zheng et al., 2023; Wang et al., 2024b). ControlNet, T2I-Adapter, and Uni-ControlNet establish the broader paradigm of attaching dense image conditions to pretrained generators (Zhang et al., 2023; Mou et al., 2024; Zhao et al., 2023). DreamRenderer uses bridge tokens to associate local text and image evidence without retraining the attention layers (Zhou et al., 2025). These methods provide strong spatial guidance, but a box, mask, or generic control image does not by itself specify both camera-relative 9-DoF pose and open-category internal structure.

3D pose and structural control. SceneDesigner represents the translation, scale, and rotation of several objects with object-coordinate surfaces rendered from the camera view (Qin et al., 2025), and SeeThrough3D conditions on translucent, orientation-colored 3D boxes to make multi-object placement occlusion-aware (Agrawal et al., 2026). HumanSD, Stable-Pose, HyperHuman, and SMPL-conditioned methods rely on human-specific joints or body priors (Ju et al., 2023; Wang et al., 2024a; Liu et al., 2024; Buchheim et al., 2025). SP-Ctrl extends sparse keypoint control (Cao et al., 2021) to

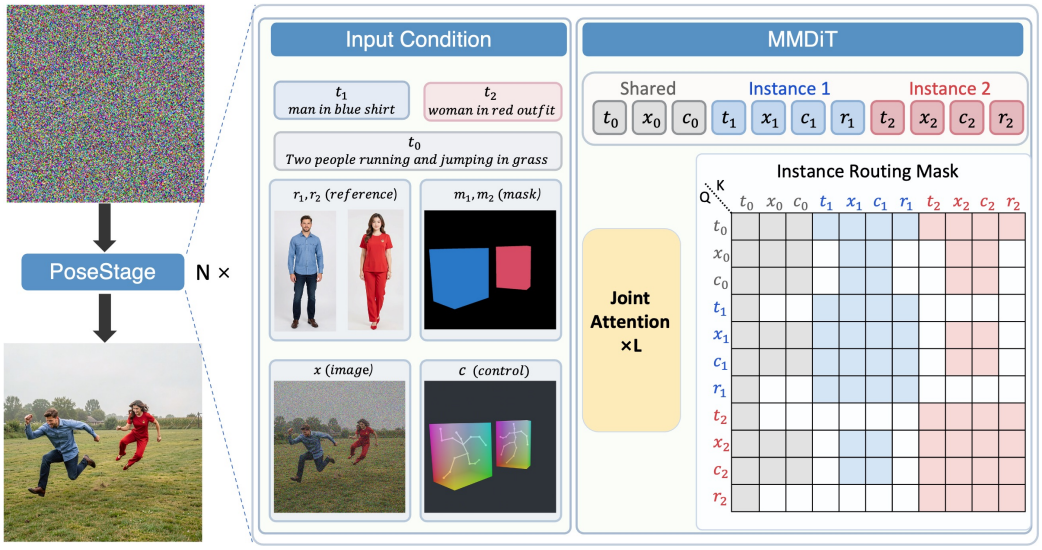

Figure 2: **Overview of PoseStage.** Each instance owns its text, control, image, and reference tokens (t_i, c_i, x_i, r_i); shared tokens (t_0, x_0, c_0) carry the background. These form one input sequence for the MMDiT, where an instance-routing mask blocks cross-instance attention so appearance and text stay bound to the correct object, while shared spatial tokens keep the scene coherent.

both human and animal categories (Xuan et al., 2025). SKDream instead extracts category-agnostic curves from meshes and renders them with coordinate colors (Xu et al., 2025). PoseStage unifies these two lines by encoding both 9-DoF external pose and internal structure in a single scene.

3D-aware image editing. 3D-aware editors seek geometrically meaningful object transformations while preserving identity and scene consistency. OBJECT 3DIT learns language-specified translation, rotation, insertion, and removal from procedurally rendered edit pairs, whereas Neural Assets disentangles appearance from target 3D pose using paired video frames for multi-object manipulation (Michel et al., 2023; Wu et al., 2024). PoseStage instead performs 3D-aware editing within the same framework it uses for generation, at inference and without task-specific training, driving both the target 9-DoF pose and the internal structure from the two geometric maps.

3 Method

3.1 Geometric Conditions

A scene contains N instances. Each instance i has a 9-DoF external pose $\mathbf{p}_i = (\mathbf{t}_i, \mathbf{s}_i, \mathbf{R}_i)$, consisting of camera-relative translation $\mathbf{t}_i \in \mathbb{R}^3$, anisotropic scale $\mathbf{s}_i \in \mathbb{R}^3_+$, and rotation $\mathbf{R}_i \in SO(3)$. Additionally, its internal structure is represented as a category-agnostic geometric graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$. Both representations are rendered as 2D maps aligned with the output image.

External map. The external pose is rendered as a camera-view bounding-box coordinate map $\mathbf{C}^{\text{ext}}$. Following the CNOCS (Cuboid Normalized Object Coordinate System) representation in SceneDesigner (Qin et al., 2025), we construct this map by rasterizing the visible surfaces of the object's 3D bounding box. The resulting map captures the object's translation, scale, and rotation.

Internal map. The internal structure is rendered as an internal map $\mathbf{C}^{\text{int}}$ by rasterizing the geometric graph $\mathcal{G}_i$. Because both the projected positions and the coordinate colors are produced by Eq. 4, this map captures the internal structure while also implicitly encoding the projected pose.

3.2 Model Architecture

As illustrated in Fig. 2, a scene contains a global text prompt t_0 describing the background, and N instances. Each instance i provides an open-vocabulary text phrase t_i, an optional appearance reference image r_i, and a spatial mask $\mathbf{M}_i$. The mask $\mathbf{M}_i$ is not an extra input the user must draw: it is the image-plane projection of instance i, which we obtain directly from the geometric maps as the region covered by the object's projected bounding box. The masks are downsampled to the latent grid resolution using nearest-neighbor interpolation, and overlapping pixels are assigned according to the scene order, yielding disjoint masks $\{\overline{\mathbf{M}}_i\}_{i=1}^{N}$.

Input sequence. We build PoseStage on FLUX.2 [klein] (Black Forest Labs, 2026), an MMDiT that natively accepts one or more reference images, encoding each with its VAE into tokens concatenated with the noisy image latents. We reuse this native channel for both per-instance appearance references and, in editing, the source image, without adding any extra encoder or adapter. Let E denote the image encoder, which is the same pretrained VAE used for encoding the image latents, and let Pack denote its subsequent patch tokenization operation. We first encode the two geometric maps separately to preserve their distinct and complementary information, $\mathbf{L}^{\mathrm{ext}} = \mathrm{Pack}(E(\mathbf{C}^{\mathrm{ext}}))$ and $\mathbf{L}^{\mathrm{int}} = \mathrm{Pack}(E(\mathbf{C}^{\mathrm{int}}))$: the external map provides an explicit global anchor for absolute position and scale, while the internal map describes the structure and, through its coordinate-as-color scheme, implicitly encodes the projected pose. A trainable fusion layer then combines them into the unified spatial control sequence c:

$$c = W_f \begin{bmatrix} \mathbf{L}^{\mathrm{ext}} \\ \mathbf{L}^{\mathrm{int}} \end{bmatrix}, \qquad W_f^{(0)} = \tfrac{1}{2} \begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix}, \tag{1}$$

where the weights of the fusion layer W_f are initialized as $\frac{1}{2}[\mathbf{I} \quad \mathbf{I}]$; W_f is then learned end-to-end, allowing the model to adjust the relative contribution of the two conditions.

We then construct the complete input sequence by concatenating four types of tokens: text (t), noisy image (x), spatial control (c), and reference image (r). The spatial control c and noisy image x are shared across the entire scene. To route the correct conditions to the corresponding regions, we explicitly partition them into mutually exclusive subsets based on the disjoint masks $\{\overline{\mathbf{M}}_i\}$. The global group (index 0) owns the global text t_0, as well as the background tokens x_0 and c_0 extracted using the inverted union mask $1 - \bigcup_{i=1}^{N} \overline{\mathbf{M}}_i$. Each local group i also owns its specific text phrase t_i and reference image r_i. Finally, we concatenate these mutually exclusive subsets to form the complete input sequence $\mathcal{Q}$ fed into the MMDiT:

$$\mathcal{Q} = [t_0, t_1, \ldots, t_N \mid x_0, x_1, \ldots, x_N \mid c_0, c_1, \ldots, c_N \mid r_1, \ldots, r_N]. \tag{2}$$

Attention mask for instance routing. We achieve instance routing entirely through an attention mask applied across all joint attention layers. Formally, let $\mathcal{T}_i = t_i \cup r_i \cup x_i \cup c_i$ denote the set of all tokens associated with instance $i \in \{1, \ldots, N\}$, and $\mathcal{T}_0 = t_0 \cup x_0 \cup c_0$ for the background. We further define the complete token set $\mathcal{U} = \bigcup_{i=0}^{N} \mathcal{T}_i$ and the spatial token set $\mathcal{S} = \bigcup_{i=0}^{N} (x_i \cup c_i)$. The allowed query–key relation $\mathcal{R} \subseteq \mathcal{U} \times \mathcal{U}$ and the corresponding binary attention mask are defined as

$$\begin{aligned} \mathcal{R} &= (t_0 \times \mathcal{U}) \cup (\mathcal{U} \times t_0) \cup (\mathcal{S} \times \mathcal{S}) \cup \bigcup_{i=1}^{N} (\mathcal{T}_i \times \mathcal{T}_i), \\ A_{u,v} &= \mathbf{1}[(u, v) \in \mathcal{R}]\,. \end{aligned} \tag{3}$$

The relation consists of three components: the global text tokens t_0 interact bidirectionally with all tokens; all spatial tokens in $\mathcal{S}$ are mutually visible; and tokens within each local instance group $\mathcal{T}_i$ interact freely. All remaining query–key pairs are masked out. As a result, no query attends *directly* to the private text or reference tokens of another instance, so an instance's appearance and text are drawn mainly from its own group, while the globally connected spatial tokens maintain the scene coherence and blend the instances into a single image. The mask operates on direct attention within each layer, with the global text t_0 and the shared spatial tokens serving as common channels for scene-level coherence; its effect on attribute mixing is quantified in Sec. 5.

3.3 Training and Inference

Training objective. We train our model on the standard conditional image generation task. The training data consists of condition-image pairs with corresponding text prompts, without any appearance-reference supervision. We do not apply the instance routing mechanism during training; instead, the model employs standard full attention.

Training-free inference. Single-object generation. For basic generation, we disable instance routing and use standard full attention. The model synthesizes the object according to the geometric control maps and text. Providing an appearance reference image is optional; if included, the model transfers the reference texture to the target geometry without additional fine-tuning.

Multi-object generation. When synthesizing multiple instances in a single scene, we enable the instance-routing mask (Eq. 3). Appearance references remain optional for each instance.

Image editing. For controllable image editing, instance routing is disabled. The original image is fed into the network as a global appearance reference alongside an editing instruction and a target geometric control map. Because the source image conditions the whole scene, the background and unedited regions are substantially preserved, while the target control map drives the specified pose and structure changes.

4 Dataset

For PoseStage to generalize to complex applications at inference without task-specific training, the model must first learn a reliable alignment between spatial geometry and image appearance. This requires high-quality training images with well-aligned annotations for both external pose and internal structure. However, existing datasets typically provide at most one type of spatial control, hindering the development of unified models. To address this issue, we construct a hybrid dataset using two complementary pipelines (illustrated in Fig. 5): one using real photographs for natural diversity, and one using synthetic renderings for precise geometry. Both pipelines output training pairs in a unified format, comprising the target RGB image, text captions, and the two coordinate-as-color rendered spatial maps (i.e., the external map and internal map).

Real subset for natural diversity. To ensure the model generalizes to diverse real-world scenes, we draw our real subset from an internal image collection. To mitigate the noise in model predictions, we enforce a strict quality control process. We automatically filter out candidates exhibiting clear inconsistencies among masks, poses, and structures. The remaining samples undergo a rigorous manual inspection to ensure the predicted geometry is consistent with the visual evidence before being converted into the rendered spatial maps.

Synthetic subset for geometric precision. We sample textured 3D assets from an internal 3D asset collection, discarding those with invalid geometry or metadata, and extract the same category-agnostic geometric graph representation used for the real objects. We then procedurally generate Blender scenes containing one to three objects, randomizing camera viewpoints, object transformations, lighting, and materials. This synthetic pipeline directly exports the exact camera and object transformations alongside the rendered spatial maps. Per-source pair counts are reported in Appendix A.6.

5 Experiments

Implementation details. We build PoseStage on FLUX.2 [klein] (9B) (Black Forest Labs, 2026), fine-tuning the MMDiT backbone with LoRA (Hu et al., 2022) of rank 32 and training the geometric fusion layer W_f (Eq. 1) with full parameters. We use dynamic-resolution training so that objects of varied aspect ratios are preserved without distortion. The remaining optimization settings are listed in Appendix A.4.

Evaluation details. We evaluate on three benchmarks constructed from data sources disjoint from training, with every selected pose manually verified. **(i) Reference-free generation** and **(ii)**

Reference-conditioned generation are each split into Single-Object and Multi-Object subsets; the reference benchmark additionally attaches, for every instance, an appearance image synthesized by a text-to-image model from an *appearance-only* prompt (no pose description), so that geometry is specified solely by our control maps and appearance solely by the reference. **(iii) Editing** is a separately constructed benchmark covering both single- and multi-object scenes, and is partitioned into a *Rigid* subset (9-DoF pose changes only) and an *Arbitrary* subset that additionally includes non-rigid deformation and shape changes.

All methods are scored through the same measurement pipeline that recovers geometry from the generated image with SAM 3 (Carion et al., 2026)/SAM 3D (Chen et al., 2026a). Since this reconstruction can fail on missing or malformed objects, we report a **Reconstruction rate** (Rec., the fraction of target instances successfully reconstructed) and compute the geometric metrics on the reconstructed instances. Unlike prior pose-controllable generators that report only the azimuth error (Min et al., 2025; Qin et al., 2025), our orientation metric covers the full 3D rotation. The routing study additionally reports a **Binding accuracy** (Bind) that measures whether each appearance is routed to the correct object. Exact metric definitions are given in Appendix A.7.

5.1 Comparisons with State-of-the-Art Methods

Because jointly controlling 9-DoF external pose and internal structure of multiple objects is a new task, no existing method accepts the same input as PoseStage. We therefore follow one principle throughout: for every baseline we supply the target geometry as accurately as its interface allows, converting our conditions into the most expressive form it can consume. A method that natively takes a structured condition (a pose map, a box, or a skeleton) receives it; for a method that takes only text, we append a detailed GPT-authored (OpenAI, 2025) description of the same target pose and structure.

Reference-free generation. We compare against three closely related controllers, each evaluated on the controls it supports: SceneDesigner (Qin et al., 2025) controls camera-relative pose, SKDream (Xu et al., 2025) controls skeleton structure (single-object only), and InstanceDiffusion (Wang et al., 2024b) handles layout-grounded multi-instance placement. As shown in Table 1 and Fig. 3, PoseStage attains the best orientation, position, and structure scores on both subsets, with lower Trans.Err and Scale.Err indicating closer agreement with target centers and box dimensions after alignment. PoseStage is the only method that controls both external pose and internal structure while maintaining competitive image quality.

Table 1: **Reference-free generation.** Geometric metrics are computed on reconstructed instances; Rec. reports how often reconstruction succeeds. SKDream does not support multi-object scenes (reported on Single-Object only) and generates each object centered at a canonical scale, so its position metrics (Trans.Err, Scale.Err) are N/A.

			Orientation		Position & Scale		Structure		Quality
Subset	Method	Rec.↑	Ang.Err↓	Acc@22.5° ↑	Trans.Err↓	Scale.Err↓	Chamfer↓	F1@0.05↑	PickScore↑
Single	SceneDesigner	100.0	34.39	52.4	0.0798	0.3724	0.0547	0.577	21.059
	SKDream	98.4	46.83	56.5	N/A	N/A	0.0470	0.659	19.233
	InstanceDiffusion	100.0	86.69	19.0	0.2750	0.7681	0.0677	0.564	19.799
	PoseStage (Ours)	100.0	**17.87**	**81.0**	**0.0315**	**0.1701**	**0.0218**	**0.890**	**21.338**
Multi	SceneDesigner	100.0	57.53	37.8	0.2769	0.3667	0.0948	0.521	**20.287**
	InstanceDiffusion	100.0	83.15	47.9	0.3525	0.4868	0.1159	0.508	19.487
	PoseStage (Ours)	100.0	**26.93**	**72.6**	**0.1248**	**0.2016**	**0.0289**	**0.868**	19.994

Reference-conditioned generation. Here we compare against multi-subject layout methods that accept both spatial grounding and reference images: MS-Diffusion (Wang et al., 2025) and LaMIC (Chen et al., 2026b). Each baseline is provided with the per-instance boxes and references it supports, along with a GPT-authored description of the target pose and structure. Table 2 shows that PoseStage attains the highest appearance fidelity (CLIP-I, DINO) while retaining the geometric accuracy of the reference-free setting, confirming that the appended reference tokens map each reference's appearance to its target object without weakening the geometric control.

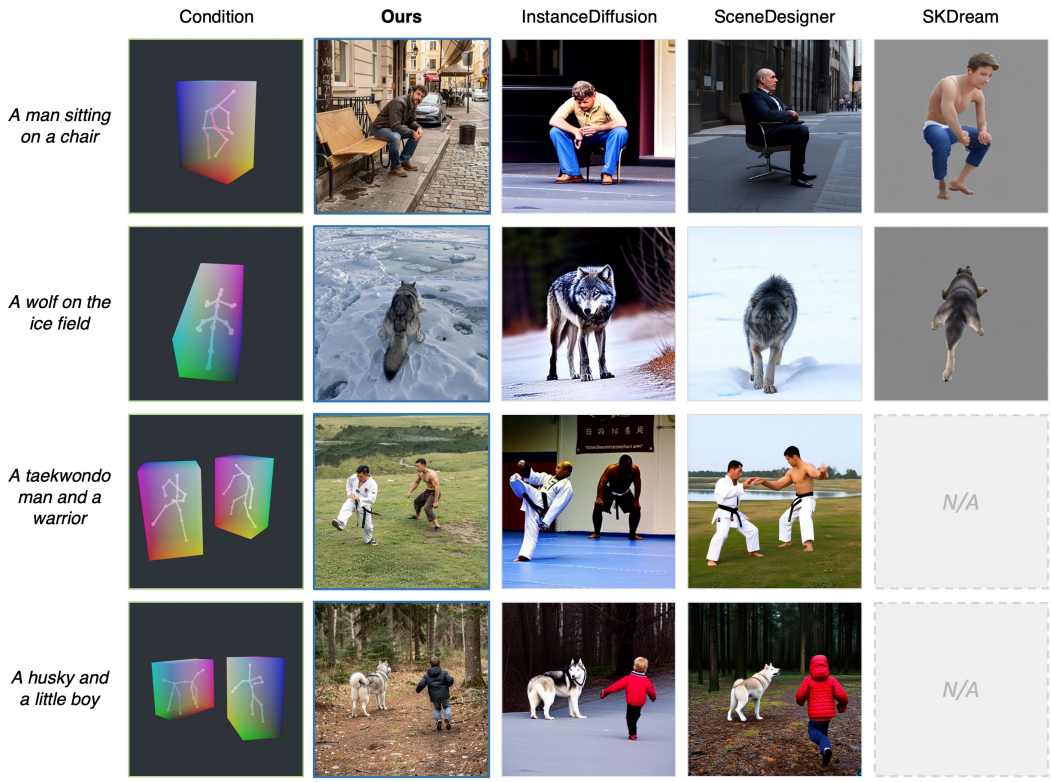

Figure 3: **Qualitative comparison on reference-free generation.** SKDream does not support multi-object scenes (bottom two rows marked N/A).

Table 2: **Reference-conditioned generation.** Appearance preservation (CLIP-I, DINO) measures identity fidelity, while the geometric columns confirm that pose and structure control is preserved when references are appended.

		Orientation		Position & Scale		Structure		Appearance	
Method	Rec.↑	Ang.Err↓	Acc@22.5° ↑	Trans.Err↓	Scale.Err↓	Chamfer↓	F1@0.05↑	CLIP-I↑	DINO↑
MS-Diffusion	99.6	84.61	14.4	0.6506	0.6590	0.1069	0.426	0.629	0.291
LaMIC	97.0	80.28	20.3	0.6568	0.6699	0.1466	0.364	0.552	0.163
PoseStage (Ours)	100.0	**21.77**	**80.7**	**0.0990**	**0.1872**	**0.0260**	**0.884**	**0.717**	**0.431**

Image editing. On the *Rigid* subset we compare against two 3D-aware editors, OBJECT 3DIT (Michel et al., 2023) and FreeFine (Zhu et al., 2025), as well as two general instruction editors, the closed-source GPT-Image-2 (OpenAI, 2026) and the open-source Qwen-Image-Edit (Wu et al., 2025), both driven by a detailed GPT-authored instruction describing the source-to-target pose difference. For multi-object scenes, OBJECT 3DIT and FreeFine edit one object at a time and are therefore applied in a multi-round manner. As reported in Table 3 and Fig. 4(a), PoseStage achieves the most accurate target orientation while preserving the background and object identity: OBJECT 3DIT and FreeFine yield less accurate poses and lower image quality, while GPT-Image-2 and Qwen-Image-Edit fail to achieve precise 9-DoF changes.

On the *Arbitrary* subset (Table 4 and Fig. 4(b)), which requires non-rigid deformation, we compare only against GPT-Image-2 and Qwen-Image-Edit, as the geometry-driven editors support rigid transforms exclusively. PoseStage reproduces the target skeleton far more faithfully, while the instruction-based editors, lacking a precise structural target, cannot transform the object to match the target pose.

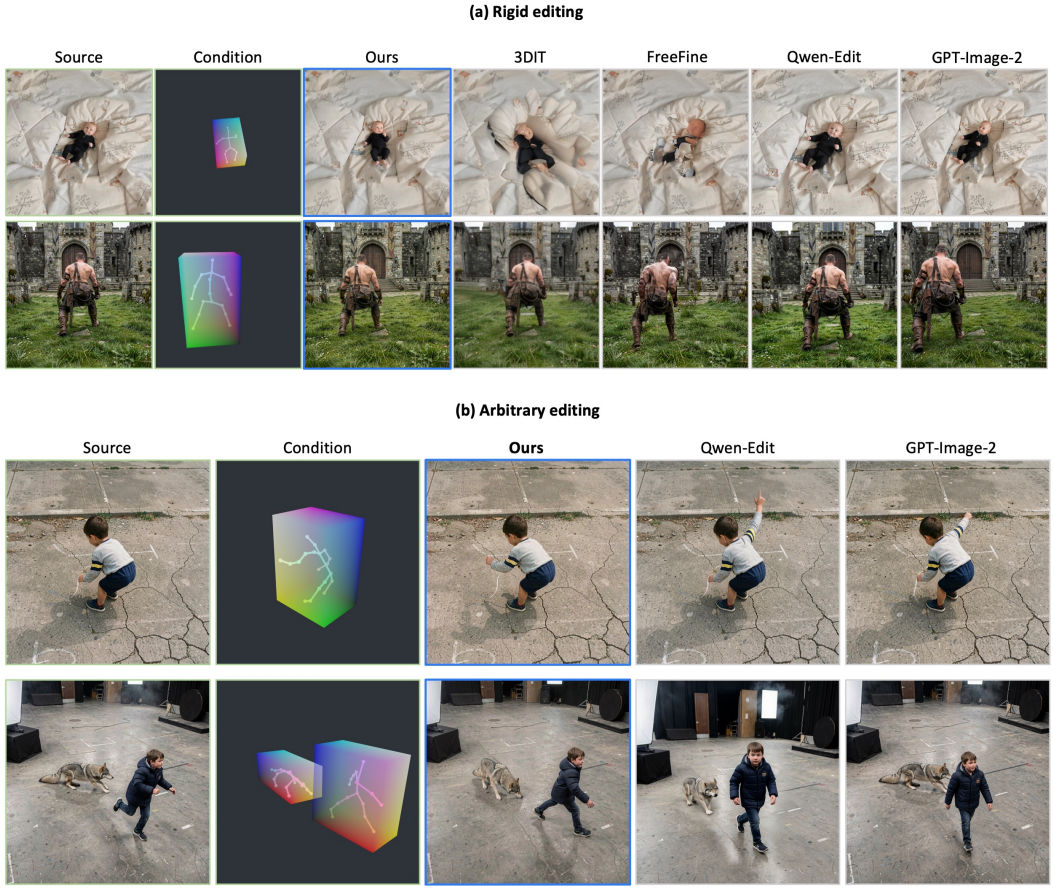

Figure 4: **Qualitative comparison on pose-and-structure editing. (a) Rigid** and **(b) Arbitrary** editing, driven by the target pose / structure. Here "3DIT" and "Qwen-Edit" abbreviate OBJECT 3DIT and Qwen-Image-Edit, respectively.

Table 3: **Image editing on the Rigid subset.** Methods marked † edit a single object per pass and are run multi-round on Multi-Object scenes.

		Orientation		Position & Scale		Preservation		Identity
Method	Rec.↑	Ang.Err↓	Acc@22.5° ↑	Trans.Err↓	Scale.Err↓	PSNR↑	SSIM↑	DINO↑
OBJECT 3DIT†	100.0	52.99	36.2	0.2227	0.3551	20.92	0.523	0.431
FreeFine†	100.0	60.23	33.0	0.1941	0.2639	21.71	0.569	0.668
Qwen-Image-Edit	100.0	56.38	29.8	0.2506	0.3212	14.22	0.374	0.787
GPT-Image-2	98.9	54.75	37.6	0.2608	0.2922	**28.85**	**0.778**	0.782
PoseStage (Ours)	100.0	**18.15**	**80.6**	**0.1464**	**0.1946**	21.87	0.760	**0.826**

5.2 Ablation Study

Geometric conditions. We ablate the two control maps on the reference-free benchmark by feeding only the external map, only the internal map, or both. Each variant is a separately trained model, retrained from scratch under the same training settings. As shown in Table 5, the internal-only variant already attains strong orientation and structure accuracy, since its coordinate-as-color curves implicitly encode the projected pose in addition to the structure. The external-only variant provides an explicit anchor for position and scale through its bounding-box map, but carries no internal signal, so its skeleton is largely random. The full model is best on orientation, position, and structure

Table 4: **Image editing on the Arbitrary subset** (non-rigid deformation). Geometry-driven editors are excluded as they support rigid transforms only.

		Orientation		Position & Scale		Structure		Preservation	
Method	Rec.↑	Ang.Err↓	Acc@22.5° ↑	Trans.Err↓	Scale.Err↓	Chamfer↓	F1@0.05↑	PSNR↑	SSIM↑
Qwen-Image-Edit	100.0	43.70	47.8	0.3350	0.4600	0.0582	0.617	19.13	0.555
GPT-Image-2	98.9	38.91	51.6	0.2124	0.3065	0.0475	0.706	**34.15**	0.769
PoseStage (Ours)	100.0	**22.32**	**79.3**	**0.1792**	**0.2133**	**0.0256**	**0.864**	22.62	**0.798**

together, confirming that the two maps encode complementary information that the fusion layer combines.

Table 5: **Ablation on geometric conditions** (reference-free generation).

		Orientation		Position & Scale		Structure		Quality
Variant	Rec.↑	Ang.Err↓	Acc@22.5° ↑	Trans.Err↓	Scale.Err↓	Chamfer↓	F1@0.05↑	PickScore↑
External map only	100.0	46.47	51.9	0.1477	0.2760	0.0491	0.650	**20.337**
Internal map only	100.0	31.14	66.7	0.1294	0.2234	0.0279	0.845	20.330
Full	100.0	**24.77**	**74.6**	**0.1025**	**0.1941**	**0.0272**	**0.873**	20.314

Instance routing. We evaluate the routing mask on multi-object scenes under both the reference-free and reference-conditioned settings, enabling and disabling it. Table 6 reports all four configurations together, alongside the orientation and position & scale metrics. Disabling routing lets instances share text or appearance from one another, which lowers binding accuracy and, in the reference-conditioned setting, appearance preservation; enabling it reduces this attribute mixing while leaving the geometric accuracy essentially unchanged. Since routing is applied only at inference, it improves multi-object results without any additional training.

Table 6: **Ablation on instance routing** (multi-object). Bind is the binding accuracy (Sec. 5); Bind and appearance preservation (CLIP-I, DINO) apply only to the reference-conditioned setting.

		Orientation		Position & Scale		Structure		Appearance		
Setting	Routing	Ang.Err↓	Acc@22.5° ↑	Trans.Err↓	Scale.Err↓	Chamfer↓	F1@0.05↑	Bind↑	CLIP-I↑	DINO↑
Ref-free	w/o route	27.22	**72.6**	**0.1137**	**0.2012**	0.0425	0.866	–	–	–
	w/ route	**26.93**	**72.6**	0.1248	0.2016	**0.0289**	**0.868**	–	–	–
Ref-cond	w/o route	24.35	77.1	**0.1200**	0.1939	0.0281	0.874	82.6	0.689	0.381
	w/ route	**24.07**	**79.1**	0.1208	**0.1896**	**0.0275**	**0.880**	**88.1**	**0.699**	**0.393**

5.3 User Study

On the reference-free benchmark, normalized human ratings follow the same ranking as our automatic metrics, with PoseStage rated highest on every axis, providing empirical support for using the automatic pipeline to compare methods. The full protocol and per-axis results are given in Appendix A.8.

6 Conclusion and Limitations

Our work has several limitations. First, the annotation pipeline for real images integrates segmentation, category-level canonicalization, monocular reconstruction, and curve extraction, so its controls are inherently noisy, and reliability degrades under severe occlusion or partial visibility as the estimators' errors compound. Second, constructing the internal structure takes more user effort than specifying the external pose. Third, a skeleton-based representation does not suit every category: objects such as vehicles or buildings have rigid, volumetric geometry that curve-like structures cannot naturally capture.

We presented PoseStage, a unified framework that controls the external pose and internal structure of multiple objects from two 2D geometric maps.

AI USE STATEMENT

We used generative AI tools to assist with translating and polishing the manuscript and to search for and identify relevant literature. We also used GPT-5 (OpenAI, 2025) to generate textual descriptions of target object poses and structures, as well as editing instructions describing source-to-target pose changes, for the baseline evaluations in Section 5. As described in that section, we used a text-to-image model to synthesize appearance reference images for the reference-conditioned generation benchmark. The authors take full responsibility for the final content of this work, including all AI-assisted text, references, instructions, and other artifacts.

REFERENCES

Vaibhav Agrawal, Rishubh Parihar, Pradhaan S Bhat, Ravi Kiran Sarvadevabhatla, and Venkatesh Babu Radhakrishnan. SeeThrough3D: Occlusion aware 3d control in text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Harry G. Barrow, Jay M. Tenenbaum, Robert C. Bolles, and Helen C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1977.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. `https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence`, 2026.

Benito Buchheim, Max Reimann, and Jürgen Döllner. Controlling human shape and pose in text-to-image diffusion models via domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. LAMIC: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026b.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. HyperHuman: Hyper-realistic human generation with latent structural diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. OBJECT 3DIT: Language-guided 3d-aware image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yunhong Min, Daehyeon Choi, Kyeongmin Yeo, Jihyun Lee, and Minhyuk Sung. ORIGEN: Zero-shot 3d orientation grounding in text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

OpenAI. Introducing GPT-5. `https://openai.com/index/introducing-gpt-5/`, 2025.

OpenAI. Introducing ChatGPT images 2.0 (GPT-Image-2). `https://openai.com/index/introducing-chatgpt-images-2-0/`, 2026.

Zhenyuan Qin, Xincheng Shuai, and Henghui Ding. SceneDesigner: Controllable multi-object image generation with 9-dof pose manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jiajun Wang, Morteza Ghahremani, Yitong Li, Björn Ommer, and Christian Wachinger. Stable-pose: Leveraging transformers for pose-guided text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations (ICLR)*, 2025.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. InstanceDiffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A. Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey R. Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuanyou Xu, Zongxin Yang, and Yi Yang. SKDream: Controllable multi-view and 3d generation with arbitrary skeletons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Wenjie Xuan, Jing Zhang, Juhua Liu, Bo Du, and Dacheng Tao. Rethink sparse signals for pose-guided text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. LayoutDiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. DreamRenderer: Taming multi-instance attribute control in large-scale text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Hanshen Zhu, Zhen Zhu, Kaile Zhang, Yiming Gong, Yuliang Liu, and Xiang Bai. Training-free geometric image editing on diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

A Supplementary Reproducibility Details

A.1 Geometric Conditions

Coordinate-as-color rendering. To render these 2D maps, we employ a coordinate-as-color strategy. Any canonical 3D point $\mathbf{p} \in [-1, 1]^3$ associated with object i is transformed by the object's external pose and projected onto the 2D image plane to obtain its pixel coordinate $\tilde{\mathbf{p}}$:

$$\tilde{\mathbf{p}} = \Pi_{\mathbf{K}}(\mathbf{R}_i \operatorname{diag}(\mathbf{s}_i)\mathbf{p} + \mathbf{t}_i)\,, \tag{4}$$

where $\operatorname{diag}(\mathbf{s}_i)$ represents the scale as a diagonal matrix, and $\Pi_{\mathbf{K}}$ denotes perspective projection with intrinsics $\mathbf{K}$. During rasterization, the canonical coordinate $\mathbf{p}$ is directly stored as the RGB color value at the projected location $\tilde{\mathbf{p}}$. This projection ensures that both geometric representations are strictly aligned in the image plane.

External map. Specifically, for each covered pixel, we cast a camera ray to find its camera-space 3D intersection $\mathbf{x}_i$ on the box surface. We then inversely compute the corresponding canonical bounding-box coordinate $\mathbf{u} = \operatorname{diag}(\mathbf{s}_i)^{-1}\mathbf{R}_i^{-1}(\mathbf{x}_i - \mathbf{t}_i) \in [-1, 1]^3$ and store it as the RGB color.

Internal map. Using Eq. 4, each node $\mathbf{v}_{ij} \in \mathcal{V}_i$ is projected to a 2D endpoint and drawn with its 3D coordinate $\mathbf{v}_{ij}$ acting as its RGB color. Every edge is then drawn as a 2D line connecting its endpoints, with the 3D coordinate of its geometric midpoint used as the color (Xu et al., 2025).

A.2 Model Architecture

Input sequence. For each instance $i \in \{1, \ldots, N\}$, its local group owns the image and control tokens whose spatial coordinates fall strictly within its mask:

$$x_i = \{x^{(p)} \mid p \in \overline{\mathbf{M}}_i\}, \qquad c_i = \{c^{(p)} \mid p \in \overline{\mathbf{M}}_i\}, \tag{5}$$

where p denotes the spatial position index.

A.3 Training and Inference

Training objective. For a target latent $\mathbf{z}_0$, noise $\boldsymbol{\epsilon}$, and flow time τ, we optimize the flow matching objective:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}\,\|v_\theta(\mathbf{z}_\tau, \tau, y, \mathcal{Q}) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\|_2^2\,, \qquad \mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}. \tag{6}$$

A.4 Optimization Settings

Beyond the summary in Sec. 5, PoseStage is trained for 30k steps with a batch size of 16 and a learning rate of 1×10^{-4}. Dynamic-resolution training caps each image at max_pixels $= 1048576$. The MMDiT backbone is fine-tuned with LoRA of rank 32, and the geometric fusion layer W_f (Eq. 1) is fully fine-tuned.

A.5 Dataset

Real subset for natural diversity. We first refine the bounding-box-derived segmentation masks using SAM 3 (Carion et al., 2026). Next, SAM 3D (Chen et al., 2026a) estimates the 9-DoF camera-relative pose and recovers a local mesh for each masked object. For the internal structure, a category-agnostic geometric graph is extracted from this mesh via voxel remeshing, mean-curvature-flow contraction, and curve simplification (Xu et al., 2025).

A.6 Training Data Statistics and Evaluation Setup

Following ORIGEN (Min et al., 2025) and SceneDesigner (Qin et al., 2025), we exclude categories whose orientation is inherently ambiguous (e.g., front–back symmetric objects) when building the benchmarks, so that the absolute angular error is well defined. The training set contains 13,862 image–condition pairs: 3,522 from the real-image pipeline built on an internal image collection,

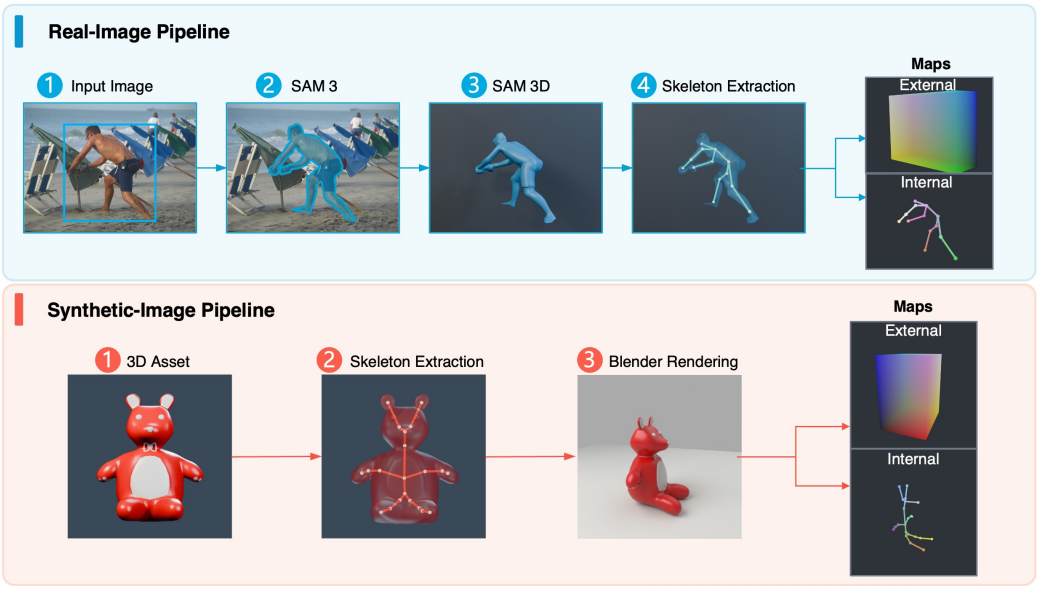

Figure 5: **Data generation pipelines.** The real-image pipeline (top) extracts pose and skeleton annotations from in-the-wild images via off-the-shelf vision models. The synthetic-image pipeline (bottom) samples 3D assets to render high-quality images with exact geometric maps.

and 10,340 from the synthetic pipeline that renders assets from an internal 3D asset collection in Blender. As described in Sec. 4, the real pairs supply natural appearance diversity and the synthetic pairs supply exact geometry, and both are exported in the same format with the external and internal maps.

A.7 Exact Metric Definitions

We evaluate five groups: **Orientation** (Ang.Err and Acc@22.5° on the relative rotation between the estimated and target 9-DoF pose), **Position & Scale** (a normalized 3D center error Trans.Err and a per-axis relative edge-length error Scale.Err, computed after the per-image alignment defined below), **Structure** (a scale-normalized Chamfer distance and a skeleton F1@0.05 on the category-agnostic curves of SKDream (Xu et al., 2025)), **Image quality** (PickScore (Kirstain et al., 2023)), and, for the reference and editing tasks, **Appearance / Non-edited preservation** (CLIP-I and DINO to each reference; PSNR and SSIM outside the edited instances). All methods are scored through the same measurement pipeline. The orientation, position, and structure metrics recover geometry from the generated image with SAM 3 (Carion et al., 2026)/SAM 3D (Chen et al., 2026a), a step that can fail on missing or malformed objects; we therefore report a **Reconstruction rate** (Rec., the fraction of target instances successfully reconstructed) and compute the geometric metrics on those instances.

Orientation. The absolute angular error (Ang.Err, in degrees) of the relative rotation between the estimated and target 9-DoF pose, and Acc@22.5°, the fraction of instances below 22.5°.

Position & Scale. We compare reconstructed object centers and box edge lengths with their targets using the following per-image alignment protocol, applied identically to all methods.

Alignment. (i) Focal-ratio rescaling. Let f_{SAM3D} denote the focal length estimated by the SAM 3D pipeline and f_{GT} the focal length used to render the control maps, both expressed in the same normalized image-coordinate convention. We set $q_f = f_{\mathrm{SAM3D}}/f_{\mathrm{GT}}$ and multiply each reconstructed center's lateral coordinates $(\hat{x}_i, \hat{y}_i)$ and all three box edge lengths by q_f, leaving its center depth $\hat{z}_i$ unchanged. *(ii) Shared scale alignment.* We compute one scale factor per image from the center-depth ratios over successfully reconstructed target instances,

$$g = \exp\Big(\operatorname{median}_i \log \tfrac{z_i}{\hat{z}_i}\Big), \tag{7}$$

where z_i and $\hat{z}_i$ are the target and reconstructed center depths. The same g is applied to every reconstructed instance in the image. The aligned center and box edge lengths are $\hat{\mathbf{c}}'_i = g\,[q_f\hat{x}_i, q_f\hat{y}_i, \hat{z}_i]^\top$ and $\hat{\mathbf{d}}'_i = g\,q_f\,(\hat{\mathbf{b}}_i \odot \hat{\mathbf{s}}_i^{\mathrm{SAM3D}})$, where $\hat{\mathbf{b}}_i$ contains the canonical bounding-box edge lengths of the SAM 3D GLB mesh and $\hat{\mathbf{s}}_i^{\mathrm{SAM3D}}$ is its predicted scale vector. We compare these quantities against the target center $\mathbf{c}_i$ and edge lengths $\mathbf{d}_i$. When only one instance is reconstructed, this alignment matches its center depth exactly.

Trans.Err is the aligned 3D center distance normalized by the target box diagonal, and *Scale.Err* the mean per-axis relative edge-length error,

$$E_{\mathrm{trans},i} = \frac{\|\hat{\mathbf{c}}'_i - \mathbf{c}_i\|_2}{\|\mathbf{d}_i\|_2}, \qquad E_{\mathrm{scale},i} = \frac{1}{3}\sum_{k=1}^{3}\frac{|\hat{d}'_{i,k} - d_{i,k}|}{d_{i,k}}, \tag{8}$$

where the target diagonal $\|\mathbf{d}_i\|_2$ is a ground-truth-only normalizer that makes Trans.Err dimensionless and comparable across object sizes.

Structure. We extract a category-agnostic curve from the reconstruction via the mean-curvature-flow procedure of SKDream (Xu et al., 2025), re-project it to the image plane, sample 256 points per skeleton by arc length, and report a scale-normalized Chamfer distance (Chamfer) (Barrow et al., 1977; Fan et al., 2017) and a skeleton F1 at a 5% threshold (F1@0.05).

Image quality. PickScore (Kirstain et al., 2023). **Appearance preservation** (reference tasks). The CLIP-I and DINO similarity between each generated instance region and its reference.

Non-edited preservation (editing). PSNR and SSIM outside the edited instances.

Binding accuracy (Bind, routing study). The fraction of instances whose region is closer to its own reference than to any other reference in DINO feature space, directly measuring whether each appearance is routed to the correct object.

A.8 USER STUDY

To verify that our automatic metrics faithfully reflect human judgment, we conduct a human study following prior practice (Min et al., 2025; Qin et al., 2025). We sample generated results from the reference-free benchmark. For each sample, annotators are shown the target condition (the rendered external and internal maps) alongside a method's output and are asked to rate how well the output matches the target on four axes separately: orientation, position and scale, internal structure, and overall image quality. We collect ratings from 20 participants. For each axis we average the ratings of each method and normalize them to $[0, 1]$, where a full score is 1.0 and higher is better (Table 7). The normalized human ratings follow the same ranking as our automatic metrics, with PoseStage rated highest on every axis, supporting the automatic pipeline as our primary measure.

Table 7: **User study** Per-axis human rating, averaged over annotators and normalized to $[0, 1]$, where a full score is 1.0 (higher is better). Human ratings match the automatic metrics on every axis.

Method	Orientation↑	Position & Scale↑	Structure↑	Image Quality↑
SceneDesigner (Qin et al., 2025)	0.77	0.81	0.43	0.75
SKDream (Xu et al., 2025)	0.74	0.39	0.80	0.31
InstanceDiffusion (Wang et al., 2024b)	0.45	0.67	0.41	0.57
PoseStage (ours)	**0.88**	**0.93**	**0.91**	**0.82**