Geometry-Preserving Image Generation for 6D Object Pose Estimation

Jiafeng Zhang[1], Rui Song[1], Zhengtai Zhang[1], Jiaojiao Li[1], Kailang Cao[1], Lizhang Peng[2], David Ferstl[3], and Yinlin Hu[3]

[1] State Key Laboratory of Integrated Services Networks, Xidian University
[2] College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[3] Magic Leap

Abstract. We propose an image generation pipeline that preserves geometric consistency for 6D object pose estimation. Most pose estimation methods rely on training with large-scale annotated real datasets, but collecting such data is often difficult and expensive in practice. To address this limitation, we introduce GenerationPose. First, we create a large-scale synthetic dataset by rendering objects from multiple viewpoints using their 3D meshes. We then convert these synthetic images into pseudo-real images using a geometry-conditioned appearance generation model trained on large-scale synthetic–real image pairs. Our experiments show that the model generalizes well to unseen objects and can generate realistic images from synthetic views only. Furthermore, we show that pose estimation models trained on the generated data achieve significantly better performance compared to baseline methods. The code is available at `https://github.com/JiafengZhang-1117/GenerationPose`.

Keywords: Geometry-Conditioned Diffusion Transformer · Synthetic-to-Real Appearance Transfer · 6D Object Pose Estimation

1 Introduction

6D object pose estimation aims to recover the 3D rotation and translation of an object from one or more images, enabling applications such as robotic grasping, augmented reality, and industrial inspection. Despite substantial progress, real-world performance remains constrained by a data bottleneck: collecting diverse real images with accurate 6D pose annotations is costly and difficult to scale [9, 16, 18, 21, 37]. Accurate pose annotation typically requires specialized capture setups and careful calibration, while existing datasets still cover only limited variations in object instances, materials, illumination, and backgrounds [6, 8, 10, 22, 33, 38]. As a result, pose estimators often suffer from a distribution gap between training data and real-world scenes, especially for objects with limited real annotations or unseen instances [3, 5, 16, 17, 21, 24].

Synthetic data rendered from CAD models or point clouds offers a scalable alternative with precise pose labels [9,33]. However, synthetic renders differ from real images in texture, lighting, and background context, creating a reality gap

Fig. 1: Generating pseudo-realistic images from synthetic inputs. The top row shows the synthetic renderings used as inputs. Rows 2 and 3 illustrate different generated results obtained with different network initializations. The generated images are more photorealistic than the synthetic inputs while maintaining the object geometry. Moreover, they introduce greater diversity in background and lighting conditions, which is useful for training robust pose estimators.

that limits generalization when models are trained purely on synthetic data [26, 34, 35]. This motivates a central question: Can geometry renders be translated into visually realistic images while preserving object pose and structure, enabling the low-cost construction of large-scale pose-labeled pseudo-real data for 6D pose estimation?

Prior work addresses the data bottleneck and the synthetic-to-real gap in two main directions. First, systematic benchmarks and datasets improve comparability and drive progress. For example, the BOP benchmark unifies multiple object-level datasets with standardized evaluation protocols and tracks performance across seen and unseen objects through annual challenges [9, 10, 22, 33]. Second, synthetic data is widely used to scale training, but closing the reality gap remains challenging. Domain randomization mitigates overfitting to specific rendering styles by randomizing textures, lighting, and backgrounds, yet its effectiveness depends on the randomization design and often fails to match the complexity of real-world appearance [26, 34, 35].

Recently, diffusion models have demonstrated strong fidelity and controllability in image synthesis [7, 32]. Latent diffusion performs denoising in the latent space of a pre-trained VAE, improving efficiency while maintaining high quality [2, 14, 27, 29]. Diffusion Transformers (DiT) further formulate diffusion modeling with scalable Transformer backbones over latent patches [1, 25]. Moreover, lightweight adapters and conditional branches enable effective structural guidance without sacrificing generation quality [20, 40, 41]. These advances suggest a promising direction for appearance transfer under strong structural constraints.

In this work, we propose GenerationPose, a geometry-conditioned appearance generation framework built on DiT. Given a geometry render x_{geom} from point clouds or CAD models and its object mask m, GenerationPose generates

a pseudo-real image x_{out} that matches real-image appearance while preserving the input pose and geometry, as illustrated in Fig. 1. Our cross-domain protocol trains on uCO3D point-cloud renders and evaluates zero-shot on both uCO3D and BOP CAD/mesh renders, explicitly testing robustness to substantial render-style shifts [19, 25, 28, 29].

GenerationPose is built around three task-critical conditions. **Texture conditioning** performs latent diffusion in the VAE latent space [14, 27, 29], where real images are encoded into latent codes, corrupted with Gaussian noise, patchified, and processed by the DiT backbone for realistic appearance generation. **Geometric conditioning** encodes the geometry render and feeds it into a lightweight ControlNetGeom branch, which produces layer-wise hints and cross-attention tokens following established conditional diffusion designs [20, 41]. **Spatial conditioning** introduces mask-aware self-attention with FG→BG blocking, preventing foreground queries from attending to background keys and optionally strengthening foreground-to-foreground aggregation. Together with diffusion objectives [7, 32], mask-restricted foreground reconstruction, perceptual consistency [13, 31], and background-aware regularization, these conditions improve appearance realism while maintaining pose-critical structure.

Our contributions are summarized as follows: (i) We propose GenerationPose, a geometry-conditioned appearance generation framework that translates geometry renders into pseudo-real images while preserving object pose and structure, enabling low-cost pose-labeled data generation for 6D pose estimation. (ii) We formulate GenerationPose with three complementary conditions: texture conditioning for latent appearance generation, geometric conditioning for pose-aware structural guidance, and spatial conditioning for explicit foreground–background control. (iii) We introduce ControlNetGeom on a DiT backbone to inject geometric priors via layer-wise hints and cross-attention tokens [1, 25, 41], and we propose mask-aware self-attention with FG→BG blocking to suppress background interference and improve structural stability. (iv) We validate the design through zero-shot BOP evaluation, appearance-quality analysis, strength-control analysis, training-and-test component ablations including ControlNetGeom, and downstream 6D pose-estimation experiments.

2 Related Work

6D Pose Benchmarks and the Data Bottleneck. 6D object pose estimation recovers the 3D rotation and translation of rigid objects from images and is crucial for robotic manipulation, augmented reality, and industrial inspection. To improve comparability, the BOP benchmark unifies multiple datasets under a common format and provides standardized evaluation protocols and public challenges [9, 10, 22, 33]. However, collecting real images with accurate 6D pose annotations remains expensive and difficult to scale, often requiring specialized setups and careful calibration [6, 8, 38]. As a result, real datasets have limited coverage of object instances, materials, illumination, and backgrounds, and this gap is particularly evident for unseen objects and scenes [16, 17, 21, 24, 37]. Recent

works explore generalizable and/or promptable pose estimation using stronger representations and foundation-model features [3,5,15,18,23], but data diversity remains a core limiting factor.

Synthetic Data and Sim-to-Real via Domain Randomization. Synthetic renders provide scalable training data with exact pose labels, but they differ from real images in texture, lighting, and background context, leading to the well-known reality gap [34,35]. Domain randomization mitigates this gap by aggressively randomizing textures, lighting, and backgrounds during rendering, aiming to make real images appear as one of many simulated variations [26,34,35]. While effective in some settings, it is sensitive to the randomization design and still struggles to capture the complexity of real-world appearance, especially high-frequency textures and fine details [35]. Recent work has also explored diffusion-enhanced synthetic data for category-level 3D pose estimation [39]. In contrast, our setting targets instance-level BOP-style data generation, where each generated image must preserve the rendered 6D pose label; therefore, pose-defining silhouettes and visible geometry consistency are central constraints rather than optional properties.

Image-to-Image Translation and Synthetic Image Refinement. Image-to-image translation is a classical paradigm for appearance transfer. Paired methods such as pix2pix learn relatively stable mappings when strict input–output pairs are available, using pixel-level and adversarial supervision [12]. Unpaired methods such as CycleGAN relax pairing requirements but may introduce geometric drift or uncontrolled changes, which is undesirable when strict pose/structure preservation is required [42]. SimGAN further frames the task as refinement: making synthetic images more realistic using adversarial supervision on unlabeled real images while regularizing changes to preserve labels [30]. Overall, GAN-based pipelines demonstrate feasibility but often face challenges in training stability, controllability, and high-fidelity detail synthesis under strong geometric/mask constraints [4].

Diffusion Models and Conditional Control: LDM, DiT, and ControlNet. Diffusion models have recently achieved strong image fidelity with stable optimization [7, 32]. Latent Diffusion Models (LDMs) perform diffusion in the latent space of a pre-trained autoencoder, substantially reducing computation while maintaining fine visual details [14, 27, 29]. Diffusion Transformers (DiT) replace U-Net backbones with Transformers and model denoising over latent patches, exhibiting favorable scalability [1, 25]. For spatially grounded control, ControlNet introduces an auxiliary conditioning branch with stable training behavior, enabling robust injection of structural cues such as edges, depth, or segmentation [41]. Recent adapter-based variants further improve controllability with lightweight plugins [20, 40]. Our approach builds on these advances by adopting a latent DiT backbone and introducing task-specific conditioning from geometry renders and object masks for pose/structure-preserving pseudo-real generation.

Perceptual Consistency and Mask-Guided Constraints. Pixel-wise losses alone often yield over-smoothed outputs and fail to capture realistic textures. Perceptual losses compute similarity in feature space, providing stronger constraints on texture and semantic consistency, and are widely used in restoration and translation tasks [13, 31]. In sim-to-real transfer, masks provide a natural prior to emphasize object regions and reduce background interference, and foreground-weighted objectives are commonly adopted [13]. Beyond loss reweighting, masks can also be used to impose spatial constraints on attention to improve structural stability; this becomes especially natural in latent attention-based diffusion backbones [25, 29, 41].

Summary. Existing sim-to-real approaches either rely on heuristic randomization or use GAN-based translation, and they often struggle to simultaneously achieve high-fidelity texture synthesis and strict pose/structure preservation. In contrast, GenerationPose combines geometry conditioning and mask-derived spatial constraints within a latent DiT framework: geometry provides stable structural priors, while mask-aware self-attention with FG→BG blocking improves foreground coherence and reduces background contamination, enabling controllable and pose-consistent pseudo-real data generation for downstream 6D pose estimation.

3 Method

We present GenerationPose, which formulates synthetic-to-real appearance transfer as geometry and mask-conditioned latent diffusion. Given a geometry render x_{geom} and an object mask m, our goal is to generate a visually realistic image $\tilde{x}$ that preserves the input pose and geometric structure while allowing realistic texture, illumination, and background. GenerationPose is built on a Diffusion Transformer (DiT) backbone and integrates three complementary conditions: (i) a texture condition that denoises latent patch tokens, (ii) a geometry condition that injects pose-aware priors via a lightweight ControlNetGeom branch, and (iii) a spatial stream that imposes mask-derived constraints through mask-aware self-attention with FG→BG blocking.

3.1 Problem Formulation and Training Pairs

Pose-aligned training triplets. We assume paired triplets $(x_{\text{real}}, x_{\text{geom}}, m)$ under the same pose, where x_{real} is a real RGB image, x_{geom} is a geometry render, and m is the object mask. The key property is pose-aligned pairing: for each real image x_{real}, we render x_{geom} using the ground-truth object pose and camera intrinsics of x_{real}. This yields a one-to-one correspondence between geometry and appearance, enabling direct supervision that encourages geometry/pose preservation while learning realistic textures and lighting.

Mask construction. We use the object segmentation mask from annotations when available; otherwise, we obtain m from rendering or projected model silhouettes under the same pose. Mask quality directly affects token-level constraints, so we keep m aligned with x_{geom} and x_{real} by construction.

Object-centric preprocessing. To reduce background bias and make supervision focus on geometry/pose, we adopt an object-centric crop protocol. For each triplet, we compute a square bounding box from m, enlarge it by a small factor, crop the same region from x_{real} and x_{geom}, and resize it to 256×256; the cropped mask is resized accordingly. This produces aligned object-centric pairs and makes the geometry constraints explicit. Importantly, the same crop/resize operation is applied to x_{geom} and x_{real}, preserving pixel-wise pose consistency within the crop.

Latent diffusion objective. We perform diffusion in the latent space of a pre-trained VAE with encoder $E(\cdot)$ and decoder $D(\cdot)$. We encode the real image into a latent representation $z_{\text{real}} = E(x_{\text{real}})$ and sample a timestep t with noise $\epsilon \sim \mathcal{N}(0, I)$ to obtain:

$$z_t = \alpha_t z_{\text{real}} + \sigma_t \epsilon. \tag{1}$$

The denoiser is trained to predict ϵ conditioned on (x_{geom}, m) through the geometry and spatial streams described below. At inference, we start from Gaussian noise and iteratively denoise under the same conditions to obtain z_0, then decode $\tilde{x} = D(z_0)$.

3.2 Architecture Overview

GenerationPose uses a DiT backbone operating on latent patch tokens. Given $z_t \in \mathbb{R}^{C\times H\times W}$, we patchify and embed it into token sequences $x \in \mathbb{R}^{N\times D}$. Conditioning enters the DiT through two routes: (i) cross-attention to geometry cond_tokens from ControlNetGeom, and (ii) mask-aware self-attention via an additive attention bias derived from the mask on the token grid. Additionally, ControlNetGeom provides early-layer residual `hints` to strengthen geometric alignment at the beginning of denoising, as shown in Fig. 2.

3.3 Geometry Conditioning: ControlNetGeom

A core challenge is to enhance realism while preserving the pose/geometry implied by x_{geom}. We address this with a lightweight ControlNet-style branch, ControlNetGeom, designed to deliver geometry priors in a token-aligned manner.

Pose-consistent geometry latent. We encode the geometry render using the same VAE encoder:

$$z_{\text{geom}} = E(x_{\text{geom}}), \tag{2}$$

where x_{geom} is rendered under the same pose and intrinsics as x_{real} during training. Thus, geometry features extracted from z_{geom} are pose-consistent with the appearance target.

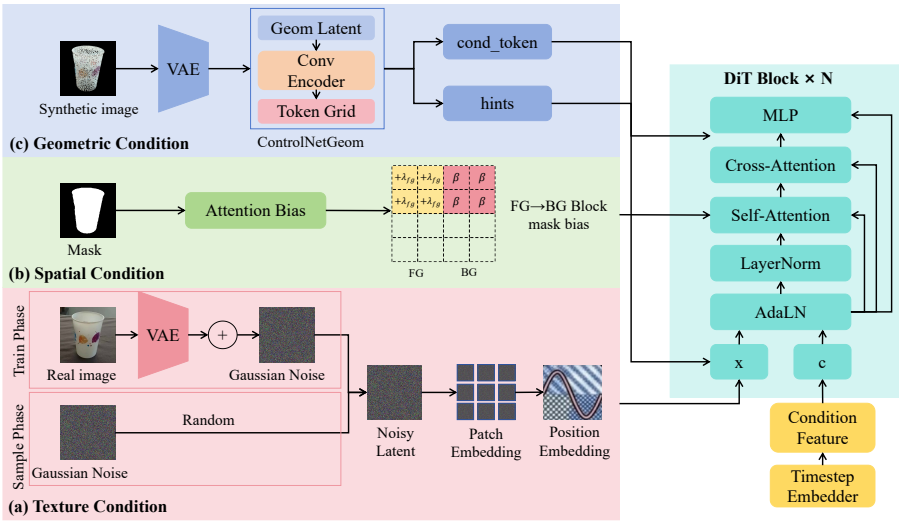

Fig. 2: GenerationPose overview. We perform latent diffusion with three conditions: (i) a texture condition that denoises the noisy latent, (ii) a geometry condition that injects geometry priors via ControlNetGeom (cond_tokens + hints), and (iii) a spatial condition that applies mask-aware self-attention with FG→BG blocking by an additive attention bias on the token grid.

Token-grid alignment. ControlNetGeom applies a small convolutional encoder to z_{geom} to produce a feature map, then interpolates it to the DiT token grid and flattens it into geometry tokens $u \in \mathbb{R}^{B\times N\times D}$. This explicit grid alignment establishes a spatial correspondence between geometry guidance and denoising tokens: each token receives geometry context from the matching location, which is critical for maintaining object structure and pose.

Two geometry injection pathways. ControlNetGeom outputs two types of signals, as shown in Fig. 2:

(i) Cross-attention cond_tokens. We project geometry tokens to conditioning tokens:

$$c_{\text{cond}} = W_{\text{cond}}\, u, \tag{3}$$

and inject them through cross-attention in each DiT block. These cond_tokens provide global geometry context throughout denoising.

(ii) Early-layer residual hints. To enforce structure when the latent is highly noisy, we produce K residual hints:

$$h_i = W_i\, u, \quad i = 0, \ldots, K-1, \tag{4}$$

where W_i are *zero-initialized* for stable training [41]. For the i-th DiT block ($i < K$), we inject:

$$x \leftarrow x + h_i. \tag{5}$$

Later blocks rely only on cross-attention. Intuitively, hints lock in pose/geometry early, while later layers remain flexible to refine texture, illumination, and background. This strong-early/weak-late design is a key contributor to geometry-preserving generation.

Hints in addition to cross-attention. Cross-attention alone provides global geometric context but can be insufficient at high-noise timesteps, where token representations are unstable and structural drift may occur. Our early residual hints act as token-aligned local geometric anchors in the first few blocks, stabilizing the denoising trajectory toward pose-consistent structures. As noise decreases, guidance naturally shifts to cross-attention, allowing the model to refine texture and illumination while maintaining the geometry established by early layers. This complementary design improves geometry preservation without overly constraining appearance diversity.

3.4 Spatial Conditioning: Mask-Aware Self-Attention

We downsample the binary mask to the token grid $\hat{m} \in \{0,1\}^{B\times N}$ and use it to build a self-attention bias that prevents foreground tokens from attending to background tokens. Specifically, we apply FG→BG blocking with a large negative bias:

$$M_{ij} = \begin{cases} \beta, & \hat{m}_i = 1,\ \hat{m}_j = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\beta \ll 0$. The bias is used only in self-attention; cross-attention to geometry `cond_tokens` remains unmasked.

3.5 Conditional DiT Blocks

The backbone consists of L DiT blocks modulated by the timestep embedding $c(t)$ via AdaLN-style shift/scale and gating [25]. Each block contains: (i) self-attention with the mask bias M, (ii) cross-attention over geometry cond_tokens c_{cond}, and (iii) an MLP. The network predicts the diffusion noise.

3.6 Training Objectives and Strategies

Dual-mask design. We use two masks because attention constraints and loss weighting prefer different mask properties. A stable attention mask m_{full} is generated by random shifting and dilation, producing a thicker mask for robust attention bias and cropping. A stricter loss mask m_{loss} is generated by random shifting and erosion, yielding a tighter foreground region for supervision.

Diffusion objective. We optimize the standard diffusion loss with an optional variance term:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\epsilon} + \lambda_{\text{var}}\mathcal{L}_{\text{vb}}. \tag{7}$$

Mask-restricted reconstruction and perceptual losses. Every K_{perc} steps, we decode the predicted $\hat{z}_0$ and compute foreground-weighted losses:

$$\mathcal{L}_{\text{L1-fg}} = \|(\hat{x} - x_{\text{real}}) \odot m_{\text{loss}}\|_1, \tag{8}$$

$$\mathcal{L}_{\text{perc-fg}} = \sum_{\ell} \left\|\phi_\ell(\hat{x} \odot m_{\text{loss}}) - \phi_\ell(x_{\text{real}} \odot m_{\text{loss}})\right\|_1, \tag{9}$$

and optionally a weak background perceptual regularizer $\mathcal{L}_{\text{perc-bg}}$. The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathbb{I}[s \bmod K_{\text{perc}} = 0] \cdot (\lambda_{\text{l1}}\mathcal{L}_{\text{L1-fg}} + \lambda_{\text{fg}}\mathcal{L}_{\text{perc-fg}} + \lambda_{\text{bg}}\mathcal{L}_{\text{perc-bg}}). \tag{10}$$

Robustness: geometry-only degradation and latent alignment. To improve robustness to rendering styles, we apply degradations *only* to x_{geom} while keeping x_{real} unchanged. To reduce train–test mismatch, we encode geometry deterministically at inference using the VAE posterior mean, consistent with how geometry guidance is used at test time.

3.7 Texture Conditioning via DDIM with Style Seeds

Given a geometry render x_{geom} and mask m, we first compute a geometry latent using the VAE posterior mean, $z_{\text{geom}} = E(x_{\text{geom}}).\text{mean}$, and build the conditioning mask m_{full} via dilation to match training. We then perform DDIM sampling [32] with two explicit controls for appearance diversity: (i) a strength parameter that sets the starting timestep t_{start}, and (ii) a style seed that initializes the sampling randomness. For the same (x_{geom}, m), changing the style seed yields diverse pseudo-real outputs with different object textures and illumination, while also producing diverse yet coherent backgrounds.

Strength-controlled sampling. We map a user-specified strength to a start index t_{start} and denoise from t_{start} to 0. A smaller t_{start} keeps the generation closer to the geometry-conditioned structure and produces more conservative appearance changes, whereas a larger t_{start} allows greater freedom in texture, lighting, and background synthesis. This provides a simple trade-off between stability and diversity.

Inference mode. By default, we use an anchor-based DDIM mode: a base sample is first generated with a full DDIM chain, then diffused forward to t_{start} and denoised back to 0. This stabilizes the global structure while still allowing controllable appearance variations through the strength and style seed. We also support a truncate mode that starts directly from random noise at t_{start}, which provides higher diversity but is less stable. Finally, we decode the final latent to obtain the pseudo-real image, $\tilde{x} = D(z_0)$. In practice, we generate multiple candidates per input by sweeping style seeds and optionally strength, exposing both style and background diversity, as shown in Fig. 3.

Fig. 3: Random style-seed diversity. Given the same synthetic input, different style-seed initializations produce pseudo-real images with varied textures and lighting on the object, while also generating diverse and coherent backgrounds.

4 Experiments

We evaluate GenerationPose under a cross-domain, zero-shot protocol and answer three questions: (i) **Pose/structure preservation:** can we translate geometry renders into pseudo-real images while preserving object pose and geometric structure? (ii) **Effect of our DiT design:** do mask-aware attention and inference-time mask dilation improve generation stability and reduce geometric drift? (iii) **Visual realism and downstream utility:** are the pseudo-real images visually plausible and beneficial for 6D pose estimation?

A key feature of our evaluation is cross-dataset and cross-render-style generalization. We train on uCO3D, where the geometry inputs are rendered from point clouds and exhibit characteristic point-cloud artifacts, and evaluate zero-shot on BOP datasets using official CAD/mesh geometry renders and masks. This setting tests robustness to substantial changes in geometry-rendering domains without finetuning.

4.1 Datasets and Cross-Domain Protocol

Training data: uCO3D. We train GenerationPose on uCO3D using pose-aligned triplets of real RGB images, point-cloud geometry renders, and object masks. The geometry inputs are rendered from uCO3D point clouds under the same camera poses as the real images, providing paired supervision for learning appearance transfer while preserving pose.

Zero-shot test data: BOP. We evaluate the trained model zero-shot on several BOP datasets using CAD/mesh-style geometry renders and masks as inputs. This creates a strong rendering-domain shift from sparse point-cloud renders during training to CAD/mesh renders during testing. No finetuning on BOP is used in Sec. 4; dataset abbreviations and preprocessing details are provided in the supplementary material.

4.2 Implementation Details

Architecture. We use a DiT-B/2 latent diffusion backbone at 256×256 resolution. Images are cropped around the object mask before resizing, and the geometry condition is encoded by the same VAE using the posterior mean to reduce train–test mismatch. The model predicts both the diffusion noise and a variance-related term.

Training and inference. We train the model on 4 NVIDIA RTX 4090 GPUs. During training, we use object-centric crops, geometry-only degradation, and the dual-mask strategy described in Sec. 3.6. During sampling, generating one 256×256 image takes about 1.72s on a single RTX 4090.

Ablation protocol. For component ablations, we train a separate model for each variant with the corresponding component removed. All variants use the same data setting, step budget, and evaluation metrics for fair comparison. Additional architectural, optimization, and timing details are provided in the supplementary material.

4.3 Evaluation Metrics

In the zero-shot BOP setting, paired real targets are unavailable for each generated crop, and the backgrounds are intentionally allowed to vary. Therefore, no single metric fully captures the objective. We report three complementary groups of metrics.

Geometry preservation. For geometry preservation, we use contour deviation **NCD** and boundary agreement **BF1@**τ / **Inlier@**τ at $\tau \in \{2, 5\}$ pixels to quantify whether the pose-defining silhouette is maintained.

Appearance realism. For appearance realism, FID and KID are computed as complementary distribution-level metrics. Since the generated images intentionally vary in texture, lighting, and background under fixed pose constraints, these metrics are interpreted together with geometry consistency.

Downstream pose utility. For downstream utility, WDR is trained under different data settings and evaluated using ADI accuracy at normalized object-diameter thresholds.

4.4 Zero-Shot Results on BOP

Pose/structure preservation. Table 1 summarizes pose/structure preservation on TYOL, LM, YCBV, and ICMI using the contour metrics described above. Across datasets, we observe low NCD and high BF1/Inlier, indicating that GenerationPose introduces limited geometric drift and largely preserves the silhouette/boundary dictated by the input geometry renders. TYOL and LM

achieve near-saturated BF1@5, suggesting very strong coarse boundary alignment, while YCBV and ICMI are more challenging, consistent with increased appearance/shape variability. Importantly, even on these harder datasets, BF1@5 remains high, supporting that the generated pseudo-real images retain the pose/structure cues needed for downstream 6D tasks.

Table 1: Pose/structure preservation on BOP. Lower NCD indicates smaller normalized contour deviation between the input geometry render and the generated pseudo-real image, while higher BF1/Inlier indicates better boundary alignment.

Dataset	NCD(%)↓	BF1@2↑	BF1@5↑	Inlier@2↑	Inlier@5↑
TYOL	0.481	0.950	0.998	0.951	0.998
YCBV	1.469	0.838	0.945	0.846	0.951
ICMI	1.090	0.844	0.969	0.846	0.970
LM	0.428	0.960	0.998	0.961	0.998

Qualitative results. Fig. 4 visualizes the mask-aware generation and composition pipeline across domains. Starting from the synthetic render and its foreground mask, our method produces pseudo-real images with more plausible textures and coherent shading while preserving the object pose and structure, consistent with the trend in Table 1. The final column further demonstrates clean object boundaries and seamless blending with the composited background.

4.5 Appearance Realism and Strength Trade-off

FID/KID analysis. Table 2 compares raw geometry renders with our pseudo-real outputs in terms of distribution-level appearance quality. Compared with raw renders, our generated crops reduce FID from 132.84 to 114.57 and KID from 0.0986 to 0.0794, indicating that the generated images move closer to the real-image distribution. Since FID/KID do not measure pose-label correctness, we interpret them together with the boundary-preservation results in Table 1, which evaluate whether pose-critical geometry is maintained.

Strength control. The DDIM strength controls the degree of deviation from the input geometry condition. As shown in Table 3, strength 0.6 achieves the best geometry preservation, with the lowest NCD and the highest BF1@2/Inlier@2. We therefore use strength 0.6 as the default setting.

4.6 Training-and-Test Ablation Studies

To directly evaluate each design choice, we train a separate model for each ablated variant with the corresponding component removed, and evaluate all

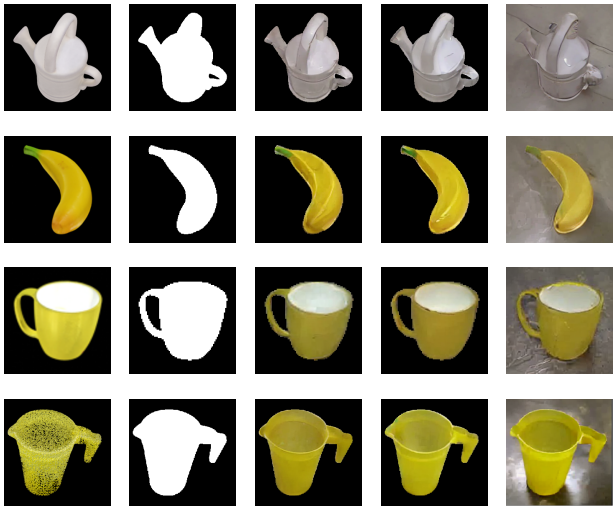

Fig. 4: Cross-domain results. From top to bottom: LM, YCBV, ICMI, and unseen uCO3D. Given a render and its mask, our method preserves pose/shape while generating diverse pseudo-real textures/lighting (S1/S2), and further produces seamless background-composited results with coherent shading (BG).

Table 2: Appearance-quality comparison on BOP. FID/KID are computed against real-image crops and reported as the macro average over LM, ICMI, TYOL, and YCBV.

Input	FID↓	KID↓
Render	132.84	0.0986
Ours	**114.57**	**0.0794**

variants under the same protocol. This avoids the train–inference mismatch that may occur when a component is disabled only at inference time.

Table 4 evaluates the contribution of each component under the default setting. Removing ControlNetGeom causes the largest degradation, increasing NCD from 0.867 to 1.318 and reducing BF1@2 from 0.898 to 0.881. This confirms that explicit geometry conditioning through token-aligned cond_tokens and early residual hints is critical for preserving pose-defining structure.

Mask-aware attention and mask dilation also improve geometry stability. Removing mask-aware attention slightly reduces boundary agreement, while removing dilation leads to worse NCD, BF1@2, and Inlier@2. These results indicate that foreground–background spatial control and a slightly enlarged attention mask help stabilize object-region generation.

Table 3: Strength trade-off on BOP. Results are reported as the macro average over TYOL, YCBV, ICMI, and LM. Increasing sampling strength allows more appearance freedom but may affect pose-critical boundary stability.

Strength	NCD(%)↓	BF1@2↑	Inlier@2↑
0.3	1.085	0.889	0.898
0.6	**0.867**	**0.898**	**0.901**
1.0	1.045	0.896	0.900

Table 4: Training-and-test component ablation on BOP. Results are reported as the macro average over TYOL, YCBV, ICMI, and LM. Each variant is trained and evaluated with the corresponding component removed.

Variant	NCD(%)↓	BF1@2↑	Inlier@2↑
Full	**0.867**	**0.898**	**0.901**
w/o ControlNetGeom	1.318	0.881	0.887
w/o mask-aware attention	0.873	0.892	0.896
w/o dilation ($r = 0$)	0.886	0.895	0.898

4.7 Downstream 6D Pose Estimation with Pseudo-Real Images

To validate the practical utility of our generated images, we use them to train downstream 6D pose estimators. We first evaluate Wide-Depth-Range Pose [11] on BOP LineMOD under five data settings.

Table 5 reports the mean ADI results over thirteen LineMOD objects. Ours alone performs close to Real, achieving 79.87 ADI.10d and 94.09 ADI.20d, while clearly outperforming PBR. When combined with Real, Real+Ours reaches 85.55 ADI.10d and 97.27 ADI.20d, improving more clearly than Real+PBR. These results suggest that our generated data provides more effective complementary supervision than directly adding PBR renders.

Comparison with representative pose estimators. Table 6 further evaluates our generated data on two downstream pose estimators. GenerationPose is not a new pose-estimation architecture, but a pose-label-preserving data generation method. Adding our data improves both WDR and the stronger GDR-Net baseline across multiple BOP datasets, indicating that GenerationPose can benefit different pose-estimation backbones.

5 Conclusion

We presented GenerationPose, a geometry-conditioned pseudo-real image generation framework for 6D object pose estimation. By combining a latent DiT backbone, ControlNetGeom for geometry injection, and mask-aware self-attention for foreground–background control, our method translates geometry renders into realistic images while preserving pose-defining structure and valid 6D pose labels.

Table 5: Mean WDR results on BOP LineMOD. We report mean ADI accuracy over thirteen LineMOD objects under different training data settings.

Setting	ADI.10d↑	ADI.20d↑
PBR	40.96	60.98
Ours	79.87	94.09
Real	83.01	96.43
Real+PBR	<u>83.54</u>	<u>96.84</u>
Real+Ours	**85.55**	**97.27**

Table 6: Contextual comparison of ADI.10d on BOP datasets. GenerationPose provides pseudo-real training data and can be combined with different downstream pose estimators.

Method	LM	LM-O	YCB-V
GDR-Net [36]	91.00	52.90	60.10
GDR-Net + Ours	**92.71**	**56.20**	**78.73**
WDR [11]	83.01	37.90	27.50
WDR + Ours	<u>85.55</u>	<u>44.76</u>	<u>41.40</u>

Experiments under a challenging cross-domain protocol show that GenerationPose maintains strong boundary consistency between input renders and generated outputs, improves appearance realism over raw renders, and provides controllable sampling through the strength parameter.

Downstream experiments demonstrate that the generated pseudo-real images provide useful supervision for 6D pose learning and improve representative pose estimators when used as additional training data.

GenerationPose currently focuses on preserving the visible geometry specified by the input render and mask. It does not explicitly model external occluders or occluder–object interactions, and severe occlusion, inaccurate masks, or unusual lighting may still introduce artifacts or small geometric drift. Future work will explore occlusion-aware generation, stronger realism supervision, and larger-scale integration with downstream pose estimators.

Acknowledgments

This work was supported in part by the Youth Innovation Team of Shaanxi Universities, the "Scientist + Engineer" Team of the Qin Chuangyuan in Shaanxi Province, the Xi'an Science and Technology Program, the "Leading the Charge" initiative for the industrialization of core technologies in key industrial chains in Shaanxi Province, the National Natural Science Foundation of China under Grant No. 62371359, the Key Research and Development Program of Shaanxi, and the Key Research Program of the Chinese Academy of Sciences.

References

1. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International conference on learning representations. vol. 2024, pp. 57611–57640 (2024)
2. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
3. Fan, Z., Pan, P., Wang, P., Jiang, Y., Xu, D., Wang, Z.: Pope: 6-dof promptable pose estimation of any object in any scene with one reference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7771–7781 (2024)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
5. He, X., Sun, J., Wang, Y., Huang, D., Bao, H., Zhou, X.: Onepose++: Keypoint-free one-shot object pose estimation without cad models. Advances in Neural Information Processing Systems **35**, 35103–35115 (2022)
6. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision. pp. 548–562. Springer (2012)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In: 2017 IEEE winter conference on applications of computer vision (WACV). pp. 880–888. IEEE (2017)
9. Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., et al.: Bop: Benchmark for 6d object pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 19–34 (2018)
10. Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J.: Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5610–5619 (2024)
11. Hu, Y., Speierer, S., Jakob, W., Fua, P., Salzmann, M.: Wide-depth-range 6d object pose estimation in space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15870–15879 (2021)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
16. Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: Megapose: 6d pose estimation of novel objects via render & compare. arXiv preprint arXiv:2212.06870 (2022)
17. Lee, J., Cabon, Y., Brégier, R., Yoo, S., Revaud, J.: Mfos: Model-free & one-shot object pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2911–2919 (2024)
18. Lin, J., Liu, L., Lu, D., Jia, K.: Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27906–27916 (2024)
19. Liu, X., Tayal, P., Wang, J., Zarzar, J., Monnier, T., Tertikas, K., Duan, J., Toisoul, A., Zhang, J.Y., Neverova, N., et al.: Uncommon objects in 3d. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14102–14113 (2025)
20. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
21. Nguyen, V.N., Groueix, T., Salzmann, M., Lepetit, V.: Gigapose: Fast and robust novel object pose estimation via one correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9903–9913 (2024)
22. Nguyen, V.N., Tyree, S., Guo, A., Fourmy, M., Gouda, A., Lee, T., Moon, S., Son, H., Ranftl, L., Tremblay, J., et al.: Bop challenge 2024 on model-based and model-free 6d object pose estimation. arXiv preprint arXiv:2504.02812 (2025)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
24. Örnek, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: Foundpose: Unseen object pose estimation with foundation features. In: European Conference on Computer Vision. pp. 163–182. Springer (2024)
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
26. Peng, X.B., Andrychowicz, M., Zaremba, W., Abbeel, P.: Sim-to-real transfer of robotic control with dynamics randomization. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 3803–3810. IEEE (2018)
27. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024)
28. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10901–10911 (2021)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

30. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2107–2116 (2017)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
33. Sundermeyer, M., Hodaň, T., Labbe, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J.: Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2785–2794 (2023)
34. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 23–30. IEEE (2017)
35. Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 969–977 (2018)
36. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16611–16621 (2021)
37. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17868–17879 (2024)
38. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)
39. Yang, J., Ma, W., Wang, A., Yuan, X., Yuille, A., Kortylewski, A.: Robust category-level 3d pose estimation from diffusion-enhanced synthetic data. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3446–3455 (2024)
40. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
41. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)