SGMHand: Structure-Guided Modulation for Structure-Aware Hand Inpainting

Chuancheng Shi[1*], Shiming Guo[1*], Ke Shui[2], Yixiang Chen[1], Fei Shen[3†]

[1]The University of Sydney
[2]Northern Arizona University
[3]National University of Singapore

Abstract

Diffusion models have demonstrated remarkable capabilities in image synthesis, yet realistic hand generation remains a persistent challenge due to complex articulations, self-occlusion, and the lack of explicit structural guidance. To address these issues, we present SGMHand, a novel structure-guided hand inpainting framework that explicitly injects topological priors to enhance structural fidelity and spatial precision. Specifically, we present a structure-guided modulation (SGM) module that synergistically combines structure spatial attention with global feature calibration, enabling fine-grained geometric control over the generative process. Then, we devise a keypoint-aware (KA) loss that enforces topological coherence by aligning attention activations with structures, thereby bridging the gap between high-level semantics and low-level geometry. By jointly optimizing over structural constraints in both representation and learning objectives, SGMHand achieves semantically consistent and geometrically plausible hand synthesis, even under severe occlusion. Extensive experiments demonstrate the effectiveness and strong generalization ability of SGMHand across various foundation models, significantly enhancing the quality and realism of human image synthesis in diverse scenarios.

Introduction

Recent advances in diffusion-based generative models (Jang et al. 2024; Hu et al. 2024; He et al. 2024; Shen et al. 2025c; Jia et al. 2024; Shen et al. 2025e; Shen et al., 2025a; Shi et al. 2025b,a) have enabled high-resolution and semantically coherent image synthesis for tasks such as text-to-image generation and inpainting (Figure 1, first row). However, for fine-grained anatomical structures such as human hands, these models still frequently produce missing or fused fingers and distorted geometry, mainly due to complex articulation, self-occlusion, and weak structural priors (Figure 1, second row). This reveals a key bottleneck: how to enforce anatomically plausible hand synthesis while maintaining overall visual quality. Reliable hand generation is crucial for digital humans and human-computer interaction, and also serves as a practical stress test for structurally grounded and trustworthy generative models.

Figure 1: Hand synthesis remains a key challenge for diffusion models. Top: high-quality results from leading models. Bottom: all models show hand-related failures, such as fused fingers or unnatural poses, revealing difficulty in modeling fine-grained hand structures.

Hand generation (Wang et al. 2025b) remains a persistent challenge in generative modeling due to three compounding factors: (a) The human hand comprises over 16 joints and 27 degrees of freedom, making its structure highly complex and error-sensitive. (b) Frequent self-occlusion, such as in clenched fists or interlaced fingers, leads to occluded critical regions and unstable data distributions. (c) Hands typically occupy a small pixel area, limiting the model's capacity to capture fine details. These constraints, coupled with the absence of explicit skeletal or topological priors, often result in artifacts such as absent fingers, misaligned joints, and unnatural forms. To overcome these limitations, it is crucial to incorporate structured supervision that enforces joint topology and spatial consistency during the generation process.

While conditional control techniques (Shen et al. 2025e; Zhang, Rao, and Agrawala 2023; Shen et al. 2025a; Gao et al. 2025b,a) have advanced face and full-body generation significantly, structured modeling tailored specifically for hands remains underexplored. A few recent works have attempted to incorporate hand priors into the generative process: HandDiffuser (Narasimhaswamy et al. 2024) and HHMR (Li et al. 2024) utilize MANO or SMPL meshes to provide 3D shape constraints, while HandRefiner (Lu et al. 2024), Giving-a-Hand (Pelykh, Sincan, and Bowden 2024), and RHands (Wang et al. 2025c) adopt local inpainting strategies guided by segmentation maps or pose estimations. However, none of these methods explicitly inject structure priors into the diffusion iterations through attention

[*]These authors contributed equally.
[†]Corresponding Author

modulation or joint-level supervision. A unified framework that systematically encodes structure or topological knowledge, while coordinating feature generation across spatial and channel dimensions, remains an open challenge.

To address the aforementioned limitations, we propose SGMHand, a structure-guided framework designed to enhance both spatial accuracy and geometric plausibility in hand generation. Central to our approach is the structure-guided modulation (SGM) module, which integrates structure priors estimations into a local inpainting network. This module modulates spatial attention across multiple receptive fields while preserving contextual coherence via global feature calibration. In parallel, we introduce a keypoint-aware (KA) loss that enforces alignment between attention peaks and structure during optimization, serving as a fine-grained structural constraint. Together, the SGM module and the KA loss form a dual-branch structural guidance mechanism, injecting explicit priors into both the representation and supervision stages. Without relying on 3D meshes or manual masks, SGMHand enables semantically consistent restoration of severely occluded or distorted hand regions. Moreover, our framework is model-agnostic and can be seamlessly integrated with various base T2I diffusion models, demonstrating strong generalization across different generation scenarios. We highlight the following contributions:

- We propose SGMHand, a structure-guided hand inpainting framework that explicitly incorporates structural priors to address the challenges of self-articulation and structural ambiguity in hand region generation.
- We design a structure-guided modulation (SGM) module that effectively combines spatial attention and global channel-wise calibration, enabling semantically consistent and detail-preserving reconstruction.
- We present a keypoint-aware (KA) loss that explicitly aligns the model's attention with the structure topology, effectively guiding the network to focus on anatomically meaningful regions under occlusion.

Related Work

Human Image Generation. Diffusion models have significantly advanced human image generation (Shen et al. 2024; Shen and Tang 2024; Shen et al. 2025d,b; Ma et al. 2024, 2025c,d,a,b), enabling photorealistic and semantically coherent full-body synthesis with strong backbones such as Stable Diffusion (Rombach et al. 2022; Esser et al. 2024), FLUX (Batifol et al. 2025), and Gemini (Comanici et al. 2025). However, fine-grained anatomical regions, especially hands, remain a common failure case, often suffering from missing/fused fingers and topological inconsistencies. To enhance structural fidelity, prior works inject external guidance into frozen diffusion backbones, e.g., ControlNet (Zhang, Rao, and Agrawala 2023) with edge/depth conditions and T2I-Adapter (Mou et al. 2024) with lightweight adapters for segmentation/keypoints. While effective for global layout and silhouette, these controls are typically too coarse to constrain finger-level topology, leaving hand deformations largely unresolved.

Image Inpainting. Early image restoration approaches such as LaMa (Suvorov et al. 2022) and RePaint (Lugmayr et al. 2022) leverage frequency-aware learning and iterative denoising to recover high-resolution textures and semantics, achieving impressive results in general inpainting tasks. However, when applied to fine-grained regions like hands, these models often produce artifacts such as finger merging, bone erasure, or anatomically implausible bending. These failures stem from the fact that fingers occupy a small spatial area and exhibit complex articulations, making them particularly challenging to restore. To address these challenges, recent methods have incorporated hand-specific structural priors. HandRefiner (Lu et al. 2024) renders depth maps from reconstructed 3D hand meshes and feeds them into a ControlNet-style refinement head. Giving-a-Hand (Pelykh, Sincan, and Bowden 2024) performs hand cropping and skeleton-guided synthesis using a conditional ControlNet. RHanDS (Wang et al. 2025c) separates hand structure and style via dual encoders and introduces structure-aware optimization during diffusion. Despite notable progress, these methods still face key limitations. Coarse segmentation masks fail to capture fine finger contours, and 3D mesh pipelines are prone to errors, relying on accurate pose estimation. Crucially, none explicitly reweight attention toward finger-specific regions, leading to persistent issues such as finger fusion, misalignment, and joint artifacts.

Methods

Overall Framework

Our framework aims to reconstruct occluded hand regions from masked RGB images by leveraging structural priors derived from depth and skeleton maps. We adopt an encoder–decoder architecture based on a frozen U-Net, where the masked image is passed through a fixed generative path. As illustrated in Figure 2 (a), to inject structural awareness, we propose a trainable structure-guided modulation (SGM) module into the encoder. The SGM module comprises two key components: a structure-aware attention mechanism that enhances spatial sensitivity using structural cues, and a feature modulator scheme that reweights structure features according to the global structural context. These modulated features are integrated into the U-Net bottleneck, providing structural guidance for the decoder to produce anatomically plausible and semantically coherent hand completions. As shown in Figure 2 (b), during inference, the masked image flows through the frozen U-Net, guided by the pretrained SGM module. Structural cues are encoded, modulated, and injected at the bottleneck, enabling high-fidelity restoration of the occluded hand region.

Structure-Guided Modulation Module

The proposed SGM module is designed to inject structural information into the generative process by modulating depth features with skeletal priors. It consists of two complementary submodules: (1) a structure-aware attention mechanism that enhances spatial sensitivity by attending to local structural cues, and (2) a feature modulator, which adaptively

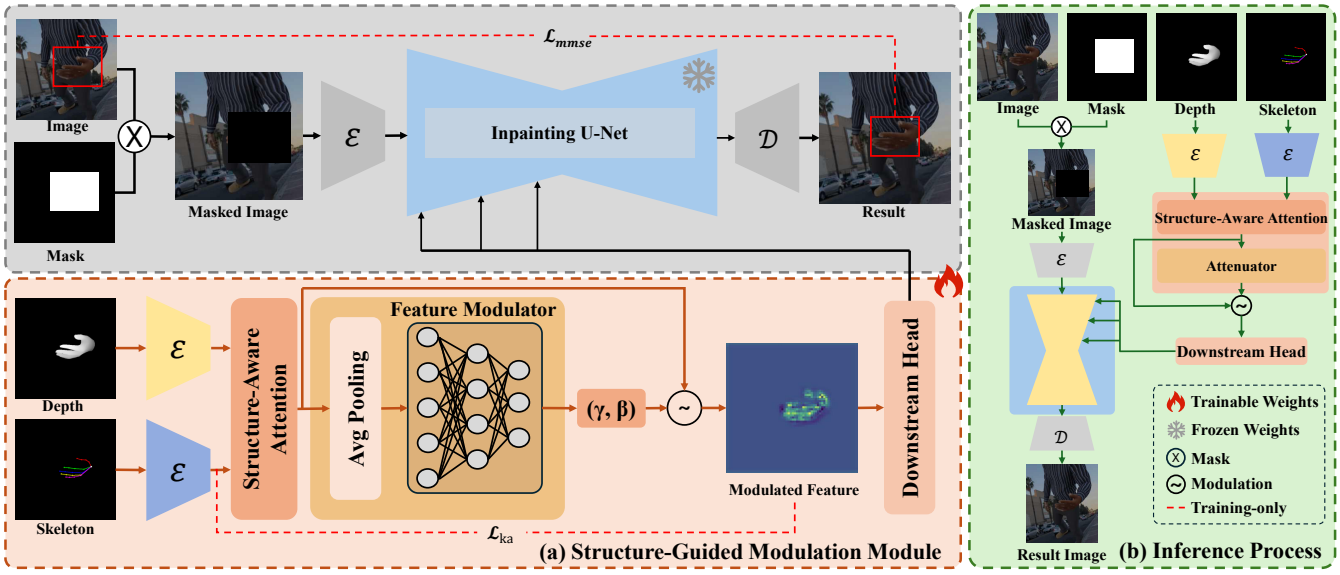

Figure 2: Overview of the proposed structure-guided hand inpainting framework. (a) During training, the structure-guided modulation (SGM) module encodes structure information to generate modulation parameters γ and β, which guide the generative process with fine-grained structural awareness. Applying the keypoint-aware (KA) loss $\mathcal{L}_{\text{ka}}$ encourages the model to align attention activations with structure, thereby enhancing topological coherence. (b) At inference time, the trained SGM module provides structural guidance for the U-Net to restore occluded hand regions.

reweights feature channels based on global structural context. Together, these submodules perform joint spatial and channel-level modulation, enabling topology-aware refinement of features. Unlike existing approaches, such as HandRefiner, which rely on coarse masks, or RealisHuman, which derives attention solely from RGB features, our SGM module directly incorporates explicit structure topology as a structural prior. This design facilitates precise semantic localization and improves the reconstruction of fine-grained hand structures, especially under occlusion or distortion.

Structure-Aware Attention. To enhance the spatial alignment between depth and skeleton features, we introduce a cross-modal attention mechanism that explicitly integrates structure topology into the depth representation. This encourages the network to focus on semantically meaningful regions during reconstruction. Let $\mathbf{F}_D \in \mathbb{R}^{B\times C_D\times H\times W}$ and $\mathbf{F}_S \in \mathbb{R}^{B\times C_S\times H\times W}$ denote the intermediate features from depth and skeleton maps, respectively. We first project them into a shared token space $\mathbf{T}_D \in \mathbb{R}^{B\times N\times d}$ and $\mathbf{T}_S \in \mathbb{R}^{B\times N\times d}$:

$$\mathbf{T}_D = \phi_D(\mathbf{F}_D), \tag{1}$$
$$\mathbf{T}_S = \phi_S(\mathbf{F}_S), \tag{2}$$

where $N = H \times W$, with H and W are representing the spatial height and width, B is the batch size. C_D and C_S respectively denote the channel of depth and skeleton, and $d = C_D$ is the token embedding dimension. The projections $\phi_D(\cdot)$ and $\phi_S(\cdot)$ (e.g., 1×1 convolutions) ensure semantic alignment across modalities. We then compute a relevance matrix $\mathbf{R} \in \mathbb{R}^{B\times N\times N}$ via scaled dot-product attention:

$$\mathbf{R} = \psi\left(\frac{\mathbf{T}_D, \mathbf{T}_S^\top}{\sqrt{d}}\right), \tag{3}$$

where $\psi(\cdot)$ denotes the row-wise softmax. This matrix encodes how each depth token attends to skeletal tokens. Using $\mathbf{R}$, we aggregate structure-guided information $\mathbf{T}_D^{\text{out}} \in \mathbb{R}^{B\times N\times d}$:

$$\mathbf{T}_D^{\text{out}} = \mathbf{R}, \mathbf{T}_S, \tag{4}$$

which is then reshaped and added to the original depth features via a residual connection:

$$\tilde{\mathbf{F}}_D = \text{LayerNorm}\left(\mathbf{F}_D + \text{Reshape}(\mathbf{T}_D^{\text{out}})\right). \tag{5}$$

The result $\tilde{\mathbf{F}}_D$ is a structure-aware depth feature map that integrates fine-grained pose guidance.

Feature Modulator. While the attention mechanism ensures local spatial alignment, it lacks global calibration across channels. To address this, we introduce a lightweight feature modulator that adaptively adjusts depth channels based on global structure context.

We first apply global average pooling (GAP) to compress spatial information from the skeleton features $\mathbf{z} \in \mathbb{R}^{B\times C_S}$:

$$\mathbf{z} = \text{GAP}(\mathbf{F}_S). \tag{6}$$

This vector encodes overall structure. Two fully connected layers then generate affine transformation parameters $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{B\times C_D}$:

$$\boldsymbol{\gamma} = \sigma(\mathbf{z}\mathbf{W}_\gamma + \mathbf{b}_\gamma), \tag{7}$$
$$\boldsymbol{\beta} = \mathbf{z}\mathbf{W}_\beta + \mathbf{b}_\beta, \tag{8}$$

Method	FID↓	KID↓	MPJPE↓
SD 1.5 Inpainting+Skeleton	26.42	0.0296	19.755
SD 1.5 Inpainting+Depth	24.19	0.0271	20.972
HandRefiner	18.73	0.0214	17.722
RealisHuman	18.47	0.0207	19.614
SGMHand (Ours)	**16.83**	**0.0187**	**15.185**

Table 1: Quantitative comparisons with SOTA methods on HaGRID dataset. SGMHand achieves the best performance.

Method	FID↓	KID↓	MPJPE↓
SD 1.5 Inpainting+Skeleton	31.51	0.0337	27.483
SD 1.5 Inpainting+Depth	30.62	0.0329	28.944
HandRefiner	26.22	0.0299	23.752
RealisHuman	24.37	0.0267	19.941
SGMHand (Ours)	**22.64**	**0.0231**	**16.185**

Table 2: Quantitative comparisons on FreiHAND shows SGMHand achieves top performance.

where $\sigma(\cdot)$ is a sigmoid activation function, and $\mathbf{W}_\gamma, \mathbf{W}_\beta$ and $\mathbf{b}_\gamma, \mathbf{b}_\beta$ are learnable parameters. The scaling vector $\boldsymbol{\gamma}$ serves as a soft gating mechanism to suppress irrelevant channels, while $\boldsymbol{\beta}$ introduces flexible offsets. The modulated depth features $\hat{\mathbf{F}}_D$ are computed as:

$$\hat{\mathbf{F}}_D = \boldsymbol{\gamma} \cdot \tilde{\mathbf{F}}_D + \boldsymbol{\beta}, \tag{9}$$

with both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ broadcast along spatial dimensions. This modulation process selectively enhances structurally relevant features while suppressing spatially incoherent or noisy channels, resulting in globally consistent and structure-aware depth representations.

Structure-Content Supervision Enhancement

To further improve the structural fidelity of the generated hand regions, we introduce a dual-branch supervision strategy that reinforces the alignment between learned representations and the structure's topology. This strategy comprises: (1) an explicit attention-level constraint to align focus regions with hand structure, and (2) a pixel-level masked reconstruction loss to guide detail restoration.

Keypoint-Aware Loss. While the structure-aware attention mechanism enables implicit interaction between structure features, it lacks explicit supervision to ensure that the attention consistently focuses on anatomically important regions. Without such structural guidance, the attention distribution may drift over unrelevant areas, leading to degraded structure fidelity or anatomically implausible completions.

To address this, we propose a keypoint-aware (KA) loss that explicitly anchors the model's internal attention to ground-truth anatomical structure positions. The core idea is to enforce alignment the network's attention map with a 2D pose-derived mask, thereby ensuring that the attention mechanism captures the spatial layout of hand joints. Given an attention map $A \in \mathbb{R}^{B\times H\times W}$ predicted by the network, we construct a binary mask $M^{\text{pose}} \in \mathbb{R}^{B\times H\times W}$ centered on detected 2D keypoints, resized to match the resolution of A. The KA loss $\mathcal{L}_{\text{ka}}$ is defined as:

$$\mathcal{L}_{\text{ka}} = \frac{1}{B}\sum_{i=1}^{B} \left\| A_i - M_i^{\text{pose}} \right\|^2. \tag{10}$$

This term encourages attention activations to correlate more closely with ground-truth joint locations, thereby enhancing topological consistency between the learned feature distribution and underlying the physical hand structure.

MMSE Loss. In addition to structural alignment, accurate reconstruction of occluded content remains essential. To this end, we adopt a masked mean squared error (MMSE) loss, which is computed exclusively over the occluded region to emphasize reconstruction of missing content. Let $\hat{\epsilon}_p$ and ϵ_p denote the predicted and ground-truth noises at pixel p. Given mask $M_p \in \{0,1\}$ and $|M| = \sum_p M_p$, which is the total number of pixels in the occluded region.

$$\mathcal{L}_{\text{mmse}} = \frac{1}{|M|}\sum_{p} M_p \left|\hat{\epsilon}_p - \epsilon_p\right|^2. \tag{11}$$

This formulation ensures that training predominantly focuses only on the missing or corrupted content, preventing the model from overfitting to unmasked regions.

Total Objective. The overall training objective combines both structural and pixel-level losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{attn}}\mathcal{L}_{\text{ka}} + \mathcal{L}_{\text{mmse}}. \tag{12}$$

Here, λ_{attn} balances the contribution of structural supervision. Empirically, we find that incorporating $\mathcal{L}_{\text{ka}}$ significantly improves the anatomical plausibility of generated hands, particularly under heavy occlusion.

Experiments And Analysis

Implementation Details

Dataset. Following HandRefiner (Lu et al. 2024), we adopt the same data processing pipeline and splitting strategy. Specifically, we resize all hand crops to a fixed resolution of 512×512, reconstruct a 3D hand mesh using MeshGraphormer (Lin, Wang, and Liu 2021), and project it into depth maps (together with the corresponding masks) under the same camera/model assumptions as HandRefiner. We further extract 2D skeletal keypoints with MediaPipe (Lugaresi et al. 2019) to produce the skeleton maps. Using this pipeline, we curate approximately 40,000 training samples from RHD (Zimmermann and Brox 2017). For evaluation, we randomly sample 12,000 real-world images from HaGRID (Kapitanov et al. 2024) and 10,000 images from FreiHAND (Zimmermann et al. 2019), applying the identical preprocessing steps to ensure domain consistency.

Metrics. Following HandRefiner (Lu et al. 2024), we adopt the same evaluation protocol to ensure fair comparison. We use three standard quantitative metrics: FID (Heusel et al. 2017) to measure distributional differences between real and generated features, KID (Bińkowski et al. 2018) as a sample-efficient alternative, and MPJPE (Ionescu et al. 2013) to assess 3D pose accuracy via joint error. To complement these, we conduct a user study incorporating three subjective metrics: R2G (real images misclassified as generated),

Figure 3: Qualitative comparison of hand generation results. Our method produces more anatomically accurate and coherent results, especially in challenging poses with occlusion or close finger interaction. Please zoom in for more details.

G2R (generated images perceived as real), and Jab (cases where real and generated images are indistinguishable). Collectively, these metrics provide a balanced evaluation of both structural fidelity and perceptual realism.

Hyperparameters. All experiments are conducted using PyTorch on a single NVIDIA A6000 GPU. We adopt Stable Diffusion 1.5 Inpainting (Rombach et al. 2022) as the backbone and freeze its weights during training. The model is trained for 10,000 steps on synthetic data using AdamW (Loshchilov and Hutter 2017) with a batch size of 8 and a learning rate of 5×10^{-5}. The keypoint-aware loss is weighted by $\lambda_{\text{attn}} = 0.3$ to balance structural alignment and reconstruction fidelity. All experiments are conducted on the same hardware setup and GPU environment to ensure a fair comparison across all methods.

Compare with SOTA Methods

We quantitatively compare the proposed SGMHand with several SOTA methods, including Stable Diffusion 1.5 Inpainting + Skeleton (SD 1.5 + Skeleton), Stable Diffusion 1.5 Inpainting + Depth (SD 1.5 + Depth), HandRefiner (Lu et al. 2024), and RealisHuman (Wang et al. 2025a).

Quantitative Results. As shown in Table 1, SGMHand consistently outperforms all state-of-the-art methods across FID, KID, and MPJPE, demonstrating superior perceptual quality and structural accuracy. Specifically, it achieves the lowest FID of 16.83, surpassing RealisHuman (18.47) by 1.64 points and the skeleton-guided baseline (26.42) by over 9.59 points. The KID drops to 0.0187 compared to 0.0207 from RealisHuman, indicating improved patch-level fidelity. For 3D accuracy, SGMHand achieves the lowest MPJPE of 15.185, reducing joint error by 4.429 compared to RealisHuman and by more than 4.07 against SD 1.5 + Skeleton (19.755). These improvements validate the effectiveness of our structure-guided modulation (SGM) module, which injects structure priors through structure-aware attention and feature modulation to enable precise and coherent hand reconstruction under complex poses and occlusions. To further assess generalization, we evaluate our model on the FreiHAND dataset, which provides hand-centric annotations for fine-grained analysis (Table 2). SGMHand again achieves the best results across all metrics. It obtains an FID of 22.64, outperforming RealisHuman (24.37), HandRefiner (26.22), and SD 1.5 Inpainting + Skeleton (31.51). On KID, it achieves 0.0231, significantly lower than RealisHuman (0.0267) and SD 1.5 Inpainting + Skeleton (0.0337). In terms of MPJPE, SGMHand achieves 16.185, reducing the error by 3.76 compared to RealisHuman (19.941) and by more than 7.5 compared to HandRefiner (23.752). These results confirm the robustness of our framework in generating anatomically accurate and visually realistic hands, especially under occlusion and pose variation.

Qualitative Results. As illustrated in Figure 3, SGMHand delivers superior results in both joint continuity and contour coherence across diverse hand poses. Compared to the SD 1.5 + Depth (first row) lacks semantic priors and often produces errors in finger count or layout. While SD 1.5 + Skeleton (second row) introduces structural guidance, it lacks shape-level cues, leading to distorted geometry and confusion in overlapping fingers. HandRefiner (third row) generates smooth textures but frequently suffers from fused fingers and deformed outlines due to weak structural encoding. RealisHuman (fourth row) improves global contours yet struggles with disproportionate fingers and broken joints,

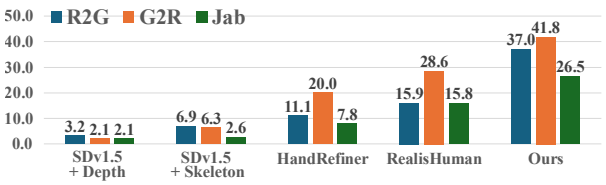

Figure 4: User study results evaluated using R2G, G2R, and Jab metrics, where higher scores indicate greater perceived realism and user preference.

SGM	KA loss	FID ↓	KID ↓	MPJPE ↓
✓	✗	18.37	0.0212	19.724
✗	✓	21.74	0.0254	18.991
✓	✓	**16.83**	**0.0187**	**15.185**

Table 3: Ablation results on proposed SGMHand. Lower FID, KID, and MPJPE are better.

especially under extreme poses. In contrast, our method (fifth row) consistently preserves anatomical plausibility and fine-grained structure. These improvements are attributed to the proposed SGM module, which combines structure-aware attention and feature modulation to effectively incorporate structure priors. Unlike naive condition concatenation, SGMHand explicitly models spatial structure, enabling coherent reconstructions even under occlusion or complex articulation. Additionally, the keypoint-aware loss encourages the network to focus on joint-relevant regions, further enhancing structural fidelity and pose alignment.

User Study. While the quantitative and qualitative results highlight the effectiveness of SGMHand, hand generation is ultimately evaluated by human perception; therefore, we also conducted subjective testing involving 30 participants. To assess perceptual quality from a human-centric perspective, we conducted a user study on the HaGRID dataset using three well-established evaluation metrics: R2G (Real-to-Generated), G2R (Generated-to-Real), and Jab (user preference). Higher values on all metrics indicate stronger visual realism and user favorability. We compared SGMHand with SD 1.5 + Skeleton, SD 1.5 + Depth, HandRefiner, and RealisHuman. As shown in Figure 4, SGMHand outperforms all baselines across all three metrics. In the G2R setting, 41.8% of SGMHand's outputs were judged as real, surpassing the second-best method by 13.2%. On user preference (Jab), SGMHand achieved 26.5%, indicating consistent favorability among participants. These results further validate the perceptual advantages of our structure-guided generation framework, particularly in scenarios requiring fine-grained structural accuracy and visual plausibility.

Ablation Studies

Effectiveness of the SGM Module. To assess the impact of the proposed SGM module, we conduct ablation experiments as shown in Table 3. Specifically, we compare model performance with and without the SGM module while keeping the keypoint-aware loss unchanged. Across all three metrics, adding the SGM module yields notable improvements, demonstrating its effectiveness in guiding feature representations toward structurally consistent regions via structure-aware attention and feature modulation. This leads to more accurate reconstructions and higher-quality image generation. To further disentangle its components, we perform a component-wise ablation by separately enabling the structure-aware attention and the feature modulator (Table 4). Each submodule contributes complementary benefits: structure-aware attention improves MPJPE to 16.578 by enhancing spatial alignment with structural priors. At the same time, the feature modulator alone reduces KID to 0.0199 by strengthening global channel relevance. Their combination yields the best overall performance, confirming that both spatial and semantic guidance are critical to achieving precise and coherent hand synthesis.

SA	FM	FID ↓	KID ↓	MPJPE ↓
✗	✓	18.14	0.0199	17.426
✓	✗	17.81	0.0193	16.578
✓	✓	**16.83**	**0.0187**	**15.185**

Table 4: Ablation study on the SGM module.

Effectiveness of the Keypoint-Aware Loss. To evaluate the contribution of the keypoint-aware (KA) loss, we conduct a controlled experiment by fixing the SGM module and comparing results with and without KA loss (Table 3). The results demonstrate consistent improvements across both perceptual quality and structural accuracy, indicating that KA loss provides effective structural supervision. By explicitly aligning attention with ground-truth structural cues, the model achieves higher anatomical fidelity without compromising visual realism, particularly under complex hand poses. To further illustrate its effect, we visualize the attention heatmaps under both settings in Figure 5. Without KA loss, attention is diffuse and lacks focus on key structural regions. In contrast, with KA loss, attention becomes significantly concentrated around anatomically important areas such as the palm center and finger joints. This observation confirms that KA loss enhances the model's ability to attend to structurally relevant regions, thereby improving topological consistency and reconstruction accuracy.

More Results

Cross-Domain Generalization. To assess the generalization ability of our method beyond the training distribution, we qualitatively evaluate its performance on images from diverse, unconstrained environments not seen during training, as shown in Figure 6. Looking at the first row from left to right: the first column shows an incorrect number of fingers in both hands, the second column generates a confusing structure in two-handed interactions; the third column generates a blurred and confusing structure due to the small size of the target; the fourth column again shows an anomaly in the number of individual fingers; and the fifth column shows an incorrect structure of the fingers in hand-object interactions, with a missing number of fingers. Despite these difficulties, SGMHand consistently reconstructs hand regions with high structural fidelity and visual realism. These results highlight

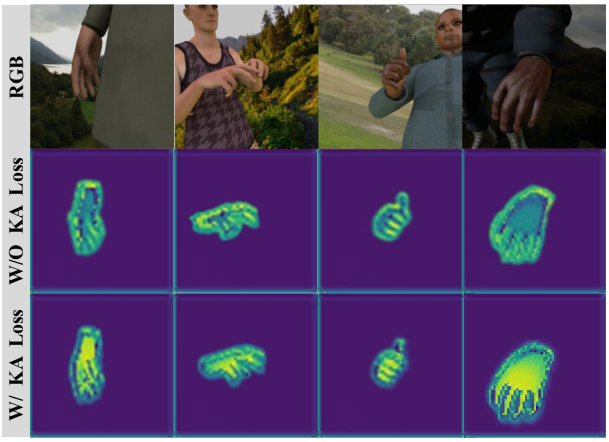

Figure 5: Visualization of the keypoint-aware (KA) loss ablation. The first row shows the original RGB inputs. The second and third rows visualize attention heatmaps without and with the KA loss, respectively. Brighter regions indicate stronger attention responses. Adding KA loss clearly enhances structural focus around hand joints.

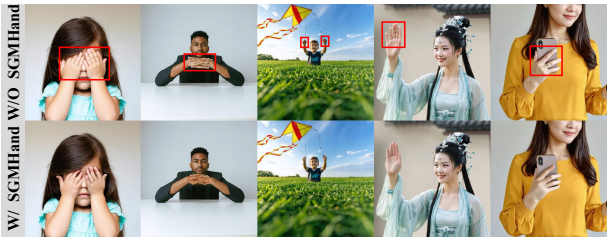

Figure 6: Cross-domain generalization results. Top row: input images from out-of-distribution domains. Bottom row: outputs from SGMHand, demonstrating strong generalization ability beyond the training domain.

the strong cross-domain generalization capability of our approach in the presence of complex interactions, severe occlusion, scale variation, and cluttered backgrounds.

Plug-and-Play Compatibility across Diffusion Models. To evaluate the model-agnostic applicability of SGM, we integrate it into diverse diffusion generators and report qualitative results in Figure 7. Specifically, we test six representative models with different architectures and scales, including Stable Diffusion 1.5 (Rombach et al. 2022), Stable Diffusion 3.5 (Esser et al. 2024), Stable Diffusion XL (Podell et al. 2023), FLUX.1 Dev (Batifol et al. 2025), FLUX.1 Kontext Pro (Batifol et al. 2025), and Gemini 2.5 Flash (Comanici et al. 2025). Across these backbones, the original generations often exhibit typical hand failures (e.g., distorted shapes, missing/extra fingers, and structural collapse under occlusion). After adding SGM, hand topology becomes noticeably more accurate and visually consistent, especially for occluded, truncated, and complex poses, demonstrating strong plug-and-play generality across diffusion frameworks.

Hyperparameter Analysis. We investigate the impact of

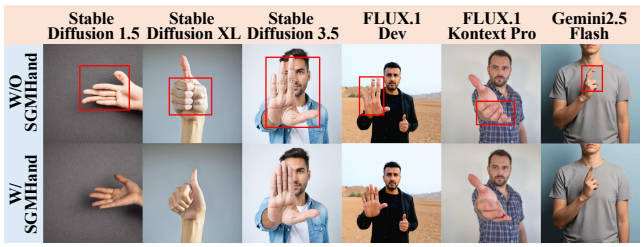

Figure 7: Plug-and-play compatibility across diffusion models. Top: Original generations with distortions, missing parts, or unnatural poses. Bottom: SGMHand results with improved anatomical completeness and realism.

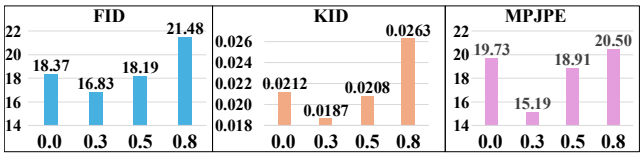

Figure 8: Hyperparameter results. Sensitivity of FID, KID, and MPJPE to the λ_{attn} hyperparameter in SGMHand. Lower FID, KID, and MPJPE are better.

the attention supervision weight λ_{attn} in Eq. 12 through a systematic ablation study. We observe that moderate structural supervision yields the best performance, as shown in Figure 8. Specifically, setting $\lambda_{\text{attn}} = 0.3$ achieves the lowest FID score of 16.83. In contrast, removing the KA loss entirely ($\lambda = 0.0$) or applying excessive weight ($\lambda = 0.8$) increases the FID to 18.37 and 21.48, respectively. Similarly, KID and MPJPE also achieve the best performance when $\lambda_{\text{attn}} = 0.3$. This indicates that a balanced supervisory signal effectively guides the model to attend to semantically meaningful structure regions, while overly strong regularization may restrict attention flexibility and degrade photorealism. Overall, the SGM module reaches optimal performance when configured with $\lambda_{\text{attn}} = 0.3$, confirming the robustness and stability of these architectural choices.

Conclusion

In this work, we propose SGMHand, a structure-guided hand inpainting framework that addresses the long-standing challenges of hand generation in diffusion models. By integrating a structure-guided modulation (SGM) module and a keypoint-aware (KA) loss, SGMHand explicitly incorporates structural priors to enhance both structural fidelity and semantic consistency. The SGM module enables spatially-aware feature modulation guided by hand topology, while the KA loss reinforces anatomical alignment through attention supervision. Extensive experiments on multiple datasets demonstrate that SGMHand significantly improves hand reconstruction quality and generalizes well across diverse foundation models. Our framework provides a plug-and-play solution for hand synthesis, paving the way for more reliable human-centric generation in real-world applications.

References

Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv e-prints*, arXiv–2506.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Gao, J.; Li, J.; Liu, W.; Zeng, Y.; Shen, F.; Chen, K.; Sun, Y.; and Zhao, C. 2025a. CharacterShot: Controllable and Consistent 4D Character Animation. *arXiv preprint arXiv:2508.07409*.

Gao, J.; Sun, Y.; Shen, F.; Jiang, X.; Xing, Z.; Chen, K.; and Zhao, C. 2025b. Faceshot: Bring any character into life. *arXiv preprint arXiv:2503.00740*.

He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; and Xie, L. 2024. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 8472–8480.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hu, T.; Zhang, J.; Yi, R.; Du, Y.; Chen, X.; Liu, L.; Wang, Y.; and Wang, C. 2024. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 8526–8534.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Jang, J.; Youn, C.-H.; Jeon, M.; and Lee, C. 2024. Rethinking Peculiar Images by Diffusion Models: Revealing Local Minima's Role. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2454–2461.

Jia, C.; Luo, M.; Dang, Z.; Dai, G.; Chang, X.; Wang, M.; and Wang, J. 2024. Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2480–2488.

Kapitanov, A.; Kvanchiani, K.; Nagaev, A.; Kraynov, R.; and Makhliarchuk, A. 2024. HaGRID–HAnd Gesture Recognition Image Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4572–4581.

Li, M.; Zhang, H.; Zhang, Y.; Shao, R.; Yu, T.; and Liu, Y. 2024. Hhmr: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 645–654.

Lin, K.; Wang, L.; and Liu, Z. 2021. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12939–12948.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, W.; Xu, Y.; Zhang, J.; Wang, C.; and Tao, D. 2024. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7085–7093.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Ma, Y.; Feng, K.; Hu, Z.; Wang, X.; Wang, Y.; Zheng, M.; He, X.; Zhu, C.; Liu, H.; He, Y.; et al. 2025a. Controllable Video Generation: A Survey. *arXiv preprint arXiv:2507.16869*.

Ma, Y.; Feng, K.; Zhang, X.; Liu, H.; Zhang, D. J.; Xing, J.; Zhang, Y.; Yang, A.; Wang, Z.; and Chen, Q. 2025b. Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. *arXiv preprint arXiv:2506.04590*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Shen, L.; Qi, C.; Ying, J.; Cai, C.; Li, Z.; Shum, H.-Y.; et al. 2025c. Follow-Your-Click: Open-domain Regional Image Animation via Motion Prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6018–6026.

Ma, Y.; Yan, Z.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; et al. 2025d. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. 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*, volume 38, 4296–4304.

Narasimhaswamy, S.; Bhattacharya, U.; Chen, X.; Dasgupta, I.; Mitra, S.; and Hoai, M. 2024. Handiffuser: Text-to-image generation with realistic hand appearances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2468–2479.

Pelykh, A.; Sincan, O. M.; and Bowden, R. 2024. Giving a hand to diffusion models: a two-stage approach to improving conditional human image generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–10. IEEE.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shen, F.; Du, X.; Gao, Y.; Yu, J.; Cao, Y.; Lei, X.; and Tang, J. 2025a. IMAGHarmony: Controllable Image Editing with Consistent Object Quantity and Layout. *arXiv preprint arXiv:2506.01949*.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2025b. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6795–6804.

Shen, F.; and Tang, J. 2024. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37: 6246–6266.

Shen, F.; Wang, C.; Gao, J.; Guo, Q.; Dang, J.; Tang, J.; and Chua, T.-S. ???? Long-Term TalkingFace Generation via Motion-Prior Conditional Diffusion Model. In *Forty-second International Conference on Machine Learning*.

Shen, F.; Xu, W.; Yan, R.; Zhang, D.; Shu, X.; and Tang, J. 2025c. IMAGEdit: Let Any Subject Transform. *arXiv preprint arXiv:2510.01186*.

Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2025d. Boosting consistency in story visualization with rich-contextual conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6785–6794.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2024. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Shen, F.; Yu, J.; Wang, C.; Jiang, X.; Du, X.; and Tang, J. 2025e. IMAGGarment-1: Fine-Grained Garment Generation for Controllable Fashion Design. *arXiv preprint arXiv:2504.13176*.

Shi, C.; Chen, Y.; Lei, B.; and Chen, J. 2025a. FashionPose: Text to Pose to Relight Image Generation for Personalized Fashion Visualization. *arXiv preprint arXiv:2507.13311*.

Shi, C.; Li, S.; Guo, S.; Xie, S.; Wu, W.; Dou, J.; Wu, C.; Xiao, C.; Wang, C.; Cheng, Z.; et al. 2025b. Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation. *arXiv preprint arXiv:2511.17282*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Wang, B.; Zhou, J.; Bai, J.; Yang, Y.; Chen, W.; Wang, F.; and Lei, Z. 2025a. Realishuman: A two-stage approach for refining malformed human parts in generated images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7509–7517.

Wang, C.; Deng, Z.; Jiang, Z.; Shen, F.; Yin, Y.; Gan, S.; Cheng, Z.; Ge, S.; and Gu, Q. 2025b. Advanced Sign Language Video Generation with Compressed and Quantized Multi-Condition Tokenization. *arXiv preprint arXiv:2506.15980*.

Wang, C.; Liu, P.; Zhou, M.; Zeng, M.; Li, X.; Ge, T.; and Zheng, B. 2025c. Rhands: Refining malformed hands for generated images with decoupled structure and style guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7573–7581.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zimmermann, C.; and Brox, T. 2017. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, 4903–4911.

Zimmermann, C.; Ceylan, D.; Yang, J.; Russell, B.; Argus, M.; and Brox, T. 2019. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 813–822.