# Attention-Enhanced Multi-ControlNet for Artist-Aligned Manga Background Generation

Louis King Deblina Bhattacharjee
University of Bath
Department of Computer Science
{lk943, db2648}@bath.ac.uk

## A. Supplementary Material

### A.1. Additional Qualitative Results

Fig. 1 provides additional qualitative comparisons across all inference variants on diverse manga scenes from Manga109 [3] and AMP-D [4]. AttnMCN consistently produces the highest-fidelity outputs across scenes with varying complexity, from simple character vignettes to dense multi-character compositions. The No-Attention MCN [2] variant (column 1) produces noisy, fragmented outputs with severe colour artefacts. Single Tile ControlNet [5] (column 2) and single Canny ControlNet [5] (column 3) both produce heavily blurred, largely featureless images, confirming that a single conditioning signal provides insufficient guidance. Our AttnMCN (column 4) reconstructs recognisable scene content with faithful structural and tonal alignment to the conditioning inputs.

### A.2. Extended Qualitative Grid

Fig. 2 shows a second set of qualitative comparisons, including highly detailed cover-art-style images that challenge all methods. AttnMCN maintains clear character outlines, text readability, and background detail even in complex compositions.

### A.3. Image-to-Image Baseline: Extended Comparison

Fig. 3 provides additional I2I baseline comparisons. The Foreground I2I column uses the foreground as image input with an auto-generated background caption; the Background I2I column uses the background image with a foreground caption. In all cases, the text-conditioned approaches produce outputs that diverge substantially from the intended scene layout, hallucinate content absent from the input, and fail to maintain consistent style. AttnMCN avoids these failure modes by conditioning exclusively on visual inputs.

### A.4. Additional I2I Examples

Fig. 4 shows further I2I comparison examples with varied scene types.

### A.5. Training Curves: Baseline Comparisons

Fig. 5 shows LPIPS [6] training curves for all baseline comparisons. The I2I baseline (a) converges to a much higher LPIPS loss (0.45–0.55) than AttnMCN. Single Tile ControlNet [5] (b) and Single Canny ControlNet [5] (c) both converge above AttnMCN but below the I2I baseline. The No-Attention MCN [2] (d) oscillates at higher LPIPS values (0.35–0.55) while AttnMCN stably achieves 0.15–0.25.

### A.6. FID Comparison Across Baselines

Fig. 6 presents the FID [1] comparison across baselines. AttnMCN achieves 35.1, compared to 100.3 (Single Tile), 133.3 (Single Canny), and 116.3 (No-Attention MCN), representing a 65%–74% reduction in distributional distance from real manga.

### A.7. Sample Generated Outputs

Figs. 7a and 7b show two high-quality outputs of AttnMCN, demonstrating that the pipeline can generate manga images with recognisable character detail, consistent lighting, and coherent backgrounds.

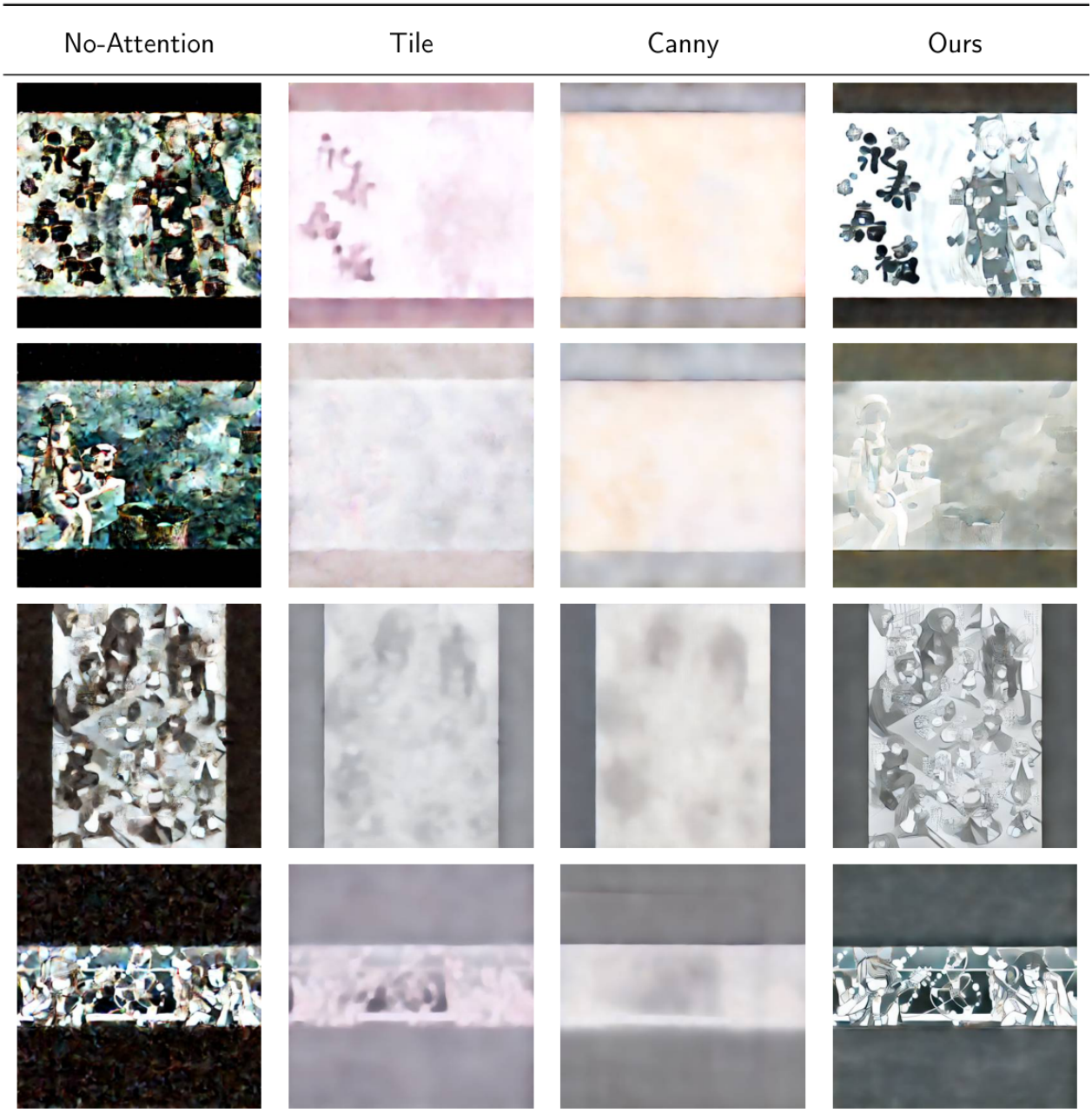



Figure 1. **Additional qualitative comparisons.** Columns: No-Attention MCN [2], Single Tile [5], Single Canny [5], AttnMCN (Ours). Rows show diverse scenes. AttnMCN preserves structural detail and colour coherence across all examples.

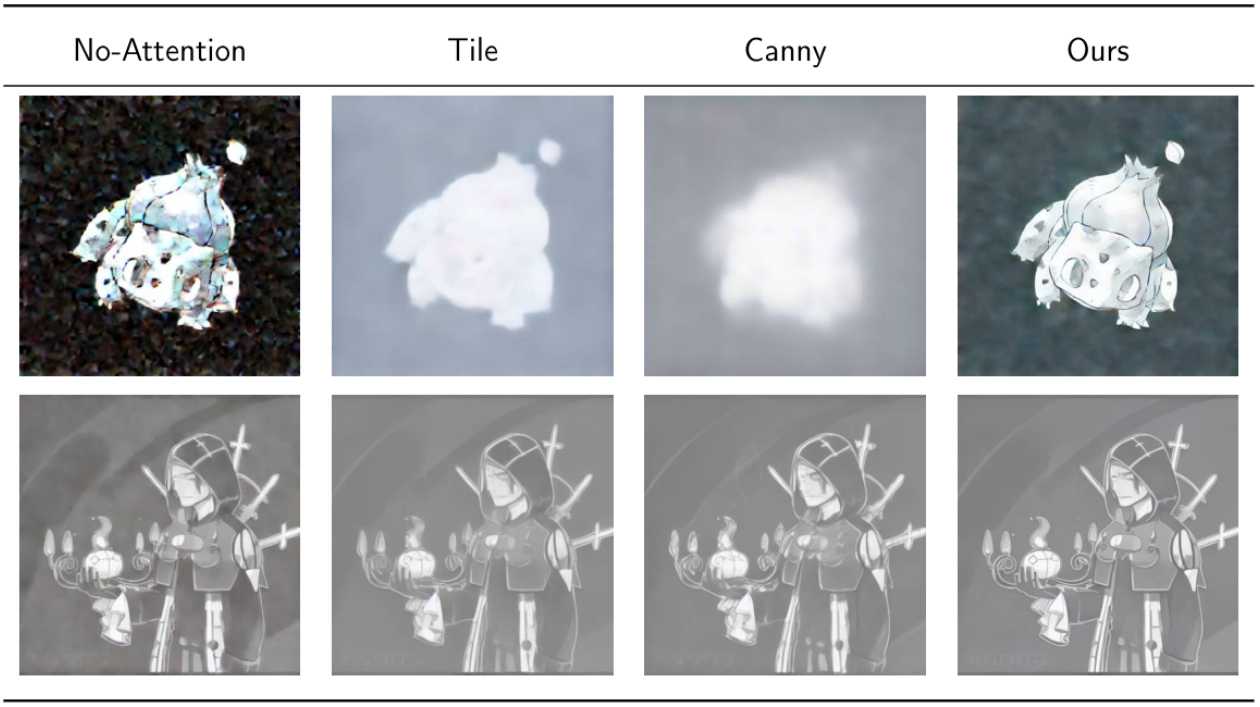



Figure 2. **Further qualitative comparisons** on cover-art-style manga panels. AttnMCN (right column) preserves fine character detail and scene structure while other variants degrade significantly.

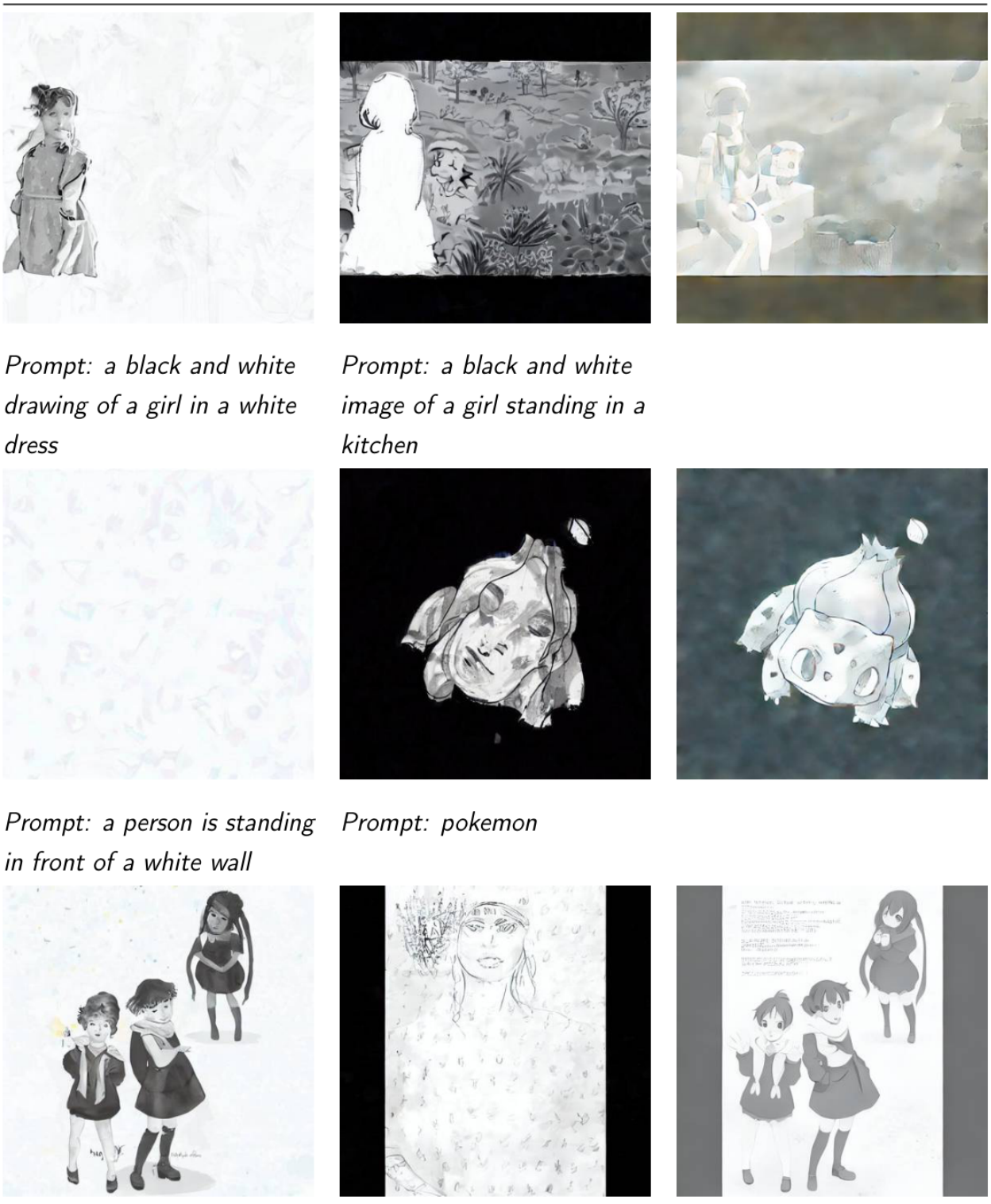



Figure 3. **Extended I2I baseline comparison.** Columns: Foreground I2I, Background I2I, AttnMCN (Ours). Auto-generated captions (*e.g.*, "*a black and white drawing of a girl in a white dress*") fail to capture spatial layout, while AttnMCN faithfully reconstructs the intended composition.

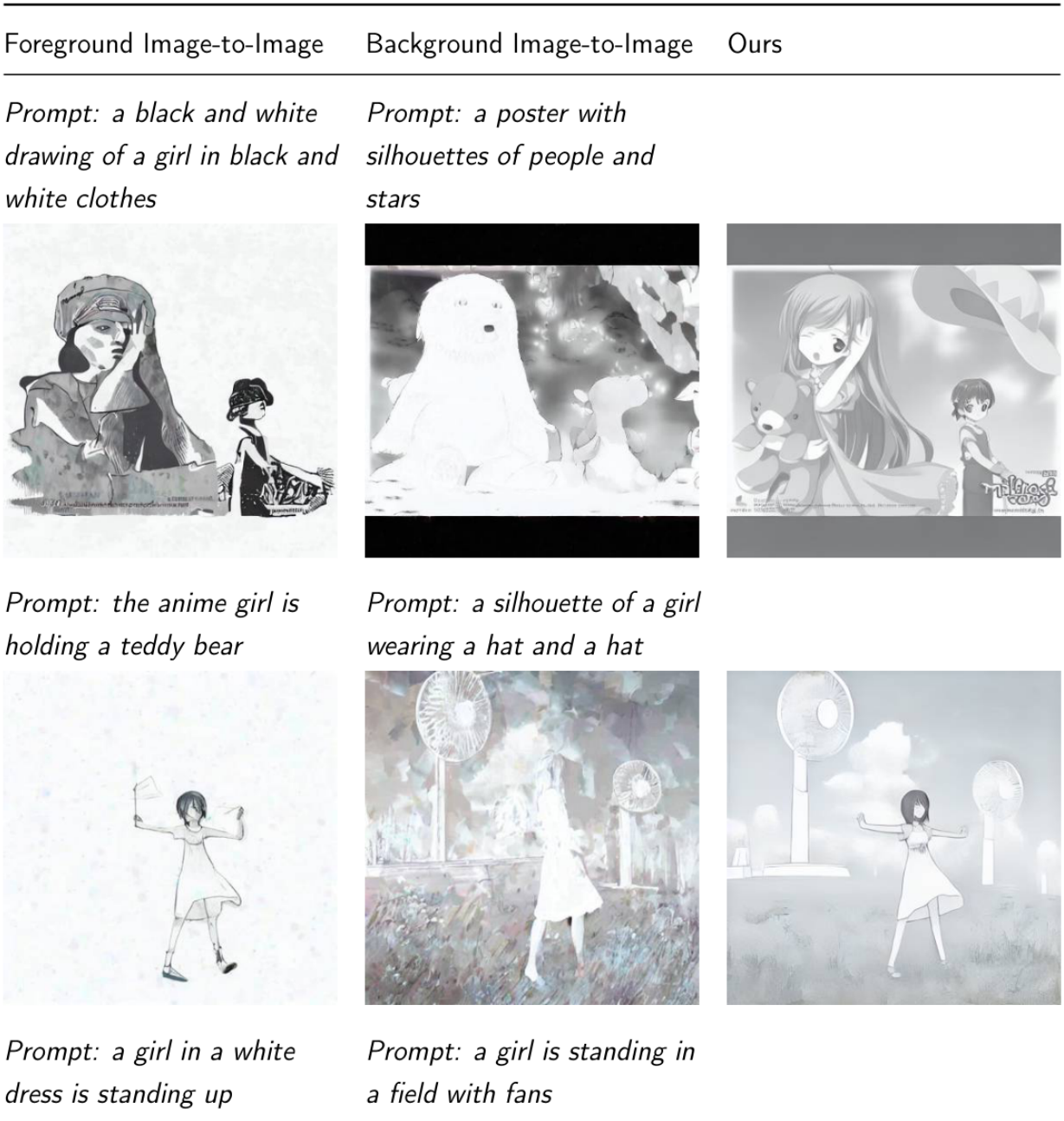
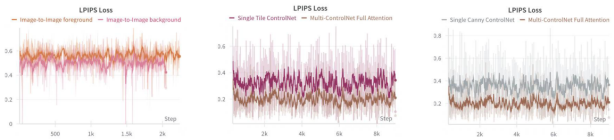
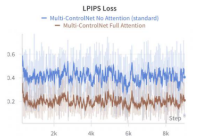



Figure 4. **Further I2I comparisons.** Each row shows Foreground I2I (with auto-generated prompt), Background I2I, and AttnMCN. Text prompts such as "*a poster with silhouettes of people and stars*" or "*a girl is standing in a field with fans*" fail to capture the spatial and stylistic details that AttnMCN preserves through direct image conditioning.

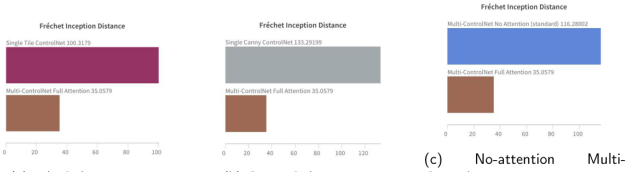



(a) Image-to-Image (b) Tile-Only Attention (c) Canny-Only Attention




(d) No-attention Multi-ControlNet

Figure 5. **LPIPS [6] training curves.** (a) I2I baseline. (b) Single Tile vs. AttnMCN. (c) Single Canny vs. AttnMCN. (d) No-Attention MCN [2] vs. AttnMCN. AttnMCN (brown) consistently achieves lower perceptual loss.




(a) Tile-Only Attention (b) Canny-Only Attention (c) No-attention Multi-ControlNet

Figure 6. **FID [1] comparison.** (a) Single Tile (100.3) vs. AttnMCN (35.1). (b) Single Canny (133.3) vs. AttnMCN. (c) No-Attention MCN [2] (116.3) vs. AttnMCN.

(a) Generated manga character. (b) Generated manga scene.

Figure 7. **Sample AttnMCN outputs.** High-quality manga generation with coherent character detail and scene structure.

## References


[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 1, 6

[2] Hugging Face. MultiControlNetModel class. https://github.com/huggingface/diffusers, 2025. 1, 2, 6

[3] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using Manga109 dataset. In *Multimedia Tools and Applications*, pages 21811–21838, 2017. 1

[4] Aasim Sani. AMPD-base. Kaggle Dataset, 2024. 1

[5] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *ICCV*, pages 3836–3847, 2023. 1, 2

[6] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 1, 6