TrioPose: Native Triple-Stream Diffusion Transformers for Pose-Guided Text-to-Image Generation

Dian Gu[1,2,3] **Zhengyi Yang**[1,2,3]

[1]Brainnetome Center, Institute of Automation,
Chinese Academy of Sciences.
[2]School of Artificial Intelligence,
University of Chinese Academy of Sciences.
[3]Beijing Key Laboratory of Brainnetome and Brain-Computer Interface,
Institute of Automation, Chinese Academy of Sciences, Beijing, China.
gudian2024@ia.ac.cn zhengyi.yang@ia.ac.cn

Abstract

Pose-guided text-to-image generation often suffers from limb distortions and feature crosstalk in complex multi-person scenarios. While existing UNet-based adapters struggle with long-range spatial dependencies, emerging Multimodal Diffusion Transformers (MM-DiTs) offer superior global modeling. However, naive signal concatenation in MM-DiTs severely disrupts pre-trained latent distributions. We propose TrioPose, a native pose-driven framework built upon SD3.5M. We introduce a Triple-Stream Pose-Aware DiT (TSPA-DiT) that treats pose as an independent modality. It employs layer-wise activation and zero-initialized dual-residual injection to incorporate geometric constraints. To resolve severe multi-instance occlusions, we design a Learnable Relational Bias Mask that categorizes topological connectivity into fine-grained physical states, mapping them into continuous attention biases to regulate inter-instance interactions. A Pose-Guided Spatial Loss Weighting strategy modulates the native diffusion objective using heatmap-derived error maps, focusing anatomical supervision on distortion-prone regions. Experiments on Human-Art, CrowdPose, and OCHuman evaluate TrioPose across diverse styles, crowded scenes, and severe occlusions. On Human-Art, TrioPose achieves an AP of 58.63 and the best CAP, FID, and KID among the compared methods. Under mixed-dataset training, TrioPose achieves the highest AP among the four compared methods on the Human-Art, CrowdPose, and OCHuman validation sets.

1 Introduction

Text-to-image (T2I) diffusion models [11, 32] have transformed digital content generation. However, high-precision applications, such as animation production [27], virtual try-on [40], and embodied AI simulations, demand control beyond natural language, which often struggles to define the complex joint positions and fine-grained dynamics of deformable entities like humans [11]. Controllable human image generation (HIG) addresses this by aligning high-level semantics with structural priors, such as human pose [24, 36], body parsing [40, 47], and multi-granularity language descriptions [19].

While earlier HIG solutions based on Generative Adversarial Networks (GANs) [8, 23, 25, 41] and Variational Auto-encoders (VAEs) [4, 15] achieved success in specific domains, they often faced training instability and limited multi-modality fusion capabilities. The field has since transitioned

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

to diffusion-based architectures [11, 31]. Earlier pose-guided diffusion methods [14, 42, 45, 17, 26, 35, 38, 39] primarily rely on CNN-based UNet backbones, utilizing frozen weights and auxiliary adapters for spatial conditioning. Although effective for standard single-person synthesis [38], the local receptive field of convolutions limits global interaction modeling. Consequently, multi-person scenarios involving overlapping poses frequently lack long-range spatial dependencies, resulting in limb malformations and anatomical distortions.

Recent methods use Multimodal Diffusion Transformers (MM-DiT) [6], which offer superior global context modeling. However, this transition challenges existing pose control strategies. Plug-in concatenation [26, 35] injects signals at shallow feature levels and can disrupt the pre-trained MM-DiT representation. Furthermore, in multi-subject interactions, global self-attention [35] induces signal crosstalk and identity confusion, while standard hard masking fails to capture true limb dependencies during complex overlaps.

To address these challenges, we propose TrioPose, a native pose modeling framework built upon the SD3.5M model [6]. We introduce the Triple-Stream Pose-Aware DiT (TSPA-DiT), treating pose as an independent modality parallel to text and image. To maintain latent space stability, TSPA-DiT employs Layer-wise Activation and Dual Residual Injection with zero-initialization, enabling a smooth transition from unconditional to constrained generation. To mitigate interference in multi-person interactions, we design a Learnable Relational Bias Mask that decouples topological connectivity into five physical states. These are mapped to continuous attention biases to adaptively regulate cross-regional interaction. Finally, we introduce Pose-Guided Spatial Loss Weighting, where a frozen ViTPose [37] extracts keypoint errors to re-weight the denoising objective, focusing anatomical supervision on distortion-prone regions. TrioPose achieves an AP of 58.63 on Human-Art [13], with the best CAP (81.46), FID (1.99), and KID (1.17) among the compared methods. The main contributions are summarized as follows:

- We propose TSPA-DiT, a novel triple-stream architecture that encodes pose as an independent modality, using zero-initialized Dual Residual Injection for pose control.
- We design a Learnable Relational Bias Mask to map fine-grained physical states into continuous attention biases, effectively mitigating feature crosstalk in complex overlaps via soft constraints.
- We introduce Pose-Guided Spatial Loss Weighting to transform anatomical supervision into latent spatial weights for the native denoising objective.
- Experiments on Human-Art demonstrate that TrioPose achieves an AP of 58.63, leading the compared methods in CAP and image quality as measured by FID and KID.

2 Related works

Controllable diffusion models: Text-to-image diffusion models excel in visual synthesis but struggle with fine-grained spatial control using text alone. To address this, ControlNet [42] introduces a dual-branch architecture with frozen pre-trained weights to encode spatial conditions for structure-aware generation. Similarly, T2I-Adapter [26] and GLIGEN [17] offer efficient feature injection and layout constraints via lightweight adapters and gated self-attention, respectively. For complex multi-condition guidance, Uni-ControlNet [45] designs a unified multi-resolution adapter, while PixelPonder [28] dynamically allocates computational priority to patches during inference, mitigating feature conflicts and accelerating generation. Earlier adapters primarily target UNet architectures. For diffusion transformers, EasyControl [44] introduces lightweight condition injection LoRA and causal attention, while OminiControl [34] jointly processes condition and image tokens through parameter reuse. Transferring spatial conditioning to MM-DiT requires preserving pre-trained representations during condition injection.

Pose-guided human image generation: Pose-guided generation synthesizes human visuals based on structural priors. Early works like PG2 [24] formulated this task as conditional image translation, though they suffered from pixel misalignment during large pose transfers. To handle spatial deformation, PATN [48] leverages progressive attention, while Liquid Warping GAN [21] and DensePose-based methods [9] utilize 3D flows and UV mapping to mitigate occlusion and texture loss. PISE [41] further employs parsing maps for part-level decoupling and local texture injection. Diffusion models have advanced synthesis fidelity; specifically, HumanSD [14] uses heatmap-guided

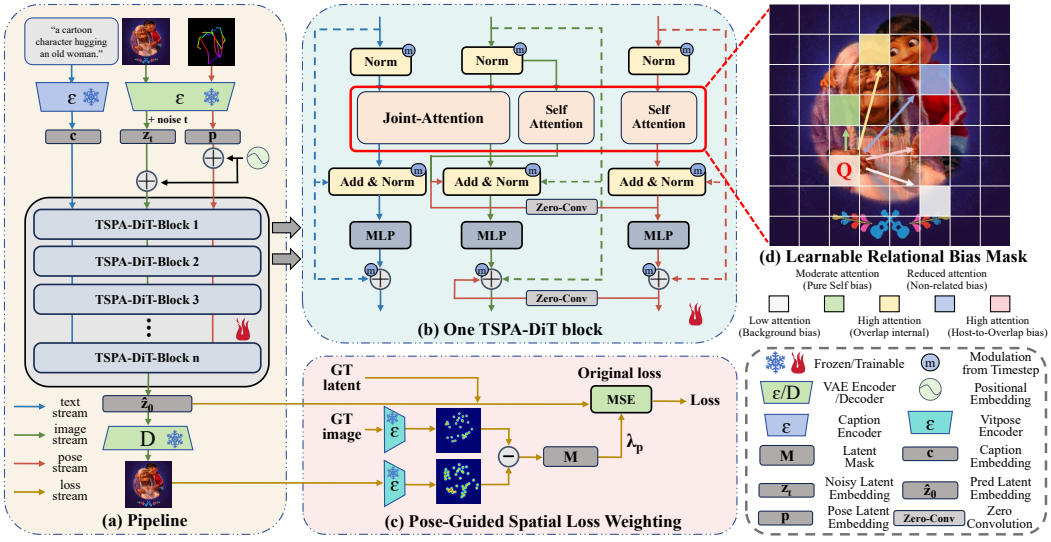

Figure 1: **Overall architecture of TrioPose.** (a) The Triple-Stream Pose-Aware DiT (TSPA-DiT) processes pose as an independent modality. (b) A TSPA-DiT block integrates spatial priors via zero-initialized dual-residual injection to preserve latent stability. (c) Pose-Guided Spatial Loss Weighting modulates the denoising objective using heatmap-derived error maps. (d) The Learnable Relational Bias Mask translates topological connectivity into physical states to decouple inter-instance interference.

losses to enhance native pose awareness. KB-DMGen [20] and Champ [46] integrate global knowledge and 3D geometric information for structural consistency. To improve sparse signal guidance, SP-Ctrl [38] and GRPose [39] leverage learnable representations and graph networks to extract topological correlations. Multi-person interactions and severe occlusions require modeling long-range dependencies and distinguishing overlapping body parts. ControlPose [5] addresses pose control in diffusion transformers by disentangling global context, instance features, and coordinate information. We develop a native pose modeling mechanism with learnable relational biases to regulate interactions between human instances.

3 Method

Preliminaries: rectified flow and MM-DiT TrioPose integrates skeletal priors into the MM-DiT via a native triple-stream architecture built upon a frozen SD3.5M [6] backbone. Optimization is restricted to Low-Rank Adaptation (LoRA) [12] modules, the novel pose branch, zero-initialized injection layers, and continuous attention biases.

SD3.5M employs continuous-time Rectified Flow (RF) [22] to connect data $z_0 \sim p_{\text{data}}$ and noise $\epsilon \sim \mathcal{N}(0, I)$ distributions via a linear trajectory $z_t = (1-t)z_0 + t\epsilon$ for $t \in [0, 1]$, optimizing a parameterized ODE vector field v_θ via:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{\epsilon, z_0, t, c}\left[w(t)\left\|v_\theta(z_t, t, c) - (\epsilon - z_0)\right\|_2^2\right], \tag{1}$$

where $w(t)$ is an empirical weighting function and c denotes the semantic condition. Larger t corresponds to higher noise levels, with $t = 0$ representing data and $t = 1$ representing noise.

Although MM-DiT's dual-stream architecture facilitates semantic alignment by processing text and image modalities independently via joint attention, this isolation complicates the native integration of high-dimensional, dense spatial controls.

Algorithm 1 Generation of attention bias mask

Input: Flattened binarized pose masks $M \in \{0,1\}^{N \times L}$
$C \leftarrow M^\top M, \quad T \leftarrow \sum_{n=1}^{N} M[n,:], \quad R \leftarrow \mathbf{0}^{L \times L}$
for (i,j) **where** $T[i] > 0 \wedge T[j] > 0$ **do**
  if $C[i,j] == 0$ **then** $R[i,j] \leftarrow 3$ ▷ Non-related
  else if $T[i] == 1 \wedge T[j] == 1$ **then** $R[i,j] \leftarrow 1$ ▷ Pure Self
  else if $T[i] \geq 2 \wedge T[j] \geq 2$ **then** $R[i,j] \leftarrow 2$ ▷ Overlap Internal
  else $R[i,j] \leftarrow 4$ ▷ Host-to-Overlap
  end if
end for
return R

Triple-stream pose-aware DiT We propose a native TSPA-DiT (Fig. 1 (a)) . Treating spatially aligned pose latents as an independent modality, a VAE encoder [15] maps them into a 1D sequence T_{pose} . To ensure strict topological alignment, the pose and image streams share identical Spatial Position Embeddings . A layer-wise activation mechanism balances geometric constraints with semantic freedom; the Pose Stream activates exclusively in the first L_p layers, reverting to the standard dual-stream backbone thereafter (Sec. 4.3) .

Abstracting layer normalization and signal modulation into pre-processing $\mathcal{F}_{\text{pre}}(\cdot)$ and post-processing $\mathcal{F}_{\text{post}}(\cdot)$ modules, we design a Dual-Residual Zero-Initialized Injection mechanism for smooth prior integration, computing the intermediate pose feature $\tilde{P}^{(l)}$ and output feature $P_{\text{out}}^{(l)}$ as:

$$\tilde{P}^{(l)} = P_{\text{in}}^{(l)} + \mathcal{F}_{\text{post}}(\text{Attn}(\mathcal{F}_{\text{pre}}(P_{\text{in}}^{(l)}))), \tag{2}$$

$$P_{\text{out}}^{(l)} = \mathcal{F}_{\text{post}}(\text{MLP}(\mathcal{F}_{\text{pre}}(\tilde{P}^{(l)}))). \tag{3}$$

Concurrently, the visual context $I_{\text{out}}^{(l)}$ extracted from the image backbone:

$$I_{\text{out}}^{(l)} = \mathcal{F}_{\text{post}}(\text{Attn}(\mathcal{F}_{\text{pre}}(X_{\text{in}}^{(l)}))) \tag{4}$$

interacts with joint text-image attention, integrating the intermediate pose prior $\tilde{P}^{(l)}$ via a zero-initialized projection (ZeroConv_1):

$$\tilde{X}^{(l)} = X_{\text{in}}^{(l)} + I_{\text{out}}^{(l)} + \mathcal{F}_{\text{post}}(\text{JointAttn}(\mathcal{F}_{\text{pre}}(X_{\text{in}}^{(l)}), \mathcal{F}_{\text{pre}}(C_{\text{in}}^{(l)}))) + \alpha \cdot \text{ZeroConv}_1(\tilde{P}^{(l)}). \tag{5}$$

During the subsequent MLP stage, the image feature undergoes non-linear transformation and fuses the final pose topology $P_{\text{out}}^{(l)}$ via a second injection (ZeroConv_2):

$$X_{\text{out}}^{(l)} = X_{\text{in}}^{(l)} + \mathcal{F}_{\text{post}}(\text{MLP}(\mathcal{F}_{\text{pre}}(\tilde{X}^{(l)}))) + \alpha \cdot \text{ZeroConv}_2(P_{\text{out}}^{(l)}). \tag{6}$$

Learnable relational bias mask In multi-subject generation, global self-attention often induces feature crosstalk, whereas hard boolean masks disrupt non-local dependencies. To resolve this, we propose a Learnable Relational Bias Mask that translates discrete topological connectivity into continuous attention biases. First, instance keypoints are rendered as independent 2D skeletal masks via adaptive thickness and Gaussian smoothing. These masks are max-pooled to the latent resolution and binarized into an occupancy matrix $M \in \{0,1\}^{N \times L}$ for N instances and L tokens. We then derive a token occupancy vector $T = \sum_{n=1}^{N} M[n,:] \in \mathbb{N}^L$ and a co-occurrence matrix $C = M^\top M \in \mathbb{N}^{L \times L}$. Based on T and C, token pairs (i,j) are classified into five physical states (Algorithm 1): (1) Background ($R = 0$), (2) Pure Self ($R = 1$), (3) Overlap Internal ($R = 2$), (4) Non-related ($R = 3$), and (5) Host-to-Overlap ($R = 4$).

Discrete relations R are mapped to a continuous bias B to smoothly modulate attention logits:

$$\text{Attn}(Q,K,V) = \text{Softmax}(QK^\top/\sqrt{d} + B)V. \tag{7}$$

We maintain two sets of five near-zero-initialized learnable scalars: one dedicated to the pose self-attention branch, and another shared by the joint and image self-attention branches. Joint bias injection is restricted to the image-to-image interaction quadrant. Finally, we employ Bias Mask Dropout [33], randomly zeroing B during training.

Pose-guided spatial loss weighting We introduce a Heatmap-Error-Guided Spatial Loss Weighting strategy that translates local anatomical deviations into continuous latent weight maps. We enable spatial reweighting at lower noise levels via a Time-step Truncation Constraint:

$$w_t = \mathbb{I}(t \leq \tau_t), \tag{8}$$

where τ_t is a predefined threshold.

Under this constraint, a frozen ViTPose [37] extracts heatmaps (H^{pred}, H^{gt}) from predicted and ground-truth single-person regions. We aggregate the pixel-wise squared error $\mathbf{E} = (H^{\text{pred}} - H^{\text{gt}})^2$ across keypoint channels, binarizing it to isolate severe distortions. This error map is then Gaussian-smoothed and downsampled into a latent mask M_{latent}:

$$M_{\text{latent}} = \text{Downsample}\big(\text{Gaussian}\big(\bigcup_{k=1}^{K} \text{RoI}_k(\mathbb{I}(\mathbf{E}_k > \tau_e))\big)\big), \tag{9}$$

where τ_e defines the error threshold across K valid RoIs.

The base spatial weight is formulated as:

$$W_{\text{base}} = 1 + \lambda \cdot M_{\text{latent}}, \tag{10}$$

with λ scaling the supervision intensity. Following intra-sample mean normalization to yield W_{final}, this term spatially modulates the original denoising loss $\mathcal{L}_{\text{original}}$ when the gate is active:

$$\mathcal{L}_{\text{total}} = \text{mean}\big([1 + w_t(W_{\text{final}} - 1)] \odot \mathcal{L}_{\text{original}}\big). \tag{11}$$

When $w_t = 0$, the spatial weight is 1 and the base denoising loss remains unchanged. When $w_t = 1$, the spatial weight is W_{final}.

4 Experiments

4.1 Experimental settings

Datasets. TrioPose is evaluated on three datasets covering diverse styles, severe occlusions, and crowded scenes. Human-Art [13] serves as the core benchmark for multi-instance decoupling in natural and artistic scenes. OCHuman [43] focuses on severe occlusion challenges, while CrowdPose [16] addresses complex pose modeling in crowded environments.

Training annotations are augmented using the Qwen3-VL vision-language model [1]. A mixed-prompt strategy is designed to accommodate multi-granularity semantic instructions. Specifically, new text labels comprise detailed long descriptions ranging from 80 to 180 words and concise short descriptions ranging from 30 to 60 words at a seven-to-three ratio. The mixed training set contains 15,500 images. This set consists of 8,000 randomly sampled Human-Art images, 5,000 CrowdPose images, and all 2,500 OCHuman images. For the mixed-dataset comparison in Table 2, all four methods use Qwen3-VL short captions for both training and evaluation. Evaluation uses the original validation splits. All human keypoints are preprocessed into the standard COCO-17 format [18], with relationship matrices generated offline based on instance spatial distributions. Main comparisons and caption ablations use HigherHRNet [3] for pose evaluation; module and hyperparameter ablations use ViTPose-H [37].

4.2 Comparison with SOTA methods

Quantitative evaluation. Table 1 reports performance on the Human-Art dataset using AP, CAP [7], PCE [4], FID [10], KID [2] (scaled by 100), and CLIP-score [30]. With original captions used for both training and evaluation, TrioPose achieves 54.11 AP and 81.00 CAP. ControlPose achieves the highest AP of 61.66, while TrioPose achieves a higher CAP of 81.00 compared with ControlPose's 74.56.

With mixed captions used for training and original captions retained for evaluation, TrioPose† achieves 58.63 AP and 81.46 CAP, improving over original-caption training by 4.52 and 0.46 points, respectively. It achieves the lowest FID (1.99) and KID (1.17) among the compared methods. Its FID is lower than those of HumanSD (10.03) [14], Stable-Pose (11.12) [35], and ControlPose (8.92).

Additionally, TrioPose† achieves a CLIP-score of 34.46, compared with 34.50 for the SD3.5M backbone, while increasing AP from 8.17 to 58.63. Across the compared methods, it ranks second in AP and CLIP-score and first in CAP, FID, and KID.

Table 1: **Results on Human-Art dataset.** Both TrioPose variants are trained on Human-Art, and all methods are evaluated on Human-Art using original captions. Methods are grouped by backbone: SD1.5, Flux.1dev, and SD3.5M. TrioPose and TrioPose† are trained with original and mixed captions, respectively. The best and second-best results for each metric across all methods, including foundation baselines, are highlighted in green and blue, respectively.

Backbone	Method	Pose Accuracy			Image Quality		T2I Alignment
		AP ↑	CAP ↑	PCE ↓	FID ↓	KID ↓	CLIP-score ↑
SD1.5	SD1.5	0.24	55.71	2.30	11.53	3.36	33.33
	T2I-Adapter	27.22	65.65	1.75	11.92	2.73	33.27
	ControlNet	39.52	69.19	1.54	11.01	2.23	32.65
	Uni-ControlNet	41.94	69.32	1.48	14.63	2.30	32.51
	GLIGEN	18.24	69.15	1.46	–	–	32.52
	HumanSD	44.57	69.68	1.37	10.03	2.70	32.24
	Stable-Pose	48.88	70.83	1.50	11.12	2.35	32.60
	GRPose	49.50	70.84	1.43	13.76	2.53	32.31
Flux.1dev	EasyControl	53.21	70.91	1.51	10.45	2.35	32.22
	OminiControl	53.28	71.08	1.47	10.08	2.32	32.54
	ControlPose	61.66	74.56	1.36	8.92	2.10	32.75
SD3.5M	SD3.5M	8.17	80.57	0.92	2.17	1.66	34.50
	ControlNet	13.53	69.06	1.49	9.81	2.31	31.21
	TrioPose (Ours)	54.11	81.00	1.39	3.22	1.29	33.50
	TrioPose† (Ours)	58.63	81.46	1.45	1.99	1.17	34.46



Figure 2: **Qualitative comparisons.** Compared to SD3.5M, HumanSD, and Stable-Pose across diverse styles and complex occlusions, TrioPose shows improved anatomical alignment and clearer instance separation in the illustrated examples.

Qualitative comparison. In scenes with entangled limbs (e.g., Drama and Cartoon in Fig. 2), TrioPose shows clearer separation of hands and legs between instances. Conversely, HumanSD and Stable-Pose frequently suffer from feature entanglement during severe overlaps, such as incorrectly attaching limb A to body B. In Dance, TrioPose preserves the relative placement of the two performers and their extended limbs, while keeping their upper-body contours distinguishable.

Despite missing limbs in the input pose image (e.g., Sketch, Movie, and Relief in Fig. 2), it reconstructs absent anatomical details by combining triple-stream structural modeling with global semantic priors. The Sketch and Relief examples retain line-based rendering and carved surface texture, respectively, while completing body regions not fully specified by the skeleton.

In scenes with numerous small figures (e.g., Digital art in Fig. 2), TrioPose maintains pose alignment and clear boundaries between figures and surrounding regions. The crowd remains concentrated along the lower image boundary, preserving the spatial separation between foreground people and background architecture.

Table 2: **Validation results after mixed-dataset training.** All methods use the same 15.5k-image Human-Art/CrowdPose/OCHuman training set, 10 epochs, and Qwen3-VL-short captions for training and evaluation. AP and CAP use HigherHRNet. KID is scaled by 100. Per-dataset best results are bold; TrioPose rows are shaded.

Dataset	Method	AP ↑	CAP ↑	PCE ↓	FID ↓	KID ↓	CLIP-score ↑
Human-Art	Stable-Pose	40.58	77.75	1.50	9.08	2.48	28.05
	HumanSD	46.42	81.41	1.38	5.51	3.31	28.25
	OminiControl	41.49	80.31	1.44	12.52	3.55	35.15
	TrioPose (Ours)	**49.91**	**81.61**	**1.31**	**1.96**	**1.03**	**36.03**
CrowdPose	Stable-Pose	46.61	72.56	2.12	**0.78**	**0.65**	34.17
	HumanSD	51.47	73.74	2.00	1.04	1.69	33.86
	OminiControl	44.11	74.88	2.39	3.62	2.30	33.24
	TrioPose (Ours)	**51.89**	**75.25**	**1.73**	0.80	0.86	**34.38**
OCHuman	Stable-Pose	38.68	68.11	3.20	**0.81**	**0.71**	**33.91**
	HumanSD	37.82	73.76	2.88	1.96	2.19	32.81
	OminiControl	45.19	74.54	3.22	2.98	2.38	32.86
	TrioPose (Ours)	**53.39**	**78.02**	**2.29**	1.05	1.06	33.59

Robustness in complex multi-human scenarios To evaluate performance in high-density and occluded scenes, we compare TrioPose with Stable-Pose, HumanSD, and OminiControl. All four methods are trained for 10 epochs on the same 15.5k-image mixture of Human-Art, CrowdPose, and OCHuman using Qwen3-VL-short captions. Evaluation uses the respective validation sets with Qwen3-VL-short captions and HigherHRNet for AP and CAP. Table 2 reports all six metrics. TrioPose achieves the highest AP and CAP and the lowest PCE on all three validation sets, indicating improved keypoint alignment, limb orientation, and person-count accuracy.

On the multi-style Human-Art validation set, TrioPose achieves 49.91 AP and 1.96 FID. HumanSD is the strongest baseline on both metrics, with 46.42 AP and 5.51 FID. TrioPose improves AP by 3.49 points and reduces FID by 3.55. It also achieves the highest CAP (81.61) and CLIP-score (36.03), and the lowest PCE (1.31) and KID (1.03), leading on all six metrics on Human-Art.

On the high-density CrowdPose validation set, TrioPose achieves 51.89 AP, compared with 51.47 for HumanSD, 46.61 for Stable-Pose, and 44.11 for OminiControl. Its CAP of 75.25 and PCE of 1.73 also improve over the baselines, with the highest CLIP-score of 34.38. Its FID (0.80) and KID (0.86) are higher than Stable-Pose's 0.78 and 0.65, respectively. On the highly occluded OCHuman validation set, TrioPose achieves 53.39 AP, exceeding the strongest baseline, OminiControl, by 8.20 points. Compared with OminiControl, CAP increases from 74.54 to 78.02 and PCE decreases from 3.22 to 2.29. These gains indicate improvements in limb orientation and person-count accuracy alongside the larger AP margin under severe occlusion. Its FID of 1.05 is lower than those of HumanSD (1.96) and OminiControl (2.98). Stable-Pose achieves the lowest FID (0.81) and KID (0.71), and the highest CLIP-score (33.91 versus 33.59 for TrioPose).

4.3 Ablation study

Table 3: Ablation study on the effectiveness of proposed modules. The Triple-Stream Pose-Aware DiT (TSPA), Learnable Relational Bias Mask (LRBM), and Pose-Guided Spatial Loss Weighting (PSLW) are progressively integrated into the baseline model.$\circ$ denotes the substitution of our proposed LRBM with a traditional binary hard mask. The best results for each metric are highlighted in bold.

Modules			Pose Accuracy			Image Quality		T2I Alignment
TSPA	LRBM	PSLW	AP ↑	CAP ↑	PCE ↓	FID ↓	KID ↓	CLIP-score ↑
×	×	×	8.17	80.57	0.92	2.17	1.66	34.50
✓	×	×	63.50	**83.32**	**0.85**	2.03	1.31	34.39
✓	✓	×	**64.34**	83.25	0.86	1.95	1.17	34.39
✓	×	✓	63.37	83.16	0.88	1.82	0.98	34.45
✓	$\circ$	✓	62.63	83.14	0.87	1.82	1.03	**34.58**
✓	✓	✓	64.33	83.23	0.86	**1.65**	**0.87**	34.46

Component-wise efficacy. In Table 3, adding TSPA within the DiT framework [29] increases AP from 8.17 to 63.50.

Incorporating LRBM further improves AP to 64.34 in multi-person scenarios. Furthermore, replacing our fine-grained 5-state relational mask with a traditional binary hard mask degrades AP to 62.63 and increases FID to 1.82, supporting the use of fine-grained relational biases over binary masking. Integrating PSLW improves visual fidelity, reducing FID from 1.95 to 1.65 and KID from 1.17 to 0.87, with a small change in AP from 64.34 to 64.33. Mean-normalized spatial reweighting redistributes denoising supervision toward distortion-prone regions.

Table 4: **Ablation study on TS-PA-DiT configuration.** Top: ablating the number of activated layers (L_p). Bottom: ablating the injection scale factor (α). The default setting is highlighted in gray.

Config		Pose Accuracy			Image Quality		T2I Alignment
L_p	α	AP ↑	CAP ↑	PCE ↓	FID ↓	KID ↓	CLIP ↑
6	0.5	63.03	**83.34**	0.88	1.78	1.03	**34.49**
12	0.5	**64.33**	83.23	0.86	**1.65**	**0.87**	34.46
24	0.5	62.70	82.97	**0.85**	2.27	1.17	34.37
12	0.3	64.19	**83.25**	0.87	2.00	0.98	34.43
12	0.5	**64.33**	83.23	0.86	**1.65**	**0.87**	**34.46**
12	0.7	62.87	83.20	**0.85**	2.01	0.93	34.37

Optimization of TSPA-DiT configuration. Sensitivity analysis of the triple-stream architecture focuses on pose stream activation layers (L_p) and injection scale (α). As shown in Table 4, $L_p = 12$ yields the highest AP (64.33) and lowest FID (1.65). Both $L_p = 6$ and $L_p = 24$ yield lower AP, with FID increasing to 2.27 at $L_p = 24$.

Similarly, $\alpha = 0.5$ yields the highest AP and lowest FID among the tested injection scales. Both lower ($\alpha = 0.3$) and higher ($\alpha = 0.7$) injection scales yield lower AP and higher FID.

Initialization strategies for relational bias mask. Table 5 compares relational bias mask initialization strategies: fixed non-learnable biases, zero initialization ($\boldsymbol{\theta} = \mathbf{0}$), and physics-derived prior-guided initialization ($\boldsymbol{\theta}_{\text{prior}}$). Learnable biases lower FID and KID compared to fixed biases; zero initialization yields 63.12 AP.

Training dynamics reveal that zero-initialized parameters naturally converge toward specific physical magnitudes, establishing the topological prior $\boldsymbol{\theta}_{\text{prior}}$. Specifically, associated states (e.g., pure self, overlap internal) receive positive initializations, while physically inhibitory connections (non-related subjects) are uniformly initialized to -0.3. Prior-guided initialization increases AP to 64.33 and reduces FID to 1.65 compared with zero initialization.

Table 5: **Ablation study on relational bias initialization.** We compare fixed (non-learnable) biases with learnable biases under different initialization strategies. By employing empirical topological priors ($\boldsymbol{\theta}_{\text{prior}}$) derived from the steady-state convergence trajectories of zero-initialized parameters, the model achieves higher AP and lower FID and KID than the zero initialization ($\mathbf{0}$) baseline.

	Pose Accuracy			Image Quality		T2I Alignment
Bias Strategy	AP ↑	CAP ↑	PCE ↓	FID ↓	KID ↓	CLIP-score ↑
Fixed Bias	62.93	83.16	0.87	2.08	1.02	34.43
Learnable ($\boldsymbol{\theta} = \mathbf{0}$)	63.12	**83.30**	0.88	1.81	0.96	34.45
Learnable ($\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{prior}}$)	**64.33**	83.23	**0.86**	**1.65**	**0.87**	**34.46**

Impact of spatial loss weighting. Sensitivity analysis on the truncation threshold τ_t (Eq. 8) and spatial weight coefficient λ (Eq. 10) investigates the trade-off between anatomical supervision (AP) and generative fidelity (FID). Fig. 3 compares these settings. For τ_t, smaller values (0.2 or 0.35) restrict spatial reweighting to a narrower low-noise interval and yield higher FID despite small gains in AP. Increasing τ_t to 0.5 lowers FID with a small decrease in AP.

A larger weight ($\lambda = 0.3$) yields lower AP and higher FID than $\lambda = 0.1$. We select $\tau_t = 0.5$ and $\lambda = 0.1$ (dual-colored star in Fig. 3) to balance pose accuracy and image quality.

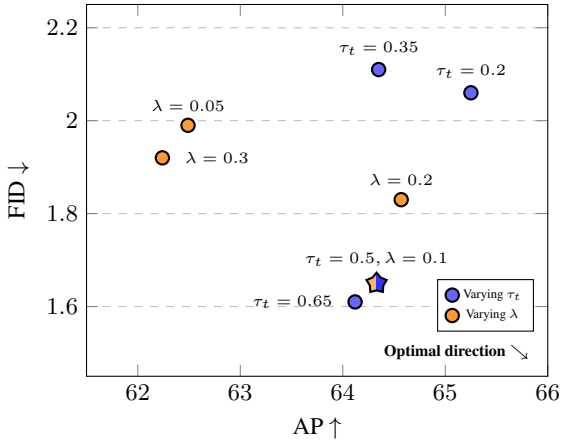

Figure 3: Impact of hyper-parameters τ_t and λ on the AP-FID trade-off.

Table 6: **Training caption ablation on Human-Art.** Only training captions vary; evaluation uses original captions. Mixed combines Qwen3-VL-short and Qwen3-VL-long captions. AP and CAP use HigherHRNet; real images for FID/KID are resized with TF.BICUBIC. Best results are bold.

	Pose Accuracy			Image Quality		T2I Alignment
Training Caption	AP ↑	CAP ↑	PCE ↓	FID ↓	KID ↓	CLIP-score ↑
Original	54.11	81.00	**1.39**	3.22	1.29	33.50
Qwen3-VL-short	58.04	**81.82**	1.42	**1.81**	**0.97**	**36.97**
Qwen3-VL-long	56.49	81.13	1.48	3.03	1.28	34.36
Mixed	**58.63**	81.46	1.45	1.99	1.17	34.46

Training caption ablation. Table 6 separates the effects of text augmentation and training caption granularity. Qwen3-VL captions improve AP by 2.38–4.52 points over original-caption training. Human-Art annotations average approximately 10–11 words and mainly contain style tags and brief phrases; augmented descriptions supply additional semantic and spatial information for text conditioning. Under original-prompt evaluation, Qwen3-VL-short training achieves higher AP (58.04 versus 56.49) and lower FID (1.81 versus 3.03) than Qwen3-VL-long training, consistent with a closer match between training and evaluation text granularity. Mixed training achieves the highest AP (58.63), while Qwen3-VL-short training leads in CAP, FID, KID, and CLIP-score. We use mixed training to expose the model to both caption granularities and evaluate its response to different prompts below.

Evaluation prompt ablation. Table 7 varies inference prompts for the same mixed-caption model. From original to Qwen3-VL-long prompts, AP increases from 58.63 to 60.74, CAP from 81.46 to 83.38, and FID decreases from 1.99 to 1.75. Qwen3-VL-short and mixed prompts also improve these metrics over original prompts. These results show that the model uses additional textual information

Table 7: **Evaluation prompt ablation on Human-Art.** Only prompts vary for a fixed mixed-caption model. AP and CAP use HigherHRNet. Best results are bold.

	Pose Accuracy			Image Quality		T2I Alignment
Evaluation Prompt	AP ↑	CAP ↑	PCE ↓	FID ↓	KID ↓	CLIP-score ↑
Original	58.63	81.46	1.45	1.99	1.17	34.46
Qwen3-VL-short	59.07	82.50	1.39	1.87	1.15	**36.58**
Mixed	60.25	83.13	1.34	1.77	1.01	35.08
Qwen3-VL-long	**60.74**	**83.38**	**1.32**	**1.75**	**0.96**	34.42

to improve pose alignment and image quality across the tested prompt granularities. Qwen3-VL-short prompts yield the highest CLIP-score (36.58). CLIP-score uses a 77-token context window, so long prompts are truncated for this metric. Together, the two ablations support mixed training for inputs of varying granularity and show the benefit of detailed prompts at inference.

5 Conclusion

We propose TrioPose, a native pose-driven framework that addresses feature crosstalk and anatomical distortion in multi-person image generation. By treating pose as an independent modality within the TSPA-DiT architecture and employing a Learnable Relational Bias Mask, our method regulates interactions between instances. On Human-Art, TrioPose achieves the best CAP, FID, and KID among the compared methods, with the second-highest AP. Under mixed-dataset training, TrioPose achieves the highest AP among the four compared methods on all three validation sets, including an 8.20-point improvement over the strongest baseline on OCHuman.

TrioPose can be misused to generate deepfakes. We plan to integrate digital watermarking into our released checkpoints for provenance tracking. Future work will explore lightweight attention mechanisms to accelerate inference and extend the framework for temporally consistent pose-guided video synthesis.

Acknowledgments and Disclosure of Funding

We thank our supervisor for guidance and the Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, for providing GPU resources. This work received no external funding. The authors received no other third-party support for this work during the preceding 36 months and declare no external financial relationships that could be perceived as influencing this work.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.

[4] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. Kpe: Keypoint pose encoding for transformer-based image generation. *arXiv preprint arXiv:2203.04907*, 2022.

[5] Zhongjing Du, Xiao Chen, Zhiwei Nie, Yuxuan Chen, Chang Liu, Xiangyang Ji, and Jie Chen. Controlpose: High-fidelity pose-controlled image generation with multi-faceted pose disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3851–3860, 2026.

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[7] freshsomebody. Posenet similarity. `https://github.com/freshsomebody/posenet-similarity`, 2023. Accessed: 2026-05-05.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[9] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2019.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[13] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–629, 2023.

[14] 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*, pages 15988–15998, 2023.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[17] 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*, pages 22511–22521, 2023.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[19] Deyin Liu, Lin Wu, Feng Zheng, Lingqiao Liu, and Meng Wang. Verbal-person nets: Pose-guided multi-granularity language-to-person generation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8589–8601, 2022.

[20] Shibang Liu, Xuemei Xie, and Guangming Shi. Kb-dmgen: Knowledge-based global guidance and dynamic pose masking for human image generation. *arXiv preprint arXiv:2507.20083*, 2025.

[21] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5903–5912. IEEE, 2019.

[22] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[23] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10806–10815, 2021.

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

[25] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020.

[26] 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 *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[27] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2920–2930, 2024.

[28] Yanjie Pan, Qingdong He, Zhengkai Jiang, Pengcheng Xu, Chaoyi Wang, Jinlong Peng, Haoxuan Wang, Yun Cao, Zhenye Gan, Mingmin Chi, et al. Pixelponder: Dynamic patch adaptation for enhanced multi-conditional text-to-image generation. *arXiv preprint arXiv:2503.06684*, 2025.

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[34] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14940–14950. IEEE, 2025.

[35] Jiajun Wang, Morteza Ghahremani, Yitong Li, Björn Ommer, and Christian Wachinger. Stable-pose: Leveraging transformers for pose-guided text-to-image generation. *Advances in Neural Information Processing Systems*, 37:65670–65698, 2024.

[36] Xiaogang Xu, Ying-Cong Chen, Xin Tao, and Jiaya Jia. Text-guided human image manipulation via image-text shared space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6486–6500, 2021.

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023.

[38] 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*, pages 15896–15906, 2025.

[39] Xiangchen Yin, Donglin Di, Lei Fan, Hao Li, Wei Chen, Yang Song, Xiao Sun, Xun Yang, et al. Grpose: Learning graph relations for human image generation with pose priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9526–9534, 2025.

[40] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10511–10520, 2019.

[41] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7982–7990, 2021.

[42] 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*, pages 3836–3847, 2023.

[43] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019.

[44] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19513–19524. IEEE, 2025.

[45] 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. *Advances in neural information processing systems*, 36:11127–11150, 2023.

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

[47] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE international conference on computer vision*, pages 1680–1688, 2017.

[48] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2347–2356, 2019.

A Appendix: data processing details

This section details our data standardization and annotation pipeline for Human-Art, OCHuman, and CrowdPose datasets.

A.1 Pose standardization

We unify all skeletal annotations into the COCO-17 format (17×3 keypoints).

OCHuman: Original COCO-style annotations are converted into instance-level `.npz` files.

CrowdPose: The 14-keypoint annotations are mapped to COCO indices 5–16. Head and neck points without direct COCO equivalents are set to zero to maintain a consistent 17×3 topology.

A.2 Multi-granularity captioning

We use Qwen3-VL-30B-A3B-Instruct to augment training images with multi-scale textual descriptions.

Inference Settings: We employ the vLLM engine with 4-bit AWQ quantization. Images are resized to a maximum side of 1280 pixels while preserving the aspect ratio.

Caption Types: For each image, we generate a *short_caption* (30–60 words) for core attributes and a *long_caption* (80–180 words) for detailed spatial and pose descriptions.

Style Preservation: For artistic datasets like Human-Art, we prepend original style tags (e.g., "oil painting") to the VLM-generated prompts.

A.3 Mixed training configuration

To balance controllability and robustness, we implement a mixed-prompt training strategy.

Mixing Strategy: We use a *long_ratio* of 0.7, meaning 70% of samples use long captions and 30% use short captions. Samples are assigned using a fixed random seed.

Dataset Scale: The final mixed training set contains **15,500** images, comprising 8,000 from Human-Art, 5,000 from CrowdPose, and 2,500 from OCHuman.

Format: All metadata is stored in a unified JSON mapping, pairing image paths with synchronized captions and pose files.

Caption mixing and dataset mixing are separate settings. The Human-Art caption ablations compare original, Qwen3-VL-short, Qwen3-VL-long, and mixed captions. The four-method mixed-dataset comparison uses Qwen3-VL-short captions for both training and evaluation.

B Implementation details

SD3.5-Medium serves as the pre-trained backbone. In the triple-stream architecture, the pose stream activation layer L_p is set to 12, and the injection scale α is set to 0.5. Training uses AdamW8bit with a learning rate of 4×10^{-5} for MM-DiT LoRA parameters. A separate learning rate of 10^{-4} with FP32 optimization is applied to the ten learnable relational bias parameters. For spatial loss weighting, $\lambda = 0.1$ and $\tau_t = 0.5$, with a linear warmup of the spatial weighting coefficient over the first 500 steps. The timestep gate enables spatial reweighting at $t \leq 0.5$; the base denoising loss is retained at higher noise levels. Text prompts are randomly dropped with probability 0.1 during training and retained during inference. Training runs for 10 epochs on three NVIDIA A40 GPUs, with bias mask dropout of 0.2. Inference uses 50 Euler steps with a discrete flow scheduler and shift 3.0.

Table 8 summarizes the optimization settings. Text embeddings are cached offline in the default training configuration. The batch size of 13 per GPU applies to model training; the computational measurements in Section F use a batch size of 1 per GPU.

Table 8: Training configuration for TrioPose. The LoRA scaling parameter is distinct from the pose injection scale.

Setting	Configuration
Resolution and seed	512×512; seed 1026
Training duration	10 epochs
Batch size	13 per GPU on 3 A40 GPUs; accumulation 1; global batch 39
LoRA	Rank 32, scaling parameter 32, multiplier 1.0
LoRA+	Learning-rate ratio 16
Trainable parameters	72.9M, approximately 2.6% of the 2.81B extended model
Optimizer	AdamW8bit; $\beta = (0.9, 0.95)$; weight decay 0.01
Learning rates	MM-DiT LoRA: 4×10^{-5}; relational biases: 10^{-4}, FP32
Schedule	Cosine decay; minimum learning-rate ratio 0.1; 500 warmup steps
Precision and memory	BF16 with gradient checkpointing; cached text embeddings
Pose injection	$L_p = 12$; injection scale $\alpha = 0.5$
Relational bias	Ten learnable scalars; mask dropout 0.2
Spatial loss weighting	$\lambda = 0.1$; $\tau_t = 0.5$; 500-step linear weight warmup
Denoising loss	L2; min-SNR weighting with $\gamma = 5$; debiased estimation
Text dropout	Probability 0.1
Inference	Euler, discrete flow scheduler, shift 3.0, 50 steps

C Evaluation metrics

We use six metrics covering pose accuracy, image quality, and semantic alignment.

C.1 Evaluation protocols

Main method comparisons and caption ablations use HigherHRNet [3] for AP and CAP. Original module, architecture-configuration, relational-bias-initialization, and spatial-loss hyperparameter ablations retain ViTPose-H [37]. ViTPose also supplies the training-time heatmap errors for PSLW. These training and evaluation roles are distinct.

Human-Art method comparisons use original evaluation captions. The training-caption ablation also fixes original evaluation captions, while the evaluation-prompt ablation varies prompts for the same mixed-caption model. Mixed-dataset comparisons use Qwen3-VL-short captions for both generation and evaluation. For the main comparisons and caption ablations, real images in FID/KID computation use TF.BICUBIC interpolation. CLIP-score uses the standard 77-token context window, truncating longer text inputs.

C.2 Pose accuracy metrics

These metrics quantify the structural fidelity of the generated human instances relative to the input pose conditions:

Average Precision (AP): Following the COCO evaluation protocol, AP compares keypoints extracted from generated images with the conditional poses across Object Keypoint Similarity (OKS) thresholds. Main comparisons and caption ablations use HigherHRNet [3], while module and hyperparameter ablations use ViTPose-H [37].

Cosine Similarity-based Average Precision (CAP): CAP measures the angular alignment of limb vectors between the generated and conditional poses. It evaluates joint orientations in multi-person scenes.

Person Count Error (PCE): PCE evaluates the model's ability to generate the exact number of human instances specified by the pose condition. It is defined as the absolute difference between the number of detected persons in the generated image and the number of instances in the input pose map.

C.3 Image fidelity and diversity

We evaluate the visual realism and distribution alignment of the synthesized results using the following:

Fréchet Inception Distance (FID): FID calculates the Fréchet distance between Gaussian distributions fitted to features extracted by a pre-trained Inception-V3 network for both real and generated image sets. Lower FID values indicate higher image fidelity and better coverage of the real data distribution.

Kernel Inception Distance (KID): KID measures the squared Maximum Mean Discrepancy (MMD) between Inception representations. Unlike FID, KID is an unbiased estimator, making it more reliable for smaller validation sets. Reported KID values are scaled by 100.

C.4 Text-image semantic alignment

CLIP Score: This metric utilizes a pre-trained CLIP (ViT-L/14) model to compute the cosine similarity between the latent embeddings of the generated image and the corresponding text prompt. It evaluates the model's success in following complex semantic instructions and text-guided attribute synthesis.

D Appendix: dataset and model assets

This section documents the datasets, pre-trained models, and their licenses.

D.1 Detailed asset descriptions

Our framework and the comparative baselines are trained and evaluated on several distinct assets as described below:

Human-Art [13]: The Human-Art dataset comprises 38,000 images distributed across 19 distinct scenarios, encompassing natural scenes, 2D artificial scenarios (e.g., oil paintings, sketches), and 3D artificial scenarios (e.g., statues, digital art). We adopt the standardized train-validation split as suggested by the authors. The dataset is licensed under the Attribution-Non Commercial-Share Alike 4.0 International License (CC BY-NC-SA 4.0).

CrowdPose [16]: This dataset is specifically designed to address human pose estimation in crowded scenes. It contains approximately 20,000 images with 80,000 human instances. The images are categorized by a "crowd index" to reflect varying degrees of occlusion and instance density. It is released under the Creative Commons Attribution 4.0 International License (CC BY 4.0) for academic and non-commercial research purposes.

OCHuman [43]: The OCHuman dataset focuses on "Occluded Human" instances, providing 5,000 images with over 10,000 instances characterized by extreme overlap and occlusion (at least 50% of the body). It serves as a rigorous benchmark for multi-instance decoupling. The annotations are licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

COCO-17 [18]: As a foundational benchmark, the MS-COCO 2017 dataset provides over 200,000 images across 80 object categories. For our task, we utilize the person category with standard 17-keypoint annotations. The dataset is licensed under the Creative Commons Attribution 4.0 License, posing no specific restrictions for research.

SD3.5-Medium [6]: This serves as our pre-trained generative backbone, utilizing a Multimodal Diffusion Transformer (MM-DiT) architecture. The model is released by Stability AI under the Stability Community License, which allows free use for academic research and small-scale commercial applications.

Qwen3-VL [1]: We employ Qwen3-VL-30B-A3B-Instruct to augment our training data with multi-granularity textual descriptions. The model weights are released under the Apache License 2.0.

ViTPose [37]: A Vision Transformer-based pose estimator used in our Pose-Guided Spatial Loss Weighting strategy to extract keypoint heatmaps. The model is available under the Apache 2.0 License.

D.2 Asset summary and licensing

Table 9 provides a quick reference for the licensing information and primary utility of the core assets used in our experiments.

Table 9: Summary of dataset and model licenses

Asset Name	Primary Use	Creator/Source	License Type
Human-Art	Benchmark	Ju et al.	CC BY-NC-SA 4.0
CrowdPose	Benchmark	Li et al.	CC BY 4.0 / Academic
OCHuman	Benchmark	Zhang et al.	CC BY-NC 4.0
COCO-17	Pre-training	Microsoft COCO Team	CC BY 4.0
SD3.5-Medium	Backbone Model	Stability AI	Stability Community
Qwen3-VL	Data Augmentation	Alibaba Qwen Team	Apache License 2.0
ViTPose	Loss Weighting	Xu et al.	Apache 2.0

D.3 Terms of use and compliance

For all assets listed above, we have ensured compliance with their respective legal frameworks:

Academic and Non-commercial Use: Benchmarks such as **Human-Art** and **OCHuman** are restricted to non-commercial research purposes. We have verified our usage aligns with these constraints and have not utilized the data for commercial product development.

Model Redistribution: The **SD3.5-Medium** model is governed by the Stability Community License, which permits free use for research. Our academic publication and proposed release of derived weights fall within these permitted boundaries.

Standardized Protocol: All datasets used in this research adhere to a standardized keypoint format featuring 17 keypoints per person, consistent with the COCO and Human-Art protocols.

Data Integrity: We have not re-distributed raw dataset files; instead, we provide scripts to acquire data directly from original sources to respect creators' distribution rights.

E Alternative pose injection designs

We examine four direct pose injection schemes on SD3.5M, covering channel concatenation, addition, and projection. Table 10 specifies their input representations and observed generation failures. Both concatenation variants produce triangular structural artifacts, while addition and projection produce severe overexposure. These configurations did not produce valid pose estimates for quantitative comparison; no AP or distributional quality scores are reported for them.

Table 10: Direct pose injection designs on SD3.5M and their observed failure modes.

Design	Pose injection	Observation
RGB concatenation	Render the skeleton as a three-channel RGB map at latent resolution; concatenate with the 16-channel noisy latent ($16 + 3 = 19$). Zero-initialize the input projection for the added channels.	Triangular structural artifacts and content collapse.
VAE concatenation	Gaussian-smooth and normalize the skeleton map, encode it into a 16-channel VAE latent, and concatenate with the noisy latent ($16 + 16 = 32$).	Triangular structural artifacts.
Pose addition	Scale and add the pose latent to the noisy latent without changing the input channel count.	Severe overexposure.
Concatenation with projection	Project the 32-channel concatenation back to 16 channels using a zero-initialized learnable convolution.	Persistent overexposure.

Concatenation changes the input representation presented to the pre-trained transformer, while addition introduces the sparse pose signal throughout the latent field, including background regions. The observed failures motivate introducing pose as a separate stream with zero-initialized residual injection and limiting structural injection to selected layers.

ControlNet-style adaptation. We also compare cloned dual-stream control branches on SD3M and SD3.5M. The SD3M variant uses InstantX pre-trained ControlNet weights and has 538.8M trainable parameters; the SD3.5M variant has 638.0M trainable parameters. The former achieves 13.17 AP, 65.27 CAP, and 11.16 FID, while the latter achieves 13.53 AP, 69.06 CAP, and 9.81 FID. Thus, the 13.17 AP result belongs to SD3M. On SD3.5M, TrioPose achieves 58.63 AP, 81.46 CAP, and 1.99 FID. Within these experiments, retaining the same backbone while changing the pose-control mechanism produces a substantially larger AP difference than changing the backbone of the ControlNet-style baseline.

F Computational efficiency

F.1 Measurement settings

Inference is measured on a single NVIDIA A40 at 512×512 resolution, batch size 1, and BF16 precision, using 5 warmup runs and a measurement window of 100 iterations. Training measurements use batch size 1 per GPU and BF16 precision. Text encoding (TE) is cached offline in the default TrioPose training setup; an online-TE configuration includes text encoding in each training step. These measurements are separate from the global-batch-39 training configuration in Table 8.

F.2 Inference efficiency

Table 11 reports runtime, memory, parameter counts, sampling configurations, and FLOPs. TrioPose takes 13.99 seconds per image at 50 sampling steps, compared with 12.38 seconds for SD3.5M+ControlNet. Its peak memory is 17.3 GiB versus 17.6 GiB. OminiControl takes 22.90 seconds at its native 28 steps; the 40.89-second value at 50 steps is a linear estimate from per-step latency. HumanSD has the lowest latency and memory use, with a smaller SD1.5 backbone.

Table 11: Inference efficiency on one A40. Latency is mean $\pm$ standard deviation. The 50-step OminiControl latency is estimated; all other per-image latencies are measured at the listed step counts. G denotes 10^9 operations. Activated parameters and denoising-model parameters use different counting scopes.

Method	Steps	Latency (s/image)	50-step latency (s)	Latency (ms/step)	Images/s	Memory (GiB)
HumanSD	50	7.328 ± 0.039	7.33	146.6	0.136	6.25
TrioPose	50	13.987 ± 0.051	13.99	279.7	0.071	17.3
SD3M+ControlNet	50	12.476 ± 0.045	12.48	249.5	0.080	16.4
SD3.5M+ControlNet	50	12.377 ± 0.097	12.38	247.5	0.081	17.6
OminiControl	28	22.900 ± 0.207	40.89 (est.)	817.9	0.044	32.1

Method	Activated parameters (B)	FLOPs (G/step)	FLOPs (G/image)	Sampler / scheduler
HumanSD	1.30	1,357.2	67,857.7	DDIM / DDIMScheduler
TrioPose	8.47	5,916.4	295,818.5	Euler / DiscreteFlow (shift 3.0)
SD3M+ControlNet	8.23	4,662.8	233,138.8	FlowMatchEulerDiscrete
SD3.5M+ControlNet	8.55	5,397.8	269,887.7	FlowMatchEulerDiscrete
OminiControl	16.89	31,467	881,081	FlowMatchEulerDiscrete

F.3 Training efficiency and text encoding

Table 12 separates online and offline text encoding. With online TE, TrioPose uses 5.98 T forward FLOPs per step, compared with 7.90 T for SD3.5M+ControlNet, a reduction of 24.3%. It trains 72.9M parameters versus 638.0M, a reduction of 88.6%. Its measured step time is higher (2.060 versus 0.926 seconds); the training pipeline also performs ViTPose-based spatial error estimation for PSLW. Parameter count and forward FLOPs therefore describe different aspects of the training cost from elapsed step time.

Offline text caching reduces TrioPose's forward FLOPs from 5.98 T to 3.47 T and peak memory from 25.2 GiB to 15.2 GiB. The extended denoising model contains 2.81B parameters, comprising the 2.24B SD3.5M backbone and a 0.57B third stream. Text encoders are counted separately in the activated-parameter measurements in Table 11.

Table 12: Training efficiency at batch size 1 per GPU and BF16 precision. TE denotes text encoding; T denotes 10^{12} operations. Denoising-model parameters exclude text encoders and include the third stream for TrioPose. ControlNet branch parameters are listed as trainable parameters separately from their backbones; HumanSD's trainable count includes its UNet and pose estimator. Step time is mean $\pm$ standard deviation.

Method	Model (B)	Trainable (M)	Forward FLOPs (T/step)	Time (s/step)	Steps/s	Memory (GiB)
HumanSD	0.8659	929.7	9.42	1.025 ± 0.024	0.98	20.2
TrioPose (offline TE)	2.81	72.9	3.47	1.913 ± 0.041	0.52	15.2
TrioPose (online TE)	2.81	72.9	5.98	2.060 ± 0.244	0.49	25.2
SD3M+ControlNet	2.03	538.8	7.41	0.634 ± 0.063	1.58	19.42
SD3.5M+ControlNet	2.24	638.0	7.90	0.926 ± 0.119	1.08	21.5
OminiControl	11.92	14.49	44.78	2.819 ± 0.097	0.36	32.8

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the end of Section 1 explicitly list the paper's core claims, including the proposed TSPA-DiT architecture, the Learnable Relational Bias Mask, and the Pose-Guided Spatial Loss Weighting. These methodological claims and the reported state-of-the-art performance are thoroughly evaluated and supported by the extensive experiments detailed in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our approach, specifically its dependency on upstream pose detection accuracy and the increased computational overhead of the architecture, in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper focuses on proposing a novel deep learning architecture and empirical methodology, detailed in Section 3, and therefore does not include formal theoretical results, theorems, or mathematical proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a comprehensive description of the proposed architecture and algorithms in Section 3, and fully disclose all datasets, hyperparameters, and training configurations in Section 4.1 and Appendix B to ensure full reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The source code and our augmented mixed-prompt dataset are currently withheld. We are fully committed to publicly releasing all code, data processing scripts, and model checkpoints immediately upon the paper's acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail the dataset composition and evaluation splits in Section 4.1, discuss hyperparameter selection in Section 4.3, and provide comprehensive training details, including the AdamW8bit optimizer, precise learning rates, and hardware configurations in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Due to the substantial computational cost required to train the large-scale diffusion transformer (TSPA-DiT) across our mixed dataset, performing multiple independent runs to calculate error bars is computationally infeasible. Instead, we follow standard practice in generative modeling by reporting exact point estimates for deterministic metrics (e.g., AP, FID) on fixed validation splits.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the hardware compute resources utilized for our experiments, specifically the deployment of three NVIDIA A40 GPUs and the total number of training steps, in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and confirm that our research fully complies with all guidelines, as our study relies exclusively on standard, publicly available benchmarks and introduces no unethical practices or unconsented human subject research.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:We discuss the potential positive impacts of streamlining production workflows in creative industries in Section 1, and address negative societal risks (e.g., Deepfakes) alongside corresponding mitigation strategies like digital watermarking in Section 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We plan to integrate digital watermarking into our model checkpoints to support provenance tracking and mitigate misuse, as discussed in Section 5.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the original creators of all utilized models and datasets in Section 4.1, and explicitly detail their respective licenses and terms of use in Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release any new foundational assets at this time; our experiments rely entirely on existing public benchmarks and their derived subsets, as detailed in Section 4.1.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve any crowdsourcing or research with human subjects; all experiments and evaluations are conducted using automated objective metrics on publicly available benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects, as all experiments are conducted using existing public benchmark datasets; therefore, IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We utilized the Qwen3-VL vision-language model to augment training annotations and generate multi-granularity textual descriptions for our mixed dataset. This usage is an integral part of our data processing pipeline and is detailed in Section A.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.