SKETCH-CONDITIONED SUPERIORITY: EMPIRICAL COMPARISON OF IMAGE-TO-IMAGE AND TEXT-TO-IMAGE GENERATION ACCURACY

Grebennik Igor, Krasnikov Vlad

System Engineering Department,
Kharkiv National University of Radio Electronics,
Ukraine

E-mail: igor.grebennik@nure.ua
vlad.krasnikov@nure.ua

Language Adviser: Associate Professor Oleg Storchak

Abstract

У статті досліджується практична перевага генерації зображень на основі скетчу над суто текстовою генерацією у дифузійному синтезі зображень. Ми порівнюємо стандартний режим text-to-image із режимом image-to-image, який використовує мінімалістичний силуетний скетч як умову, за однакових налаштувань семплінгу в моделі Stable Diffusion XL. Якісна експертна оцінка, доповнена невеликою пілотною перевіркою на основі CLIP-схожості, свідчить, що використання скетчу забезпечує стабільнішу позу, пропорції та просторове розташування ключових об'єктів, водночас зберігаючи стилістичне різноманіття базової генеративної моделі. Результати підкреслюють практичну значущість робочих процесів зі скетч-кондиціонуванням для індустріального та ігрового дизайну, технічної ілюстрації, прототипування інтерфейсів користувача та окреслюють напрями для майбутніх досліджень кількісних метрик, орієнтованих на геометрію

Introduction

Modern generative artificial intelligence models make it possible to synthesize images with a high degree of photorealism based on a textual description of a scene. However, in practical use of the text-to-image paradigm, the operator inevitably encounters the inherent subjectivity of natural-language prompts and semantic under-specification, which together result in highly variable and not always predictable image compositions, as well as systematic errors in object scale, pose, and spatial arrangement.

In recent years, generative models based on diffusion architectures have become a key tool in digital design, multimedia, and content engineering. Their ability to produce visually coherent and stylistically diverse images from textual scene descriptions has significantly expanded the capabilities of both professional artists and users with no artistic background. However, as the visual quality of synthesized images has increased, a fundamental limitation of text-to-image generation has become more apparent: the high sensitivity to prompt phrasing and the strong dependence of the result on the model's internal stochastic processes. These factors lead to ambiguity in image composition making it considerably more difficult to obtain images with precise object placement, which is critical in applications such as industrial design, technical visualization, and UI prototyping.

In response to this problem, methods for controllable generation are being actively developed, allowing additional constraints to be imposed on top of the textual description. One of the most effective approaches is the use of a sketch as a structural prior, which specifies the geometry of the scene and provides the model with essential information about object pose, scale, and overall composition. Sketch-conditioned image-to-image generation combines the flexibility of diffusion models with the advantages of a rigid compositional scaffold, substantially reducing the likelihood of shape distortions or violations of spatial relationships between objects. As a result, this approach becomes particularly relevant for tasks that simultaneously require the creative potential of the model and strict structural accuracy.

Taken together, these factors make the comparative analysis of text-to-image and sketch-conditioned image-to-image regimes an important component in the study of contemporary generative AI methods. Understanding the differences between them makes it possible to optimally select tools for specific practical applications, as well as to formulate recommendations for improving the controllability and predictability of generation outcomes.

The aim of this work is to demonstrate that using a preliminary sketch in the image-to-image regime makes it possible to reduce compositional variability, increase structural accuracy (in terms of object shape, pose, and spatial arrangement), and at the same time preserve the stylistic freedom characteristic of generative models. In this context, a sketch is understood as a simple contour or silhouette image that specifies the geometry of the scene and its key elements, but does not impose a final artistic style.

Experimental Methodology

For the comparison, we employed two operating modes of a single diffusion model: (1) text-to-image generation based solely on the textual description, and (2) sketch-conditioned image-to-image generation using the same description, augmented with a minimalist sketch of the target scene. The sampling parameters, image resolution and seed were fixed for different steps (i.e. best result for base generation was at 5 CFG Scale and was fixed at that point). In the image-to-image mode, the noise coefficient (denoising strength) was varied in order to find a balance between preserving geometry and maintaining the model's freedom of stylistic variation.

Experimental Setup

To compare generation quality, we selected two operating modes of a single Stable Diffusion–based architecture:

(1) text-to-image generation from a textual prompt;

(2) sketch-conditioned image-to-image generation from the same prompt, augmented with a silhouette sketch of the target composition.

In both modes, we used identical sampling parameters (sampling steps), the same denoising algorithm, image resolution, and a fixed set of random seeds to ensure a fair comparison of the results. The basic and extended textual prompts used in the experiment are shown in Fig. 1 and Fig. 2, respectively.

Generation Parameters

In the text-to-image mode:

- Number of diffusion steps: 29;
- Sampler algorithm: DPM++ 2M Karras;
- CFG scale: 5;
- Resolution: 832 × 1216 px.

These parameters correspond to a standard procedural protocol that provides a balance between image quality and stylistic variability.

In the image-to-image mode, an additional minimalist sketch was supplied — a knight silhouette (Fig. 3), serving as a structural prior.

The key parameter is the **denoising strength**, which determines the degree to which the original structure is preserved:

- 0.2–0.3 - almost complete preservation of pose and geometry;
- 0.4–0.5 - a balance between structural fidelity and stylization;
- above 0.6 - noticeable deviations from the initial sketch.

During the experiment, we were able to obtain a successful result with a denoising strength of approximately 0.9, which futher confirmed the hypothesis even under a setting that allows the model to deviate substantially from the original sketch.

Experimental Procedure

1. For each text prompt, 10 images were generated in the T2I mode (text-to-image) with a fixed set of random seeds.
2. The same seeds were then reused in the I2I mode (image-to-image) conditioned on the silhouette sketch.
3. Each pair of results (T2I and I2I) was compared with a pre-prepared reference layout of the composition (a contour drawing of the target scene).
4. An expert group (3 participants) evaluated the accuracy of correspondence on a five-point scale, separately for the following criteria:

- ○ pose consistency;
- ○ correctness of proportions;
- ○ placement of key objects (sword and shield);

- ○ preservation of the overall composition.

5. The automatic metrics included CLIP similarity between the reference and the generated image;

Examples of I2I generation results for the text-only condition and for the text-plus-sketch condition are shown in Fig. 4 and Fig. 5.

Experimental Design Explanation

This protocol makes it possible to:

- minimize the influence of randomness by fixing the random seeds;
- isolate and assess the impact of a single factor, namely the presence of a structural prior;
- compare the results both visually and quantitatively;
- measure the model's robustness to prompt variability.

Parameters and Metrics Description

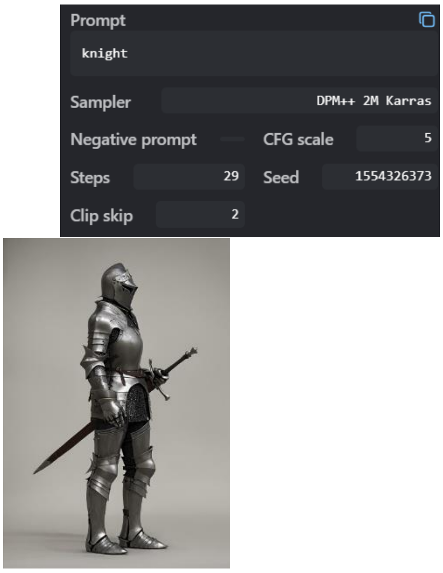
All experiments were conducted using the Stable Diffusion XL model. This choice was largely motivated by personal experience and availability, yet the setup is representative for modern diffusion models in general: their main differences lie in training data and optimization details rather than in the underlying sampling mechanics.

Diffusion models are a family of generative models trained on pairs of clean images and their corrupted, noise-perturbed versions. During training, the model learns to progressively denoise an input so that the result matches the original image. After training, the model can start from random noise and iteratively denoise it to synthesize novel images.

In this work we focus on three key configuration parameters:

1) Diffusion steps.
The number of steps used to iteratively "denoise" the latent representation. At each step the model slightly reduces the noise level and sharpens the emerging structure. Too few steps may result in artifacts and unstable geometry, whereas very large step counts increase generation time with diminishing returns in quality. In our experiments we used 29 steps, chosen empirically based on prior experience and the model authors' recommendations.
2) Sampler (sampling algorithm).
A numerical procedure that specifies how the model traverses the trajectory from pure noise to the final image (e.g. Euler, DPM++, Heun). Different samplers distribute the steps along the noise schedule and approximate the diffusion process in different ways, which affects noise level, detail stability, and smoothness. We used the DPM++ 2M Karras sampler as a compromise between image quality and speed.
3) CFG scale (classifier-free guidance scale).
A coefficient that controls the balance between strict adherence to the text prompt and the model's generative freedom. Lower CFG values tend to produce more diverse images that may deviate from the prompt, whereas higher values enforce the prompt more strongly but can introduce artifacts and reduce realism. In our setup we used a moderate guidance scale (around 5), which provides a reasonable trade-off between faithfulness to the description and visual plausibility.

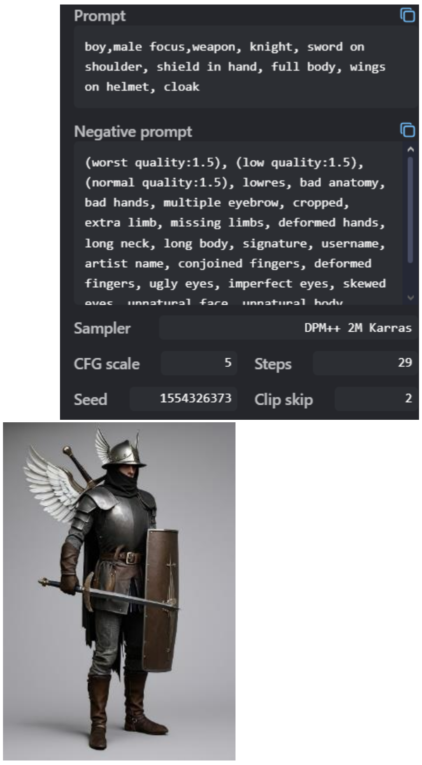
As for evaluation, it is not straightforward to measure generation quality in a purely objective way. The primary focus of this study is therefore on visual expert assessment of pose consistency, proportions, and the placement of key elements such as the weapon, shield, and armor details. In addition, we performed a small pilot, mathematics-based check using CLIP-based similarity. A short Python script was used to compute cosine similarity between CLIP embeddings of the reference sketch and the generated images, providing a single numerical score for each image. In the illustrative example considered, all modes achieved high similarity values, however highest score was obtained for the image-to-image generation with a sketch (~0.83 versus ~0.77 for the corresponding text-only result), which is consistent with our qualitative observations. However, given the small sample size and the limited sensitivity of CLIP to fine-grained geometry, this analysis should be regarded as preliminary. In future work we plan to construct a larger and more diverse dataset and to formalize geometry-oriented metrics (e.g. keypoint alignment and contour overlap) in order to more rigorously demonstrate the advantages of sketch-conditioned generation.



Prompt
```
knight
```

Sampler: DPM++ 2M Karras
Negative prompt:
CFG scale: 5
Steps: 29
Seed: 1554326373
Clip skip: 2

Fig. 1 – Base prompt generation example (T2I).



Prompt
```
boy,male focus,weapon, knight, sword on
shoulder, shield in hand, full body, wings
on helmet, cloak
```

Negative prompt
```
(worst quality:1.5), (low quality:1.5),
(normal quality:1.5), lowres, bad anatomy,
bad hands, multiple eyebrow, cropped,
extra limb, missing limbs, deformed hands,
long neck, long body, signature, username,
artist name, conjoined fingers, deformed
fingers, ugly eyes, imperfect eyes, skewed
eyes, unnatural face, unnatural body
```

Sampler: DPM++ 2M Karras
CFG scale: 5
Steps: 29
Seed: 1554326373
Clip skip: 2

Fig. 2 – Advanced prompt generation example (T2I).

Result evaluation

Accuracy was assessed using a combination of visual expert evaluation and automated metrics. The experts rated the correspondence of the generated results to the target layout on a scale from 1 to 5. In addition, we analyzed semantic similarity based on CLIP embeddings, the preservation of key elements (including presence and position of the sword and shield). Empirically, we observed that in the text-to-image mode the model tends to exhibit deviations in pose and proportions, whereas sketch-conditioned image-to-image generation provides more stable geometry while still allowing for stylistic diversity.

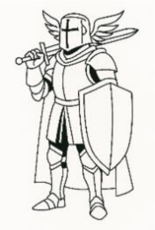


Fig. 3. Minimal sketch of a knight to be used as a template

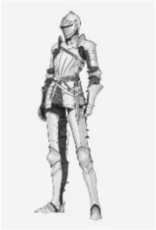


Fig. 4. Image-to-image sketch based result with simple prompt (I2I).

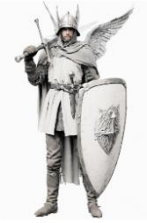


Fig. 5. Image-to-image generation result with advanced prompt (I2I).

Conclusion

As a result of the study, it was established that sketch-conditioned image-to-image generation demonstrates practical superiority in the accuracy of placement and shaping key objects compared to purely text-based generation. A minimalist silhouette is sufficient to control the pose and the type of object, making the approach accessible to users without artistic training.

This approach appears promising for applications in industrial and game design, technical illustration, and user interface development, where controlled geometry and composition are required. In future work, we plan to formalize quantitative metrics (CLIP similarity, keypoint/edge overlap, structural similarity) and to compare different architectures for controllable generation (ControlNet, T2I-Adapters, and others).

Reference List

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

2. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS).

3. Podell, D., et al. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

4. Zhang, L., et al. (2023). Adding conditional control to text-to-image diffusion models (ControlNet). arXiv preprint arXiv:2302.05543.

5. Saharia, C., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding (Imagen). arXiv preprint arXiv:2205.11487.