This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Rethink Sparse Signals for Pose-guided Text-to-image Generation

Wenjie Xuan[1], Jing Zhang[1]*, Juhua Liu[1]*, Bo Du[1], Dacheng Tao[2]
[1] School of Computer Science,Wuhan University [2] Nanyang Technological University
{dreamxwj,liujuhua,dubo}@whu.edu.cn, {jingzhang.cv,dacheng.tao}@gmail.com

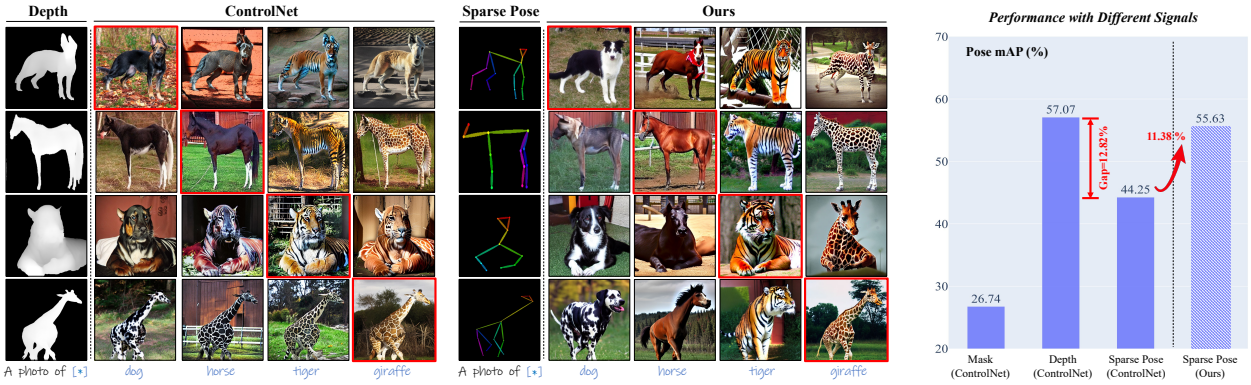


Figure 1. Performance of dense (*i.e.*, depth) and sparse signals for pose-guided text-to-image generation. Though sparse pose signals are simple, they are supposed to suffer from poor pose control than dense signals. Our work improves the pose control of sparse signals, achieving competitive pose alignment with dense signals while enjoying image diversity and cross-species generalization capability.

Abstract

Recent works favored dense signals (e.g., depth, DensePose), as an alternative to sparse signals (e.g., OpenPose), to provide detailed spatial guidance for pose-guided text-to-image generation. However, dense representations raised new challenges, including editing difficulties and potential inconsistencies with textual prompts. This fact motivates us to revisit sparse signals for pose guidance, owing to their simplicity and shape-agnostic nature, which remains underexplored. This paper proposes a novel **S***patial-***P***ose ControlNet (SP-Ctrl), equipping sparse signals with robust controllability for pose-guided image generation. Specifically, we extend OpenPose to a learnable spatial representation, making keypoint embeddings discriminative and expressive. Additionally, we introduce keypoint concept learning, which encourages keypoint tokens to attend to the spatial positions of each keypoint, thus improving pose alignment. Experiments on animal- and human-centric image generation tasks demonstrate that our method outperforms recent spatially controllable text-to-image generation approaches under the sparse-pose guidance and even matches the performance of dense signal-based methods. Moreover, SP-Ctrl shows promising capabilities in diverse and cross-species generation through sparse signals. Code is available here.*

* Corresponding author

1. Introduction

The recent success of pose-guided text-to-image (T2I) generation [25, 37, 39, 41, 42, 59, 60] has encouraged applications including animal- and human-centric image generation [17, 20, 26, 48, 52, 56], pose-driven image animation [5, 14, 54, 63], and conditional 3D generation [28, 38]. Recent efforts tend to employ dense conditions (*e.g.*, depth and normal maps, DensePose, SMPL) [21, 24, 31, 34, 54] or richer pose representations (*e.g.*, textual prompts, graphs, ViT features) [16, 26, 56] for more precise pose control. However, as validated in Fig. 1, such dense signals raised new challenges: 1) **Inflexibility.** The dense pose signals are commonly extracted from reference images with pretrained detectors, not only restricting the creation of new conditions but also raising the difficulties for editing. 2) **Contradictions with textual prompts.** Dense signals impose strong constraints on objects' shapes and contours. If these priors are violated, it would severely degrade fidelity and produce undesired results [18, 32, 40, 55].

In contrast, sparse signals like OpenPose [4] inherently mitigate the above risks due to three appealing properties: 1) **Shape-agnostic.** The sparse pose signal is an abstract of objects (*e.g.*, humans, animals) in anatomy. It is usually presented in graph structures, comprising a series of pre-defined keypoints and skeletons with topological structures. 2) **Category-agnostic.** The definitions of keypoints

are commonly shared in natural language, especially for animals. The topology structures are also similar among species like mammals despite subtle variances in skeleton proportions and actions, which provide a unified pose representation [57]. 3) **Operability.** The sparse signals do not necessarily rely on reference images, offering high freedom in creation and manipulation. However, though early methods like ControlNet [59] and T2I-Adapter [37] adopt OpenPose for spatial pose guidance, the sparsity of sparse signals raised great challenges for precise pose control, leading to poor spatial alignment with the given pose. Thus, a critical question raises: *Is it possible to achieve precise and robust pose-controllable generation using sparse signals?*

This paper investigates sparse signals, *i.e.,* OpenPose, for pose-guided T2I generation and identifies two major bottlenecks for precise pose control. First, since the OpenPose representation was originally designed for visualization purposes, its effectiveness for spatial pose guidance remains unclear. As shown in Fig. 3, although OpenPose encodes keypoint locations and topological skeletons to the skeleton image, its color channels provide little information, which would even confuse the perception of distinctive keypoints as discussed in §4.4. Second, the sparsity of pose signals, where keypoints and skeletons are usually denoted as vertices and edges separately, raises challenges for models to perceive and follow point-like spatial instructions [27]. Therefore, extra efforts are required to enhance the focus on keypoint locations and semantics.

In light of this, this paper proposes a spatial-pose ControlNet (*SP-Ctrl*), making it possible for precise pose-guided T2I generation through sparse signals. Specifically, we propose an expressive Spatial-Pose Representation (SPR), which extends the plain RGB keypoint embeddings in OpenPose to learnable embeddings. Besides, to encourage pose alignment, we design a novel Keypoint Concept Learning (KCL) strategy inspired by the associations between textual prompts and cross-attention maps [49]. We introduce new textual tokens to learn keypoint concepts through textual-inversion [9], where a heatmap constraint is derived to encourage attention to each distinct keypoint for spatial alignment. Combining both strategies, our method largely increases the pose alignment of sparse-pose-guided T2I generation with little additional computational cost for inference. Our method realizes performance comparable to dense signals and exhibits superior image diversity and cross-species generalization capability.

Our contributions are summarized as follows.

1) We rethink the sparse signals for pose-guided T2I generation and propose *SP-Ctrl*, a spatial-pose ControlNet that enables precise pose alignment with sparse signals. It reveals the potential of sparse signals in spatially controllable generations.
2) We introduce a Spatial-Pose Representation with learnable keypoint embeddings to enhance the expressiveness of sparse pose signals.
3) We propose Keypoint Concept Learning, a novel strategy that enhances keypoint focus and discrimination, enhancing details and improving pose alignment.
4) Experiments on animal- and human-centric T2I generation validate that our method achieves performance comparable to dense signals. Moreover, our method advances in diversity and cross-species generation.

2. Related Work

Spatial Controllable T2I Diffusion Models. With the popularity of text-to-image diffusion models like Stable Diffusion (SD) [43], recent works explored multi-modal image guidance (*e.g.*, edges, depths, pose, *etc.*) besides textual prompts to achieve spatially controllable synthesis. Early attempts [1, 8, 15] trained models from scratch with spatially aligned image-condition pairs that extracted by pretrained detectors. Another category of works [25, 37, 42, 59, 60] introduced adapters to inject spatial control into the pre-trained T2I models, making it plug-and-play and composable for different conditions. Some efforts also explore training-free approaches [29, 36]. Among these works, ControlNet [59] stands out for its exceptional spatial controllability and high fidelity, and its structure has been widely adopted in downstream applications, particularly for human image generation [6, 20] and animation [14, 52, 54]. This paper aims to enhance ControlNet's capability of interpreting sparse pose signals, enabling diverse outputs and cross-species generation with precise pose alignment.

Guidance for Pose Controllable Generation. As an anatomical abstraction of objects with positions and definitions, pose signals can be provided in various formats to guide spatially controllable T2I generation. Early GAN-based methods [35, 58] and diffusion-based models [13, 33, 37, 46, 59] focused on the popular sparse OpenPose signals to drive image synthesis. However, recent studies revealed the limited guidance of sparse signals and adopted dense representations such as DensePose [10, 21], parsing masks [6, 62], and SMPL [31]. While offering more precise control, these methods increased training complexity and limited the flexibility of creating and editing conditions. Meanwhile, a few works tended to enhance the controllability of sparse pose signals. For instance, HumanSD [20] introduced a perceptual loss with pre-trained pose estimators to improve alignment. Subsequent works explored strengthening sparse signals via textual prompts [26], graph learning [56], and ViT features [16]. In contrast, this paper enriches sparse pose representations and introduces a novel keypoint concept learning strategy, realizing pose alignment without relying on pre-trained pose estimators.

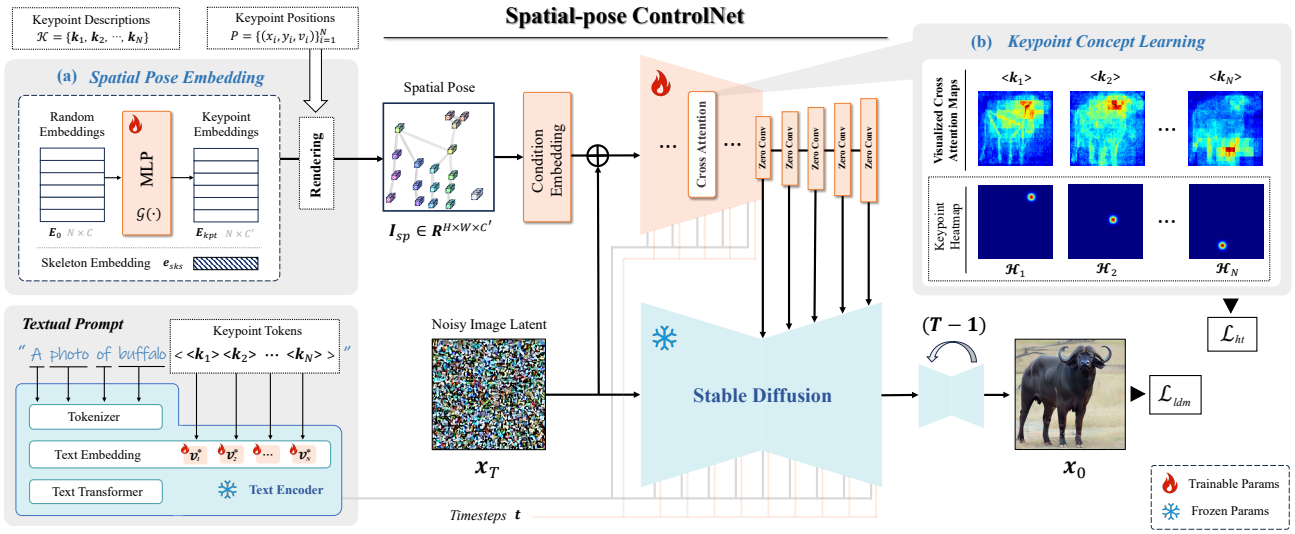


Figure 2. The overall architecture of our spatial-pose ControlNet (*SP-Ctrl*), mainly comprising: 1) a spatial-pose embedding module $\mathcal{G}(\cdot)$ to provide expressive sparse pose representations I_{sp}, 2) a set of new tokens $\{\boldsymbol{k}_i\}_{i=1}^{N}$ to learn the keypoint concept guided by heatmap constraints $\mathcal{L}_{ht}$ on the cross-attention maps, 3) a frozen SD model with a ControlNet adapter as a basic spatial controllable diffusion model. Our proposed model can synthesize images in high fidelity, which align precisely with the given sparse pose signals.

3. Spatial-Pose ControlNet

Given a sparse pose condition $\boldsymbol{c}_p$ and a textual prompt $\boldsymbol{c}_t$, our Spatial-pose ControlNet (*SP-Ctrl*) aims to synthesize images following the prompt with aligned poses. We first reveal two core designs, *i.e.*, Spatial-Pose Representation (§3.1) and Keypoint Concept Learning (§3.2), followed by an instruction of the overall pipeline (§3.3). The architecture of our proposed *SP-Ctrl* is illustrated in Fig. 2.

3.1. Spatial-pose Representation

Since OpenPose [4] was originally designed for visualization purposes, its functions for pose guidance have not been fully examined yet. Therefore, we first revisit OpenPose in the context of pose-guided generations. Given N predefined keypoints $P = \{(x_i, y_i, v_i)\}_{i=1}^{N}$, (x_i, y_i) represents spatial coordinates and $v_i \in \{0, 1, 2\}$ indicates visibility, where $v_i \geq 1$ means the keypoint exists and is labeled in the image. The keypoint descriptions are denoted as $\mathcal{K} = \{\boldsymbol{k}_i\}_{i=1}^{N}$, *e.g.*, eye, nose, elbow, etc. (See Appendix §B for examples) The OpenPose renders skeleton pose images with uniformly sampled RGB colors as Fig. 3, where vertices for keypoints and edges for skeletons. Though it is suitable for visual clarity, this representation ignores the semantic priors of $\mathcal{K}$, causing limited spatial guidance.

To this end, we develop a spatial pose representation (SPR) for sparse-pose-guided generations. While inheriting the setting of OpenPose to render skeleton pose images indicating keypoint and skeleton positions, we extend plain RGB embeddings, representing keypoint and skeleton descriptions, to learnable embeddings inspired by Spa-

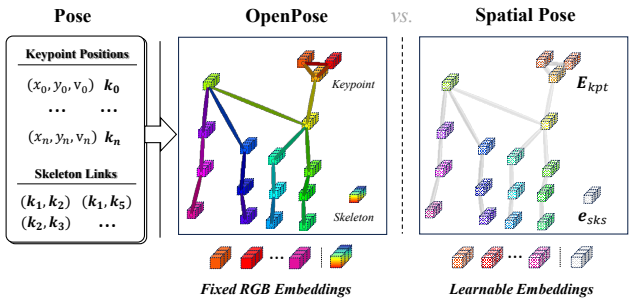


Figure 3. An illustration of OpenPose and our spatial pose.

Text [1]. This helps to encode more semantic insights into the sparse pose signals. Specifically, we employ a spatial-pose embedding module $\mathcal{G}(\cdot;\phi)$, which accepts random initialized vectors $\boldsymbol{E}_0 = \{\boldsymbol{e}_k \in \mathbb{R}^{1\times C}\}_{k=1}^{N}$ as the inputs and produces embeddings for each keypoint, denoted as $\boldsymbol{E}_{kpt} = \{\boldsymbol{e}_k \in \mathbb{R}^{1\times C'}\}_{k=1}^{N}$. This process is expressed as,

$$\boldsymbol{E}_{kpt} = \mathcal{G}(\boldsymbol{E}_0; \phi), \tag{1}$$

where the spatial-pose embedding module $\mathcal{G}(\cdot)$ is an MLP parameterized by ϕ with stacked linear layers and layer norm. The detailed structure is presented in Appendix §A. For skeleton embeddings, we use all-ones embeddings, denoted as $\boldsymbol{e}_{sks} = \boldsymbol{1}^{1\times C'}$, which is proved to work well as discussed in §4.4. Then, we render our spatial-pose representations with the learned embeddings $\boldsymbol{E}_{kpt}$ and $\boldsymbol{e}_{sks}$, resulting in a multi-channel skeleton pose image $\boldsymbol{I}_{sp} \in \mathbb{R}^{H\times W\times C'}$ for pose guidance. Note that the spatial-pose embedding module $\mathcal{G}(\cdot)$ is optimized directly for the training target of denoising diffusion models, thus it learns spatially expres-

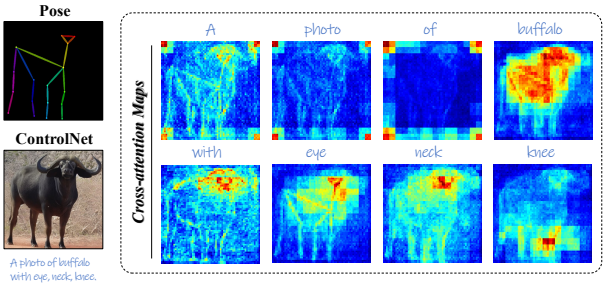


Figure 4. Visualized cross-attention maps of ControlNet at time step 799, where the model attends to positions according to keypoint description. Full results refer to Appendix Fig. S7.

sive keypoint embeddings $\boldsymbol{E}_{kpt}$ adaptively. Validations in §4.5 prove that the learned $\boldsymbol{E}_{kpt}$ effectively encodes semantic priors, where semantic-similar keypoints are pulled together, thus providing richer insight for pose guidance.

3.2. Keypoint Concept Learning

Assuming that the pose alignment relies on both spatial perception and semantic understanding of keypoints, we analyze cross-attention maps to investigate the relevance between keypoint textual prompts and spatial responses in spatially controllable diffusion models, *i.e.*, ControlNet. As illustrated in Fig. 4, consistent with findings of prior studies [11, 23, 49], the cross-attention maps exhibit strong correlations between nouns (*e.g.*, `buffalo`) and pixel positions. Furthermore, the spatial responses of certain keypoint descriptions (*e.g.*, `eye`, `knee`) coarsely align with keypoint positions in OpenPose, especially at early denoising steps. This observation suggests that it is promising to improve pose alignment via strengthening the exact connection between spatial attention maps and each keypoint description.

Following this idea, we propose a novel keypoint concept learning (KCL) strategy to improve alignment with sparse signals. We introduce a series of new keypoint tokens $\{\langle \boldsymbol{k_i} \rangle\}_{i=1}^{N}$ to the text encoder according to descriptions $\mathcal{K} = \{\boldsymbol{k}_i\}_{i=1}^{N}$. These tokens are placeholders of new text embeddings $\mathcal{V}_{kpt} = \{\boldsymbol{v}_i^* \in R^{768}\}_{i=1}^{N}$, representing new keypoint concepts. These new tokens are appended to captions if the corresponding keypoint concept exists in the pose, *i.e.*, $v_i \geq 1$, and we obtain the enriched textual prompt $\boldsymbol{c}_t'$. Then, we extract spatial attention maps from the cross-attention layers. Given the noisy image x_t and textual prompt $\boldsymbol{c}_t'$, the spatial features of the noisy image are projected to a query matrix Q, and the text embedding is projected to a key matrix K and a value matrix V via linear projections. The attention map $\mathcal{M}$ is computed as,

$$\mathcal{M} = \text{Softmax}(\frac{Q \cdot K^T}{\sqrt{d}}), \tag{2}$$

where d is the latent projection dimension of keys and queries. We select the attention maps of keypoint tokens $\mathcal{M}_{kpt} = \{\mathcal{M}_i \in \mathbb{R}^{H' \times W'} | v_i \geq 1\}$ and encourage them to align with the keypoint heatmap $\mathcal{H} = \{\mathcal{H}_i \in \mathbb{R}^{H' \times W'} | v_i \geq 1\}$ [47]. This heatmap constraint $\mathcal{L}_{ht}$ is formulated as,

$$\mathcal{L}_{ht} = \frac{1}{|\mathcal{M}_{kpt}|} \cdot \frac{1}{H'W'} \sum_{v_i \geq 1} ||(\mathcal{M}_i - \mathcal{H}_i)||^2, \tag{3}$$

where $|\mathcal{M}_{kpt}|$ denotes the number of valid keypoints. As discussed in §4.4, we notice that the attention maps from the 3^{rd} transformer block among the $250 \sim 500$ time steps are critical for forming these concepts. Empirical studies in §4.5 prove that the cross-attention maps of keypoint tokens exhibit high responses spatially aligning with the sparse pose signals, thereby improving the pose alignment with sparse pose signals.

3.3. Overall Pipeline

As illustrated in Fig. 2, our *SP-Ctrl* is built on SD [43], consisting of three main components: 1) a *spatial embedding module* to render the spatial pose representation $\boldsymbol{c}_p$, 2) a *text encoder* to encode textual prompts $\boldsymbol{c}_t'$ with learnable text embeddings of keypoints, 3) a *ControlNet adapter* [59] to inject spatial conditions into frozen SD model.

Specifically, the spatial pose representation $\boldsymbol{c}_p$ is first encoded through a conditional embedding module into the latent space. The ControlNet adapter $\mathcal{F}(\cdot;\Theta)$, a trainable copy of the SD encoder, modulates the spatial signals $\boldsymbol{c}_p$ into the diffusion model. Supposing one encoder block $F(\cdot;\theta)$ of SD parameterized by θ, it accepts input feature $\boldsymbol{x}$ and output $\boldsymbol{y} = F(\boldsymbol{x};\theta)$. The ControlNet block $F(\cdot;\theta')$ with parameter θ' injects spatial condition $\boldsymbol{c}_p$ with zero convolution to the SD decoder blocks, which is formulated as,

$$\boldsymbol{y_c} = F(\boldsymbol{x};\theta) + \lambda * Z(F(\boldsymbol{x} + Z(\boldsymbol{c}_p;w_2);\theta');w_1), \tag{4}$$

where $Z(\cdot;w_1)$ and $Z(\cdot;w_2)$ are two 1×1 zero convolution layers with parameters w_1, w_2 initialized with zeros for stable training. $\boldsymbol{y_c}$ is the output feature modulated by spatial pose signals. λ is the *conditioning scale* to adjust the condition strength for inference.

We run the diffusion model in ϵ-prediction mode, which predicts the noise ϵ from the noisy latent $\boldsymbol{x}_t$ at time step t. While training, we optimize the spatial pose embedding module $\mathcal{G}(\cdot;\phi)$, learnable keypoint embeddings $\mathcal{V}_{kpt}$, and the ControlNet adapter $\mathcal{F}(\cdot;\Theta)$ jointly to minimize a combination of the denoising loss $\mathcal{L}_{ldm}$ and heatmap loss $\mathcal{L}_{ht}$. This process is formulated as,

$$\phi^*, \boldsymbol{V}_{kpt}^*, \Theta^* = \arg\min_{\phi, v_i^*, \Theta} \mathcal{L}_{ldm} + \eta \cdot \mathcal{L}_{ht}, \tag{5}$$

where η is a coefficient to balance the two loss items. The denoising loss $\mathcal{L}_{ldm}$ is defined as,

$$\mathcal{L}_{ldm} = \mathbb{E}_{\boldsymbol{x},y,\epsilon,t}[||\epsilon - \epsilon(\boldsymbol{x}_t, t, \boldsymbol{c}_p, \boldsymbol{c}_t';\Theta)||^2], \tag{6}$$

where $\epsilon(\cdot)$ is the denoising network. Since we observe appearance collapse when introducing learnable text embeddings $\mathcal{V}_{kpt}$, we detach the gradient from the noisy image query Q for $\mathcal{L}_{ht}$ to avoid potential information leakage.

4. Experiments

4.1. Datasets and Metrics

AP-10K. AP-10K [57] dataset contains $10,105$ images collected and filtered from 23 animal families and 54 species of mammals. It defines 17 keypoints sharing across all species (see Appendix §B). We adopt the training set for training and conduct evaluation on the validation set. Since AP-10K does not contain textual descriptions, we employ several prompt templates including species and background definitions following [20], as presented in Appendix §B.

Human-Art. Human-Art [19] dataset comprises $50,000$ high-quality human-centric images from 5 real-world and 15 virtual scenarios, including both 2D and 3D scenarios. It provides bounding boxes, keypoint annotations, and textual descriptions. We employ the 17 keypoint definitions in experiments, as presented in Appendix §B. We use the training set for training and report the metrics on the validation set.

Metrics. The evaluation metrics include: 1) *Pose mAP.* We employ the SOTA pose estimator ViTPose++-H [51, 53], pretrained on AP-10K and Human-Art [7], to detect the keypoints of generated images and report OKS-based mAP metric [44] to evaluate pose alignment. The higher mAP means better pose alignment. 2) *CLIP-Score.* We report the category-averaged CLIP-Score (ViT-L/14) [12] to measure the image-text alignment. 3) *FID.* We employ FID to evaluate the image fidelity. 4) *Detection AP*.75. We utilize an open-set object detector, Grounding-DINO [30], to detect the objects in the images by prompting the category name, which reflects the appearance fidelity of generated images. Considering the possible variance of the objects' shape and positions, especially when using sparse signals for pose guidance, we report the AP.75 for reference.

4.2. Implementation Details

We adopt a frozen SD v1.5 [45] as the base model, and the ControlNet adapter is trained from scratch using SD encoder parameters for initialization in all experiments. The learning rate is set to $1e$-5. We randomly drop 50% prompts for classifier-free guidance. We initialize the keypoint embeddings $\boldsymbol{E}_0$ with randomly fixed vectors and select the cross-attention maps from the 3^{rd} transformer block among the $250\sim500$ time steps for computing $\mathcal{L}_{ht}$ as discussed in §4.5. We set $\eta = 0.1$ according to the discussions in Appendix §C. The $\mathcal{V}_{kpt}$ are initialized with their keypoint descriptions following [9]. We train the model for 800 epochs for AP-10K and 300 epochs for Human-Art. To inference, we employ a UniPC [61] sampler with 50 sampling steps with a CFG scale of 7.5 and conditioning scale $\lambda = 1.0$. No negative prompt is used. We generate three images for each prompt with different sampled initial noise for evaluation. All experiments are conducted on NVIDIA A100 40G GPUs.

4.3. Comparison with Other Methods

Performance on AP-10K. The performance is reported in Tab. 1. Our method obtains the best pose mAP and FID, outperforming the baseline ControlNet by 11.38% and 0.88. Note that the newly introduced keypoint tokens $\mathcal{V}_{kpt}$ in §3.2 are new to the pretrained CLIP model, *i.e.*, such new tokens are not included in CLIP model's vocabulary. So, we take the keypoint descriptions, which are employed to initialize such new tokens in §3.2, as an alternative to compute the CLIP-Score for estimation. Our CLIP-Score is comparable to that of ControlNet. The competitive detection AP.75 indicates the faithfulness to animal species. Visualized examples are presented in Fig. 5. Our method can discriminate the keypoint and align with detailed pose strictly, such as paws and nose with even a single point for instruction, while other methods may misinterpret some keypoints and cause misalignment.

Performance on Human-Art. In Tab. 1, our method also achieves promising pose control over human-centric generation tasks with the best pose mAP 51.11—matching the performance of the SOTA GRPose [56]—even without introducing extra pretrained pose estimators for computing perceptive losses during training. Moreover, our method brings little additional computational cost to inference compared to ControlNet. The slight decrement of CLIP-Score is considered acceptable. The marginal discrepancy in FID is supposed to be negligible, and the resulting image quality remains high in Fig. 5. Our method produces images with

Dataset	Method	Pose mAP↑	FID↓	CLIP-Score↑	Detection AP.75↑
AP-10K	Real Image	83.00	—	24.45	50.81
	T2I-Adapter	48.16	27.29	**25.52**	24.23
	ControlNet	44.25	19.40	24.77	24.35
	Ours	**55.63**	**18.52**	24.51	**25.10**
Human-Art	Real Image	79.97	—	26.87	6.78
	T2I-Adapter	38.37	40.01	27.36	6.78
	ControlNet	45.26	**26.69**	27.84	8.18
	HumanSD†	49.92	35.18	27.35	8.29
	GRPose†	50.93	28.85	**27.95**	6.51
	Ours	**51.11**	29.30	26.30	**9.11**

Table 1. Performance comparison with other popular pose-guided T2I methods on AP-10K and Human-Art dataset. † means additional pretrained pose estimators are required when training.

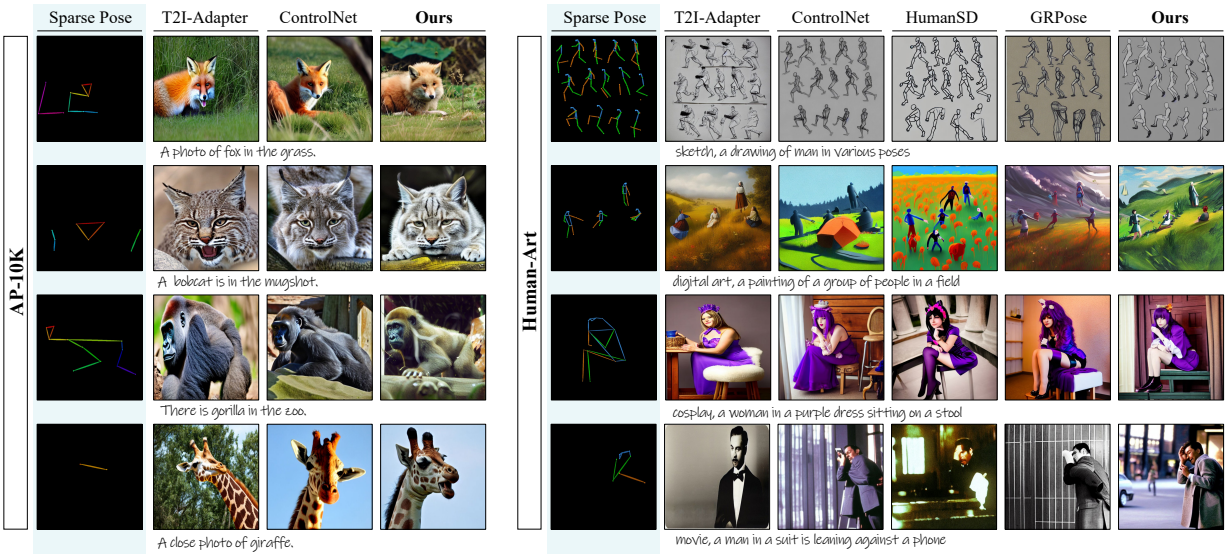


Figure 5. Visualized comparison of our *SP-Ctrl* and other popular methods. See Appendix Fig. S9 and Fig. S10 for more examples.

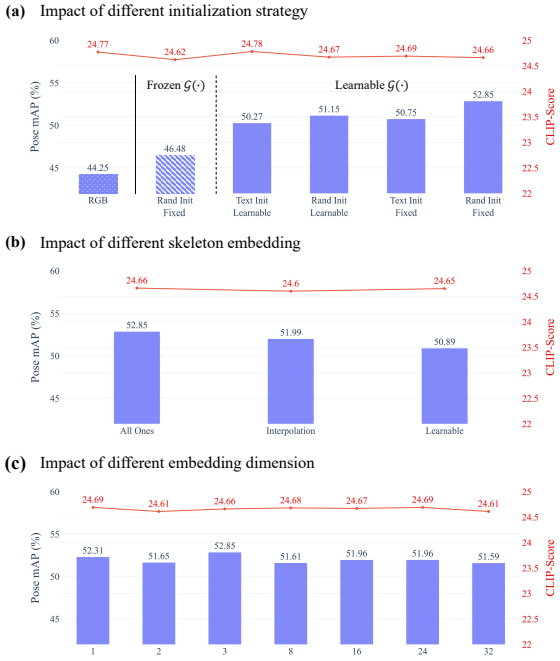


Figure 6. Discussions on the spatial-pose representation.

distinguishable and well-aligned poses, excelling at local structures like arms and legs.

4.4. Ablation Studies

All experiments in this section were conducted on the AP-10K dataset. The baseline is ControlNet with OpenPose.

Discussions on spatial-pose representation. We report the results in Fig. 6, including three parts as follows.

Effects of initialization of $\boldsymbol{E}_0$. An intuitive idea is using text embeddings of keypoint descriptions to initialize $\boldsymbol{E}_0(C{=}768)$, which introduces semantic priors explicitly [1]. We compare this setting with random initialization and compare the learnable and fixed $\boldsymbol{E}_0$ when training. By default, we set the skeleton embedding $\boldsymbol{e}_0 = \mathbf{1}^{1\times C'}$ and the keypoint embedding dimension $C'{=}3$. The baseline is the RGB embeddings of OpenPose. In Fig. 6 (a), the RGB embeddings show the lowest pose mAP, even lower than randomly initialized $\boldsymbol{E}_0$ with a frozen embedding module $\mathcal{G}(\cdot)$, indicating that OpenPose even confuses pose guidance. Using a learnable $\mathcal{G}(\cdot)$ to optimize the keypoint embeddings $\boldsymbol{E}_{kpt}$ brings remarkable improvement on mAP. When we set $\boldsymbol{E}_0$ as learnable, using randomly initialized $\boldsymbol{E}_0$ shows 0.88% higher mAP than that initialized with text embeddings, and so does the fixed $\boldsymbol{E}_0$. This implies possible discrepancies between text-embedding spaces and spatial-pose representations, and the random initialization provides approximately orthogonal vectors for optimizing keypoint embeddings. Meanwhile, fixed $\boldsymbol{E}_0$ brings consistently better mAP than the learnable one owing to optimization stability. The difference in CLIP-Score is negligible. Thus, we employ fixed, randomly initialized $\boldsymbol{E}_0$ with learnable $\mathcal{G}(\cdot)$.

Effects of different skeleton embedding $\boldsymbol{e}_{sks}$. We discuss the impact of different skeleton embedding $\boldsymbol{e}_{sks}$, including **1)** all-one embedding, **2)** interpolation of adjacent keypoint embeddings, and **3)** learnable embeddings. As shown in Fig. 6 (b), it achieves the best mAP when simply setting $\boldsymbol{e}_{sks}$ to all ones, which is 0.86% higher than interpolation and 1.96% higher than learnable vectors. We suppose that the all-ones vector is enough to provide useful spatial guidance like the topology of poses. The difference in CLIP-Score is negligible. So, we set $\boldsymbol{e}_{sks} = \mathbf{1}^{1\times C'}$ by default.

Effects of embedding dimension C'. We further explore the embedding dimension C' of spatial-pose representation. As reported in Fig. 6 (c), we surprisingly find that different settings of C' exhibit competitive performance on both pose

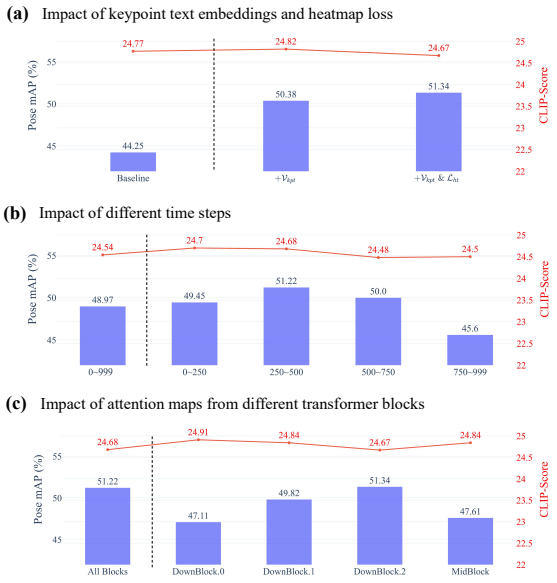


Figure 7. Discussions on the keypoint concept learning strategy.

mAP and CLIP-Sore, and even one-channel keypoint embedding can obtain competitive performance. This reflects the expressiveness of our spatial-pose representation. To compare with Openpose, we adopt $C' = 3$ for consistency. Note that the higher embedding dimension C' would cause increased computation in the condition embedding module.

Discussions on keypoint concept learning. We explore the functions of new keypoint tokens $\mathcal{V}_{kpt}$ and heatmap loss $\mathcal{L}_{ht}$ with OpenPose, which is presented in Fig. 7.

Effects of keypoint embeddings $\mathcal{V}_{kpt}$ and heatmap loss $\mathcal{L}_{ht}$. As shown in Fig. 7 (a), the $\mathcal{V}_{kpt}$ alone improves mAP by 6.13%, which learns the keypoint concepts implicitly through the existence of keypoints. With the heatmap constraint $\mathcal{L}_{ht}$, our method obtains further increments, achieving the best mAP at 51.34. Given that the pretrained CLIP model's vocabulary does not contain the new tokens of $\mathcal{V}_{kpt}$, it would not be feasible to evaluate text-image alignment of $\mathcal{V}_{kpt}$ properly. As an alternative, we take their initialized tokens, *i.e.*, keypoint descriptions, to compute CLIP-Score for estimation. Details refer to Appendix §C. Our CLIP-Score is 24.67, comparable to the baseline 24.77. We also check the metrics of those generated by our method without keypoint tokens, where CLIP-Score is 24.94 and mAP is 45.07. It also validates the importance of new keypoint tokens in pose accuracy. Meanwhile, the visualized results in Fig. 5 maintain faithful text-image alignments, and we suppose the subtle decrease of CLIP-Score is acceptable.

Effects of different time steps. As shown in Fig. 7 (b), different time steps contribute differently to learning new keypoint concepts with $\mathcal{L}_{ht}$ during the diffusion process. The 250∼500 time steps are the most important, and the 750∼999 time steps are useless. The model also achieves the second highest CLIP-Score when implementing $\mathcal{L}_{ht}$ at 250∼500 time steps. This observation provides valuable insight into the role of time steps in forming the concepts of subjects during the denoising process [2, 37].

Effects of attention maps from different transformer blocks. We further explore the functions of attention maps from different transformer blocks of the denoising U-Net encoder, especially those from the 250∼500 time steps. As shown in Fig. 7 (c), the 3^{rd} transformer block of the denoising U-Net encoder, *i.e.*, `DownBlock.2`, contributes mostly, which achieves almost the same pose mAP as that using all blocks. This shows the crucial role of the 3^{rd} transformer block in forming the keypoint concepts. The discrepancy of CLIP-Score is negligible. Thus, we adopt cross-attention maps from `DownBlock.2` during the 250∼500 time steps for computing $\mathcal{L}_{ht}$.

Method	Pose mAP↑	FID↓	CLIP-Score↑
ControlNet	44.25	19.40	**24.77**
+ Spatial Pose	52.85	19.67	24.62
+ KCL	51.34	18.94	24.67
Ours	**55.63**	**18.52**	24.51

Table 2. Ablations of proposed modules, *i.e.*, the spatial-pose representation and keypoint concept learning strategy.

Condition	Method	Pose mAP↑	FID↓	CLIP-Score↑
Mask	ControlNet	26.74	21.15	24.39
Depth	ControlNet	**57.07**	19.85	24.44
Sparse Pose	ControlNet	44.25	19.40	**24.77**
	Ours	55.63	**18.52**	24.51

Table 3. Performance of different conditional guidance.

Ablations for proposed modules. As listed in Tab. 2, both proposed modules demonstrate effectiveness compared to the baseline, where spatial-pose representation improves pose mAP by 8.60%, and keypoint concept learning increases 7.09%. Combining both, our *SP-Ctrl* obtains the best mAP 55.63, showing their compatibility. Our method also gets the best FID and comparable CLIP-Score.

4.5. Discussions

Validations of spatial pose. Fig. 8 shows the distribution of RGB embeddings from OpenPose and the learned embeddings $\boldsymbol{E}_{kpt}$ from our spatial-pose representation after t-SNE [50]. Compared to plain RGB embeddings, the learned $\boldsymbol{E}_{kpt}$ pulls semantic-similar and symmetric keypoints together, thus encoding semantic priors into the channels of sparse signals and making each keypoint expressive. This proves the discriminative guidance of our representation.

Validations of keypoint tokens. We check the cross-attention maps of our method in Fig. 9, especially for these introduced keypoint tokens $\{\langle \boldsymbol{k_i} \rangle\}_{i=1}^{N}$. Under the keypoint-heatmap guidance, these new tokens show high spatial re-

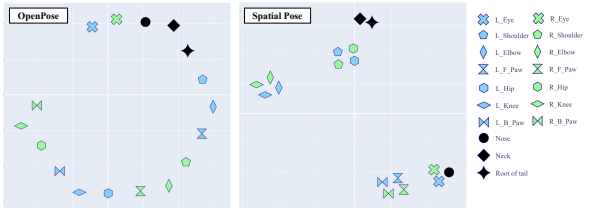


Figure 8. Comparison of the RGB embeddings and learned keypoint embeddings $\boldsymbol{E}_{kpt}$ after t-SNE [50].

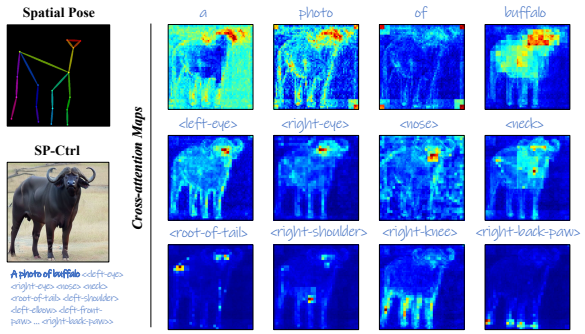


Figure 9. Visualized cross-attention maps of *SP-Ctrl* at time step 399, where the net keypoint tokens $\{\langle \boldsymbol{k_i} \rangle\}_{i=1}^{N}$ attend to spatial positions of keypoints. Full results refer to Appendix Fig. S8.

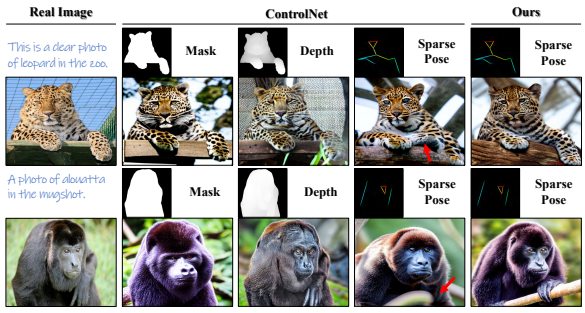


Figure 10. Examples of different guidance, where ours exhibits high pose alignment. More examples refer to Appendix Fig. S11.

sponses at keypoint positions. It encourages pose alignment and increases focus on spatial pose signals. Additional discussions refer to Appendix §C and Fig. S5.

Comparison with other conditions. Tab. 3 reports the results of pose-controllable generation through various signals including masks, depth, and OpenPose. We employ SAM-HQ [22] to extract semantic masks and ZoeDepth [3] for depth maps. Our method surpasses the baseline ControlNet by 11.38% mAP, almost matching the performance of using depth maps. The semantic masks fails to control the exact pose. Visualized results are shown in Fig. 10.

Advantages of sparse-pose guidance. We showcase the potential of sparse-pose signals in several application scenarios. Given the precise pose control with sparse pose, our SP-Ctrl enjoys diverse synthesized images in various shapes (Fig. 11) and cross-species generation capability (Fig. 12). Due to the simplicity of sparse pose, our



Figure 11. Examples of shape diversity, where our method synthesizes images with precisely aligned poses and diverse shapes. More examples refer to Appendix Fig. S12.

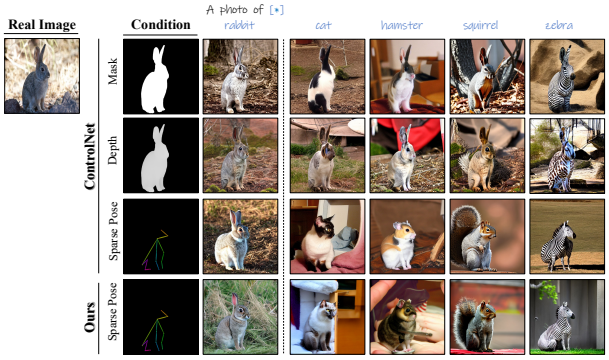


Figure 12. Examples of cross-species pose-guided generation results. More examples refer to Appendix Fig. S13.

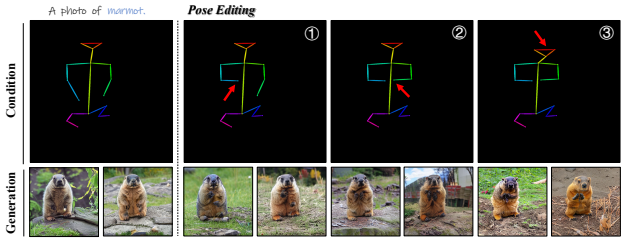


Figure 13. Examples of synthesized images with edited sparse pose signals. More examples refer to Appendix Fig. S14.

method also offers high flexibility in pose editing and allows the creation (Fig. 13) of new poses at a low cost.

5. Conclusion

This paper rethinks sparse signals, *i.e.*, OpenPose, for pose-guided text-to-image generation and proposes *SP-Ctrl*, which realizes precise pose control with sparse signals. To cover the limited guidance of sparse signals, a learnable spatial pose representation is first developed to encode keypoint priors into pose guidance. Additionally, we design a novel keypoint concept learning strategy, which achieves improved pose alignment by enhancing the spatial response of each keypoint token on the cross-attention maps. Extensive experiments demonstrate the effectiveness of our method in pose-controllable generation, even matching the performance of dense signal-based methods. Moreover, our method exhibits the advantages of sparse pose signals in image diversity, cross-species generalization, and pose editing.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC2705700, in part by the National Natural Science Foundation of China under Grants U23B2048, 623B2076, and 62225113, in part by the Innovative Research Group Project of Hubei Province under Grant 2024AFA017, and in part by the Science and Technology Major Project of Hubei Province under Grant 2024BAB046. This project is also supported by the National Research Foundation, Singapore, under its NRF Professorship Award No. RF-P2024-001. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 2, 3, 6

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. Ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 7

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 8

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 1, 3

[5] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *41st International Conference on Machine Learning*, 2024. 1

[6] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. Upgpt: Universal diffusion model for person image generation, editing and pose transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4173–4182, 2023. 2

[7] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 5

[8] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 5

[10] Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22768–22777, 2023. 2

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The 11th International Conference on Learning Representations*, 2023. 4

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 5

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[14] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 2

[15] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 13753–13773, 2023. 2

[16] Wang Jiajun, Morteza Ghahremani, Li Yitong, Björn Ommer, and Christian Wachinger. Stable-pose: Leveraging transformers for pose-guided text-to-image generation. *Advances in Neural Information Processing Systems*, 2024. 1, 2

[17] Le Jiang and Sarah Ostadabbas. Spac-net: Synthetic pose-aware animal controlnet for enhanced pose estimation. *arXiv preprint arXiv:2305.17845*, 2023. 1

[18] Kyungmin Jo and Jaegul Choo. Skip-and-play: Depth-driven pose-preserved image generation for any objects. *arXiv preprint arXiv:2409.02653*, 2024. 1

[19] 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*, 2023. 5

[20] 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. 1, 2, 5

[21] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22623–22633. IEEE, 2023. 1, 2

[22] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in

high quality. In *Advances in Neural Information Processing Systems*, 2023. 8

[23] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 4

[24] Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*, 2024. 1

[25] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2024. 1, 2

[26] Sicheng Li, Keqiang Sun, Zhixin Lai, Xiaoshi Wu, Feng Qiu, Haoran Xie, Kazunori Miyata, and Hongsheng Li. Ecnet: Effective controllable text-to-image diffusion models. *arXiv preprint arXiv:2403.18417*, 2024. 1, 2

[27] 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. 2

[28] Zhiqi Li, Yiming Chen, Lingzhe Zhao, and Peidong Liu. Controllable text-to-3d generation via surface-aligned gaussian splatting. In *International Conference on 3D Vision*, 2025. 1

[29] Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. In *Advances in Neural Information Processing Systems*, 2024. 2

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 5

[31] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. In *The 12th International Conference on Learning Representations*, 2024. 1, 2

[32] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 1

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pages 1–22, 2025. 2

[34] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7085–7093, 2024. 1

[35] 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*, 2020. 2

[36] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2

[37] 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*, pages 4296–4304, 2024. 1, 2, 7

[38] Or Patashnik, Rinon Gal, Daniel Cohen-Or, Jun-Yan Zhu, and Fernando De la Torre. Consolidating attention features for multi-view image editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 1

[39] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 1

[40] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[41] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1505–1514, 2019. 1

[42] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023. 1, 2

[43] 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. 2, 4

[44] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 369–378, 2017. 5

[45] Runway. Stable Diffusion v1-5, 2022. 5

[46] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. Advancing pose-guided image synthesis with progressive conditional diffusion models. In *The 12th International Conference on Learning Representations*, 2024. 2

[47] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 4

[48] Xiaohua Sun and Juexiao Qin. Ai for supporting the freedom of drawing. *Machine Intelligence Research*, 21(1):63–88, 2024. 1

[49] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5644–5659, 2023. 2, 4

[50] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 7, 8

[51] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 5

[52] Yuanfeng Xu, Yuhao Chen, Zhongzhan Huang, Zijian He, Guangrun Wang, Philip Torr, and Liang Lin. Animatezoo: Zero-shot video generation of cross-species animation via subject alignment. *arXiv preprint arXiv:2404.04946*, 2024. 1, 2

[53] 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, 2024. 5

[54] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 1, 2

[55] Wenjie Xuan, Yufei Xu, Shanshan Zhao, Chaoyue Wang, Juhua Liu, Bo Du, and Dacheng Tao. When controlnet meets inexplicit masks: A case study of controlnet on its contour-following ability. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6979–6988, 2024. 1

[56] Xiangchen Yin, Donglin Di, Lei Fan, Hao Li, Chen Wei, Xiaofei Gou, Yang Song, Xiao Sun, and Xun Yang. Grpose: Learning graph relations for human image generation with pose priors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 1, 2, 5

[57] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2, 5

[58] 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. 2

[59] 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. 1, 2, 4

[60] 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, 2024. 1, 2

[61] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 2023. 5

[62] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision*, pages 161–178. Springer, 2022. 2

[63] 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. 1