From Text to Pose to Image: Improving Diffusion Model Control and Quality

Clément Bonnet[*] **Ariel N. Lee** Raive **Franck Wertel** Raive **Antoine Tamano** Raive **Tanguy Cizain** Raive

Pablo Ducru Raive

Abstract

In the last two years, text-to-image diffusion models have become extremely popular. As their quality and usage increase, a major concern has been the need for better output control. In addition to prompt engineering, one effective method to improve the controllability of diffusion models has been to condition them on additional modalities such as image style, depth map, or keypoints. This forms the basis of ControlNets or Adapters. When attempting to apply these methods to control human poses in outputs of text-to-image diffusion models, two main challenges have arisen. The first challenge is generating poses following a wide range of semantic text descriptions, for which previous methods involved searching for a pose within a dataset of (caption, pose) pairs. The second challenge is conditioning image generation on a specified pose while keeping both high aesthetic and high pose fidelity. In this article, we fix these two main issues by introducing a text-to-pose (T2P) generative model alongside a new sampling algorithm, and a new pose adapter that incorporates more pose keypoints for higher pose fidelity. Together, these two new state-of-the-art models enable, for the first time, a generative text-to-pose-to-image framework for higher pose control in diffusion models. We release all models and the code used for the experiments at https://github.com/clement-bonnet/text-to-pose.

1 Introduction

Text-to-image diffusion models have recently shown impressive results in the space of image generation [14, 5, 15, 6, 13]. As the use and quality of these models have increased, so have the requirements from users to strengthen the controllability of the outputs. Various methods have emerged to yield better output control, from fine-tuning [21], to sophisticated prompt-engineering [3, 9], to adding new modalities as conditions, e.g. ControlNets [24] and Adapters [10, 23].

In the context of enhancing human pose fidelity in text-to-image diffusion models, two key problems have emerged:

- (1) Obtaining poses to cover the wide variety of situations that can be described semantically through the text description. Previous methods required selecting poses from a dataset, either by giving a picture, extracting its pose using pose-estimation models such as DWPose [22], and transferring it to the new image with systems such as GANs [8].

[*]Correspondence to: <clement.bonnet16@gmail.com>. Work done while at Raive.

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

- (2) Pose-conditioned image generation such as the previous SOTA SDXL-Tencent adaptor [10] does not include faces or hands, hindering pose fidelity, suffers from an aesthetic quality drastically lower than that of the original SDXL model.

In this article, we overcome these two main challenges of image generation conditioned on text and pose by first training a text-to-pose generative model, which is to the best of our knowledge the first of its kind, and then by training a new SOTA pose adapter for diffusion models, which incorporates both facial and hand gestures. Combining these two new SOTA models enables a new text-to-pose-to-image generative framework for higher pose control in diffusion models (see figure 4 for a visual in the appendix).

2 Text-to-Pose Generative Model

We describe a human pose with a series of points locating key positions of body parts: 18 points for the body, 42 points for the hands, and 68 points for the face (see Figure 5 in the appendix for examples of poses). To train a text-to-pose model, we first design a metric inspired by CLIP [11] to assess the quality of generated poses during training. We then design a transformer architecture for prompt-conditioned pose generation, annotate a dataset of high-quality images, and train the model on it.

2.1 CLaPP: A Contrastive Text-Pose Metric

Before training a generative text-to-pose model, we first create a metric that will guide training regarding matching text to poses. To do so, we train a contrastive model – akin to CLIP [11, 12] – by embedding both text and pose image into a joint latent space and optimizing the projection such as to minimize the angular distance between poses that corresponds to the same text description while maximizing the distance between those that do not. Inspired by CLIP, we call our model *Contrastive Language-Pose Pretraining* (CLaPP).

We train our CLaPP metric on a dataset of 500k (image, prompt) pairs from JourneyDB [16] which we annotate to extract poses (body, face, and hands) using DWPose [22]. Both prompts and poses (images of poses on a black background) are first encoded using CLIP. The CLaPP layers then map each CLIP representation to a new text-pose embedding from which to compute the CLaPP score. The resulting CLaPP scores between different poses and texts are reported in Figure 5 in the appendix. As expected, the diagonal of the score matrix has high CLaPP scores since it corresponds to the exact captions of each image.

2.2 T2P: Text-to-Pose Transformer

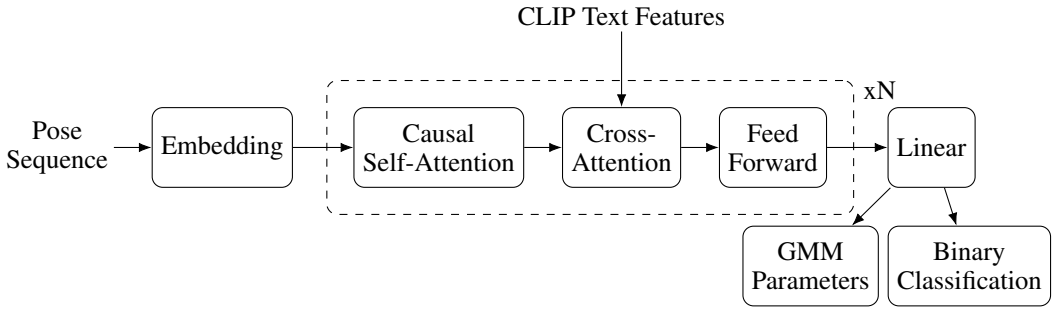


Figure 1: Text-to-Pose transformer architecture.

A pose is defined as an ordered sequence of key points, i.e. (x, y) coordinates of points in the image that correspond to e.g. the right thumb, the left shoulder, the nose, etc. Given this, it makes sense to design a sequence model that can embed the whole pose conditioned on text features. For this, we use a decoder-only Transformer [17] architecture (see fig. 1) to auto-regressively predict the next point in a pose with cross-attention on the text features produced by CLIP. We use positional

embeddings of length 128 (18 body points, 68 face points, and $2*21$ for the hands). Contrary to language models that have tokens, the pose input space is continuous, therefore, we model it with a Gaussian Mixture Model (GMM) along with a binary classifier. The former allows us to learn rather complex non-normal probability distributions over the next pose keypoint, while the latter predicts if the next keypoint exists. Indeed, poses are not always complete, e.g. if a pose describes a portrait, it is likely that only the 68 points corresponding to the face will be present. We found the use of mixtures necessary since a single Gaussian cannot represent well the multi-modal properties of the conditional probability distribution of the next keypoint, e.g. if the pose represents two people on either side, a GMM can capture this bi-modality of where the next point can be (left or right).

2.3 Training

We train the model in a self-supervised fashion by predicting the next pose point in the sequence. The GMM output is trained with maximum likelihood (the GMM likelihood can be computed analytically), and the binary classifier with binary cross-entropy on the existence of the next point or not. We train the model on the 4M (pose, prompt) pairs obtained from JourneyDB by annotating each image with DWPose. We found that a mixture of 6 different Gaussian mixtures and a transformer with 4 layers worked best for our dataset.

2.4 Inference: Tempered Distribution Sampling

After training, the auto-regressive Gaussian mixture model suffers from too high of an entropy which makes sampling from the model quickly diverge out-of-distribution at inference time. Traditionally, language models have their logits divided by a temperature at inference time to decrease the model's entropy [4, 19, 18]. In our case, the model outputs GMM parameters and although the tempered likelihood function can be analytically computed, to our knowledge there is no known algorithm to sample from it. Therefore, we generalize the approach of tempered distribution sampling to any distribution in theorem 1. We then applied tempered sampling to our GMM and found generated poses to be much more precise with lower temperatures, at the expense of being less diverse (see figure 7 in the appendix).

2.5 Analysis & Performance

To analyze the performance of our text-to-pose (T2P) model, we compiled a "COCO-Pose" benchmark dataset, where we extracted the poses of 100 image-caption pairs from the standard COCO 2017 benchmark [7, 1]. For each of the text labels of the COCO-Pose dataset, we then compared the CLaPP score of either: (a) selecting the closest pose from the training dataset (KNN search [2] from CLIP embeddings); or (b) generating a pose using T2P. The results are compiled in figure 2, where we can see that T2P outperforms a KNN in the training dataset 78% of the time. This somewhat proves the local generalization capabilities of T2P and its alignment with respect to prompts describing poses.

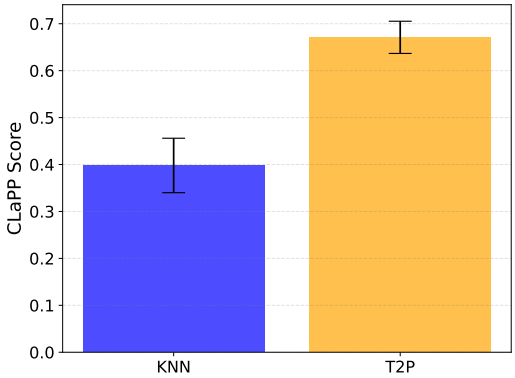


Figure 2: CLaPP scores with 95% confidence intervals. The win-rate ratio of T2P over KNN is 78%. We use a subset of 100 (caption, pose) pairs from the COCO 2017 validation dataset

3 Pose Adapter for Image Generation

Several pose-conditioning systems have been built for diffusion models, notably ControlNets [24], and Adapters [10, 23]. The previous SOTA model for pose-conditioning is the SDXL-Tencent adapter [10] which suffers from two key drawbacks. First, it does not include keypoints related to faces and hands. Second, it suffered from lower image aesthetic due to training on low-quality images or optimization issues.

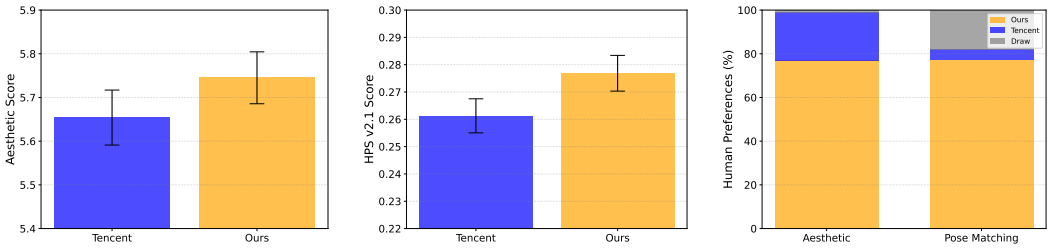
3.1 Adapter Training

We start from the same architecture as the previous SOTA adapter [10], which consists of a ResNet-like model, and we train it on the same dataset used to train T2P. We use the same hyper-parameters as the original model for a total of 7600 training steps with a batch size of 256.

3.2 Analysis & Performance

To measure the performance of our adapter compared to the previous SOTA, we conducted a series of benchmarking tests. For these tests, we generate pose-conditioned images using 30 steps of diffusion with both adapters and then 10 steps of the SDXL refiner model to correct small distortions. The fact that our adapter now has faces and hands to condition on yields better pose accuracy, as shown in figure 8 in the appendix. However, one can see the poses are never perfectly matched and the image quality is still below that of the base SDXL model.

We also measure the aesthetic score [2], and Human Preference Score (HPS) v2 [20], of both the SDXL-Tencent adapter and ours on the COCO-Pose dataset. As shown in figure 3, our adapter outperforms the Tencent one 70% of the time for the aesthetic score, and 76% of the time for the HPS score on the COCO-Pose benchmark.



(a) Aesthetic Score, win ratio: 70% (b) HPS v2, win ratio: 76% (c) Human Preferences

Figure 3: Performance of pose-conditioned image generation for the Tencent adapter and ours. **(a)**: Aesthetic score (ML based). **(b)**: Human Preference Score v2 [20]. **(c)**: Human preferences (manually annotating). Error bars represent two standard deviations. We use a subset of 100 (caption, pose) pairs from the COCO 2017 validation dataset to serve as conditions for image generation.

4 Discussion

In this article, we tackle the problem of conditioning on human poses the image generation process of diffusion models. We first solve the issue of finding the right pose given a prompt by training a text-to-pose auto-regressive model, which outperforms searching the training database (using KNN) 78% of the time. We then tackle the task of conditioning on poses in a way that guarantees both high fidelity to the pose and high image aesthetics by training a pose adapter on high-quality images with more pose keypoints than previous models. Our adapter has a win-rate ratio of 70% on the aesthetic score and 76% on the COCO-Pose benchmark compared to the previous SOTA.

This work encourages new paradigms for improved user experiences. By creating this intermediate modifiable image semantics (human poses), one can imagine slightly altering the pose while keeping the seed constant so as to modify the human pose directly on the generated image, turning the pixel image somewhat vectorized.

Improved image fidelity and human pose control bring our attention to the broader ethical impacts that such technologies can have. While our system does not guarantee photorealism, it is important to keep in mind that such AI-generated images should not be misused and mixed up with real photos of humans.

[2] see `https://github.com/christophschuhmann/improved-aesthetic-predictor`

References

[1] COCO 2017 Dataset.

[2] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, Jan. 1967.

[3] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[4] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network, Mar. 2015. arXiv:1503.02531 [cs, stat].

[5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[6] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the Design Space of Diffusion-Based Generative Models, Oct. 2022. arXiv:2206.00364 [cs, stat].

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common Objects in Context, Feb. 2015. arXiv:1405.0312 [cs].

[8] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose Guided Person Image Generation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[9] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[10] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models, Mar. 2023. arXiv:2302.08453 [cs].

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, Apr. 2022. arXiv:2204.06125 [cs].

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[14] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, June 2015. ISSN: 1938-7228.

[15] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations, Feb. 2021. arXiv:2011.13456 [cs, stat].

[16] K. Sun, J. Pan, Y. Ge, H. Li, H. Duan, X. Wu, R. Zhang, A. Zhou, Z. Qin, Y. Wang, J. Dai, Y. Qiao, L. Wang, and H. Li. Journeydb: A benchmark for generative image understanding, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[18] C. Wang, S. X. Liu, and A. H. Awadallah. Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference, Aug. 2023. arXiv:2303.04673 [cs].

[19] P.-H. Wang, S.-I. Hsieh, S.-C. Chang, Y.-T. Chen, J.-Y. Pan, W. Wei, and D.-C. Juan. Contextual Temperature for Language Modeling, Dec. 2020. arXiv:2012.13575 [cs].

[20] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[21] Z. Xiao, K. Kreis, and A. Vahdat. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs, Apr. 2022. arXiv:2112.07804 [cs, stat].

[22] Z. Yang, A. Zeng, C. Yuan, and Y. Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023.

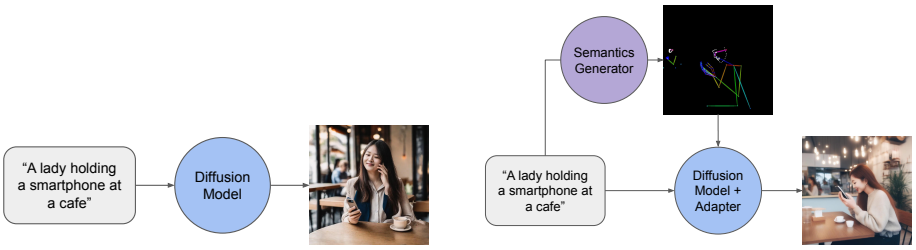
[23] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, Aug. 2023. arXiv:2308.06721 [cs].

[24] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models, Oct. 2023. arXiv:2305.16322 [cs].

A Appendix

A.1 Text-to-pose-to-image framework

We propose factorizing text-to-image generation into semantics generation (human poses) and then semantics-conditioned image generation (see figure 4). This allows for better control over the semantics and higher overall quality.



(a) Standard *text-to-image* generation (b) Ours: *text-to-pose-to-image* generation

Figure 4: Text-to-pose-to-image framework.

A.2 CLaPP metric

Our contrastive model, called CLaPP, can predict compatibility scores between a prompt and a human pose. Figure 5 demonstrates a few CLaPP scores of samples from the COCO dataset. The closer to 1, the more similar a pose and a prompt are, the worst score being -1.

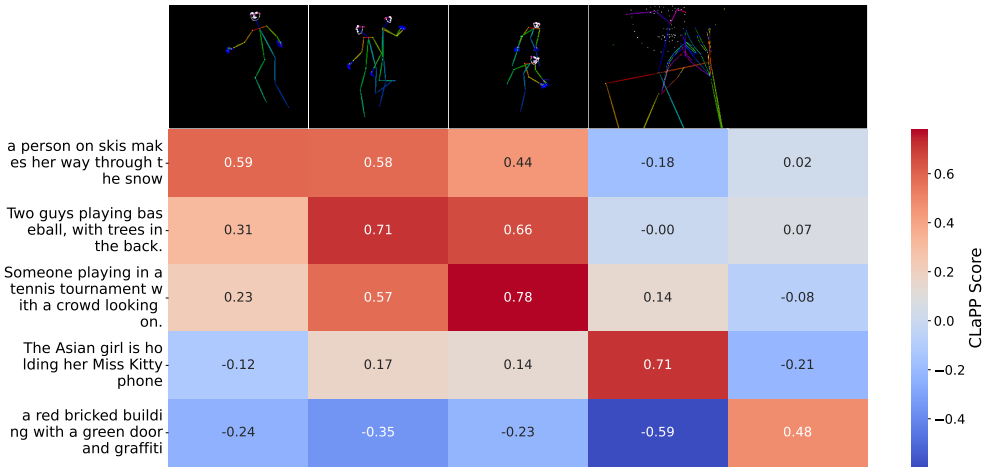


Figure 5: CLaPP scores on 5 poses and corresponding captions from the COCO dataset. The scores measure the compatibility between text and poses.

A.3 Tempered distribution sampling

We prove some properties of our generalization of tempered sampling to any distribution in theorem 1. We also describe the sampling algorithm used to do inference with the T2P model.

Theorem 1 (Tempered distribution sampling) *Let X be a random variable, with a probability distribution density $p(X)$, i.e. $X \sim p(X)$. Let $T \in \mathbb{R}_+$ be a real positive "temperature" parameter. We define the "tempered distribution transform" as $X_T \sim p_T(X)$, where:*

$$p_T(x) \triangleq \frac{p(x)^{\frac{1}{T}}}{\int_X p(x)^{\frac{1}{T}} \mathrm{d}x} = \frac{\mathrm{e}^{\frac{\ln\, p(x)}{T}}}{\int_X \mathrm{e}^{\frac{\ln\, p(x)}{T}} \mathrm{d}x} \tag{1}$$

Properties – the tempered distribution p_T is related to the original one p by the following properties:

- *Conservation of modes (and inflection points): if $\partial_x p = 0$, then $\partial_x p_T = 0$*
- *Temperature one invariance: $p_{T=1}(x) = p(x)$*
- *Mode selection with smaller temperatures: as T goes to zero, the tempered distributions p_T tends towards a dirac distribution selecting the mode of the p distribution, i.e.*

$$p_T(x) \underset{T \to 0}{\longrightarrow} \max_{p(x)} \delta(x) \tag{2}$$

- *Uniform thermalization with high temperatures: as T goes to infinity, the tempered distribution p_T tends towards a uniform distribution over the support of p, i.e.*

$$p_T(x) \underset{T \to \infty}{\longrightarrow} \underset{X}{\mathrm{Uniform}}(x) \tag{3}$$

- *Temperature score scaling: the score of the tempered distribution $\nabla_x \ln p_T(x)$ is proportional to the score of the original distribution $\nabla_x \ln p(x)$, scaled by the temperature, i.e.*

$$\nabla_x \ln p_T(x) = \frac{\nabla_x \ln p(x)}{T} \tag{4}$$

Sampling scheme – to sample from the tempered distribution, one can (c.f. Figure 6):

- *Sample N points x_i from $p(X)$ to cover its support*
- *Sample from the following softmax distribution:*

$$\mathrm{Softmax}\left\{\left[\frac{1}{T} - 1\right] \ln p(x_i)\right\} \triangleq \frac{\mathrm{e}^{\left[\frac{1}{T}-1\right]\ln p(x_i)}}{\sum_{j=1}^{N} \mathrm{e}^{\left[\frac{1}{T}-1\right]\ln p(x_j)}} \tag{5}$$

Proof 1 *The proofs of the mode-conserving and score-scaling properties are established by taking the derivatives from the tempered distribution definitions. The mode-selecting property is established by taking the temperature to zero. The proof of the sampling scheme is derived by importance sampling: $\int_X \frac{\mathrm{e}^{\frac{\ln p(x)}{T}}}{p(x)} p(x) \mathrm{d}x = \int_X \mathrm{e}^{\left[\frac{1}{T}-1\right]\ln p(x)} p(x) \mathrm{d}x$, which is then approximated at large N by Monte Carlo estimate $\sum_{j=1}^{N} \mathrm{e}^{\left[\frac{1}{T}-1\right]\ln p(x_j)}$.*

We can see in figure 6 the influence of the temperature on a tempered GMM in a toy case with two Gaussian mixtures. The closer to 0 the temperature, the closer to the mode the tempered distribution peaks.

Tempered sampling for a GMM is crucial for our T2P model because the model with a temperature of 1 has too high entropy and ends up generating out-of-distribution high-noise pose sequences. The effect of the temperature on the T2P inference can be observed in figure 7 with the poses failing to shake hands at high temperature but eventually succeeding if the temperature is lowered.

A.4 Pose adapter

We train our adapter on high-quality images annotated with full poses (including bodies, hands, and faces). This leads to a pose-conditioned image generation of higher quality and fidelity as shown in figure 8.

A.5 Limits

CLaPP metric The metric we designed to automatically assess the quality of generated poses has some shortcomings. First, it was trained on a not-so-large dataset which means it may consider out-of-distribution captions or poses that are quite different from the dataset it was trained on, biasing a text-pose matching score. Then, we used CLIP as a backbone, whose text and image encoders may have some pose-agnostic representations that hinder the quality of embeddings as much as human poses are concerned. Finally, using images as representations of poses is highly inefficient, one could design a better model by working with the sequence of (x, y) points directly (much lower dimension), for instance using the backbone of the T2P model.

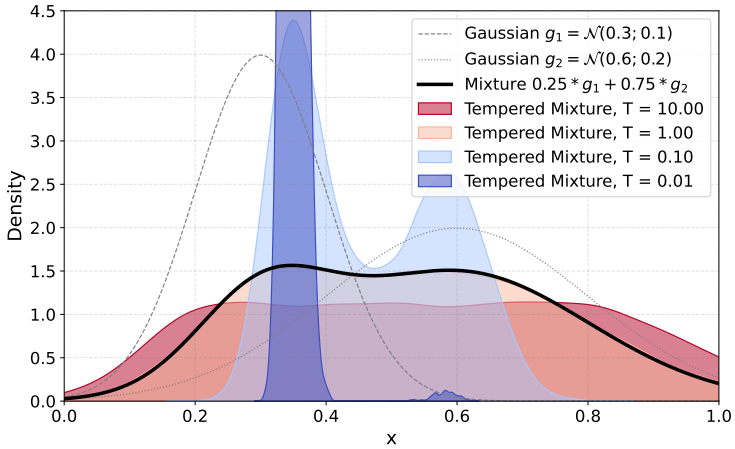


Figure 6: Tempered distribution sampling of Gaussian Mixture Model. $N = 10,000$ Monte-Carlo samples.

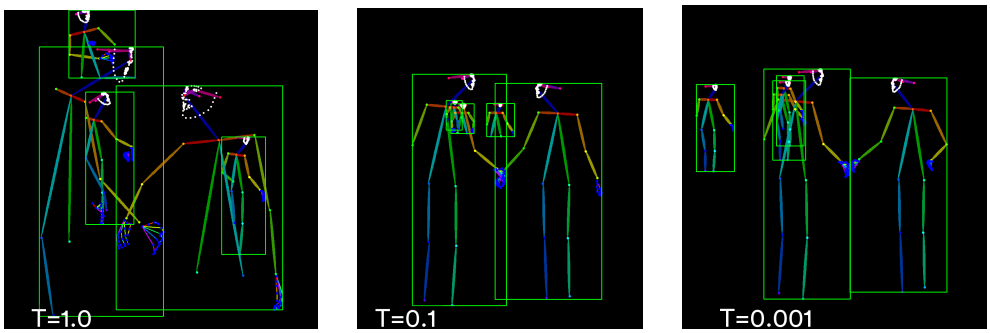


Figure 7: Effect of the temperature on GMM tempered sampling for text-to-pose generation. Prompt: "two politicians shaking hands in a lobby".

T2P Although the trained T2P model can generate rather precise poses, it tends to lack diversity and is really just a reflection of the data it was trained on. Moreover, generating poses in an auto-regressive fashion is quite costly and adds considerable overhead at inference time.

Pose adapter Although our newly trained pose adapter has higher pose fidelity and image aesthetics, it does not always conform exactly to pose conditions and still suffers from lower image quality compared to the base SDXL. More data would likely help mitigate both these issues.

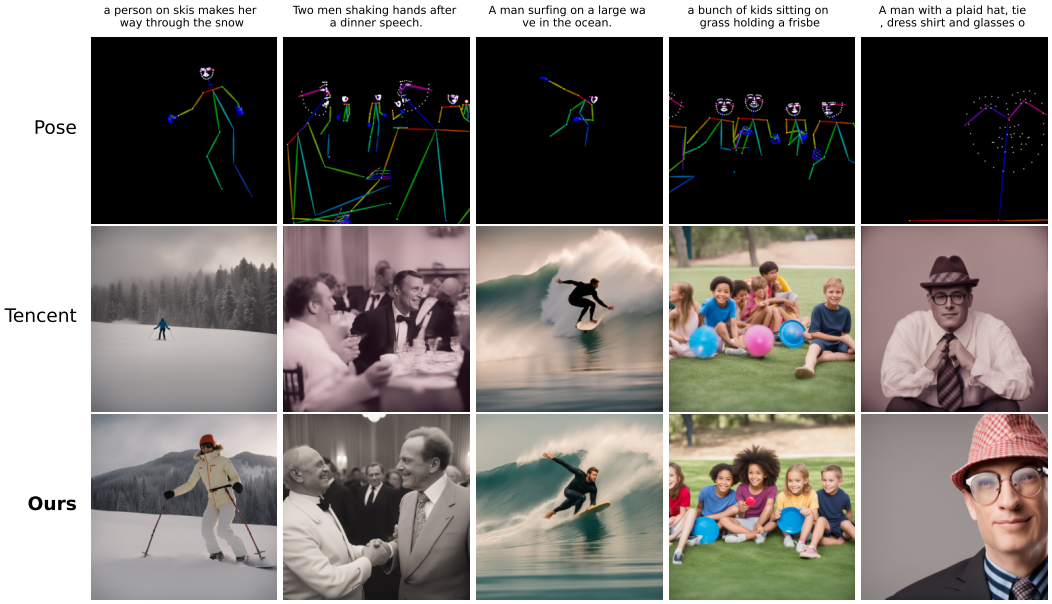


Figure 8: Pose-conditioned image generation using the SDXL-Tencent adapter, and ours (which includes hands and faces).