Scaling Down Text Encoders of Text-to-Image Diffusion Models

Supplementary Material

In this supplementary material, we present more details, experiments, qualitative results and disccusions that not covered in the main text.

- Sec. 1 provides **more details** for *mode collapse*, *qualitative evaluation*, *model architecture*, *training data* (Fig. 10), and *figure details*.
- Sec. 2 describes **more experiments** conducted on PixArt-Alpha (Tabs. 6 and 7, Fig. 11) to assess the *generalizability* of our method.
- Sec. 3 provides **more qualitative results** on *image quality* (Fig. 12), *semantic understanding* (Fig. 13), and *text rendering* (Fig. 14) comparison across different T5 size.
- Sec. 4 discusses our **limitations & social impact**.
- Sec. 5 lists the **prompts used** to generate the images featured in the main text.

1. More Details

Visualization of Mode Collapse. When text embeddings collapse into several modes, the text encoder will represent two different concepts, such as rat and man, with the same embedding. As illustrated in Fig. 9, naïve distillation makes T5 predict the wrong embedding and therefore the DM generates a failed image.

Qualitative evaluation. We conducted a user study with 20 participants, presenting them with 15 two-image-text pairs for comparison. Participants were asked to evaluate which pair aligned better with the prompt: A, B, or if they were tied. The results indicated that 82.7% believed the pairs were tied, 12.0% favored T5-XXL, and 5.3% favored T5-Base.

Model Architecture. To project the student encoder's embedding to the T5-XXL's embedding space, we use a Multi-Layer Perceptron (MLP). The MLP consists of a Linear layer with input dimension of student's embedding dimension (512 for T5-Small , 768 for T5-Base, 1024 for T5-Large, and 2048 for T5-XL) and output dimension of 4096, followed by a ReLU [8] activation layer, a dropout layer [11] with dropout rate of 0.1, and a final Linear layer with input dimension 4096 and output dimension 4096.

Dataset Samples. We visualize the three components of training prompt data in Figure 10. Each component highlights a specific ability of the T5-XXL model, providing a comprehensive guide for our student model to learn diverse and complex embeddings.

Figure Details. In Figure 1, the y-axis represents the score of student models as a percentage of the T5-XXL model's score, which is set as 100%. For score of T2I-CompBench and CommonText validation set, we take the average across all categories. The x-axis displays logarithm of the number of parameters in each student model.

For Figure 6, we randomly sample 500 prompts from DiffusionDB [13] and pass them to our T5-Base and T5-XXL respectively. we compute the mean of the embeddings across the sequence dimension, resulting in 500 data points for each model, each with 4096 dimensions. To facilitate visualization, we apply t-SNE [12] to reduce these high-dimensional data points to 2 dimensions. The reduced data are then plotted on a 2D plane, with the t-SNE components on the x and y axes, allowing us to compare how the T5-Base and T5-XXL models represent the prompts in a lower-dimensional space.

For Figure 7, we extract the attention proability of the attention layers of the 10^{th} inference step. Flux utilizes the same architecture as SD3 [3], which combines self-attention and cross-attention into a unified large attention matrix. As such, the matrix is of shape $[B, S, H+E, H+E]$, where B is batch size, S is sequence length, H is hidden states dimension, and E is encoder hidden states dimension. We extract the upper right corner ($[B, S, :E, E:]$) of the attention map, which corresponds to cross-attention in previous UNet/DiT [2, 10] structured diffusion models. For the visualization of a specific token, we extract the attention map for that token by indexing the sequence dimension. We then upsample the attention map from the latent space to the image space and overlay it onto the generated image, providing a clear view of how attention is distributed for that token in the final output.

(a) (b) (c)

Figure 9. **Mode collapse caused by naive distillation.** The images were generated using the prompts (a) 'A rat,' (b) 'A cat,' and (c) 'A monkey.' Due to mode collapse, T5 represents 'rat,' 'cat,' and 'man' with the same embedding, and 'monkey' and 'woman' with the same embedding.

2. More Experiments

Generalizability of Base Models. In addition to Flux, we also evaluated our method on PixArt-Alpha using T5-Base as the student encoder. For training, we employed the LAION-Aesthetics-6.5+ and T2I-CompBench datasets, excluding CommonText due to PixArt-Alpha's difficulty in

Model	Attribute Binding			Spatial Relationship		Numeracy (↑)
	Color (↑)	Shape (↑)	Texture (↑)	2d-spatial (↑)	3d-spatial (↑)	
Pixart-Alpha w/ T5-Base	39.89	45.00	51.58	18.75	35.59	49.75
Pixart-Alpha w/ T5-XXL (teacher)	38.91	40.78	46.18	20.24	34.63	50.29

Table 6. **Semantic Understanding Comparison for PixArt-Alpha.** We compare our T5-Base and T5-XXL across the six categories in T2I-CompBench [4].

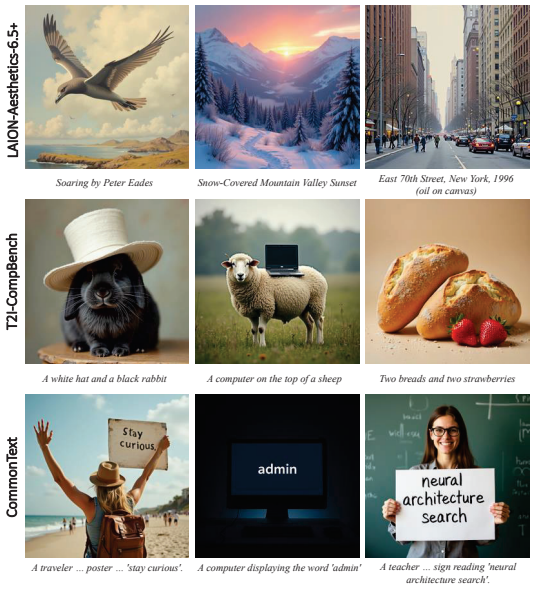

Figure 10. **Training data samples.** Prompts are sampled from each dataset and images are generated using Flux.

Model	FID (↓)	CLIP-Score (↑)
Pixart-Alpha w/ T5-Base	33.58	28.22
Pixart-Alpha w/ T5-XXL (teacher)	29.59	30.75

Table 7. **FID/CLIP-Score Comparison for PixArt-Alpha.** We compare our T5-Base and T5-XXL on the full MSCOCO-2014 validation set.

rendering text. We use the 512 checkpoint [9]. The training was conducted on 8 A100 GPUs with a total batch size of 32. The guidance scale was randomly sampled between 5 and 10, in intervals of 0.5, and we used 20 steps for iterative denoising. We applied the AdamW [7] optimizer with default PyTorch parameters, along with a linear learning rate scheduler and a learning rate of $1e-4$.

The results, summarized in Tables 6 and 7, show that our T5-Base model surpasses T5-XXL in categories related to semantic understanding, such as color, shape, texture, and 3D spatial, with only a minimal decrease in performance for 2D spatial and numeracy tasks. In terms of image quality, while T5-Base has a lower CLIP score and a higher FID score compared to T5-XXL, it still demonstrates the ability to generate high-quality images, as depicted in Figure 11.

3. More Qualitative Results

We provide more qualitative comparison between T5 of different size in Figs. 12 to 14. While T5-Small is capable of generating images with reasonable quality, it occasionally fails to capture the precise semantics of prompts, such as row 2 of Figure 12 and row 4 of Figure 13. It completely fails in text rendering as shown in Figure 14. Models larger than T5-Small generally performs well in all three categories, indicating that smaller text encoders such as T5-Base suffice for general image synthesis. However, T5-XXL still excels in generating fine details. When computational resources are not a constraint, T5-XXL is still the preferred text encoder for diffusion models.

4. Limitations & Social Impact

Limitations. Although our method reduces the memory requirement of GPUs to run large diffusion models, they may not capture the full depth and complexity of larger datasets, potentially leading to a loss in the richness and accuracy of generated content, especially for complex tasks. The process of model compression and knowledge distillation can also introduce or amplify existing biases, as the distilled model might overfit to specific training data characteristics [1, 5]. Additionally, the training time is relatively long due to the iterative nature of our step-following distillation. A potential solution is to use real images for pre-training. For future work, since LLM are becoming increasing popular for image synthesis [6, 14], it is worth investigating how we can distill LLM for image generation.

Social Impact. The development of smaller, efficient text encoders democratizes access to advanced diffusion models, fostering innovation in image synthesis by reducing computational barriers. However, these smaller models might lose some depth and accuracy, potentially introducing or exacerbating biases if not carefully managed. Additionally, increased accessibility might lead to over-reliance on AI-generated content, impacting creativity and originality. Balancing these benefits and drawbacks is crucial for maximizing positive social impact.

(a) A majestic dragon ... crystal mountain during a vibrant aurora.

(b) A mystical underwater city made of coral and bioluminescent structures.

(c) An ancient Chinese palace ... ponds under a crescent moon.

(d) A steampunk airship sailing through a sky filled with clouds.

(e) A couple dancing ... with golden sunlight streaming through trees.

Figure 11. **Qualitative results for PixArt-Alpha.** Comparison of T5-XXL (top) and T5-Base (bottom).

Figure 12. **Image Quality Comparison across T5 of Different Size.** We use the same seed and guidance scale of 3.5. (row 1): A bustling cyberpunk metropolis at night, illuminated by a kaleidoscope of neon lights and holographic advertisements. The streets are crowded with people wearing futuristic attire. (row 2): Portrait of a stylish young woman wearing a futuristic golden bodysuit that creates a metallic, mirror-like effect. She is wearing large, reflective blue-tinted aviator sunglasses. Over her head, she wears headphones with metallic accents. (row 3): Baroque ship, beautiful golden mirror sail, golden cumulus clouds, black sky, golden rock, surreal, vivid colors, chiaroscuro lighting, 50mm lens. (row 4): An astronaut exploring a mysterious alien landscape, with strange vegetation and a planet rising in the sky. (row 5): Haunting dark fantasy illustration of an ancient, twisted statue standing atop a steep cliff, overseeing a decaying metropolis shrouded in mist. The sky churns with ominous clouds and flashes of lightning.

Figure 13. **Semantic Understanding Comparison.** Each prompt corresponds to a category of T2I-CompBench, specifically color, shape, texture, 3D-relationship, and numeracy. (row 1): A man in a gray jacket standing in a kitchen next to a black dog. (row 2): A triangular sign and a small sculpture. (row 3): A metallic necklace and a leather chair. (row 4): An airplane in front of a clock. (row 5): Four people gathered for a picnic.

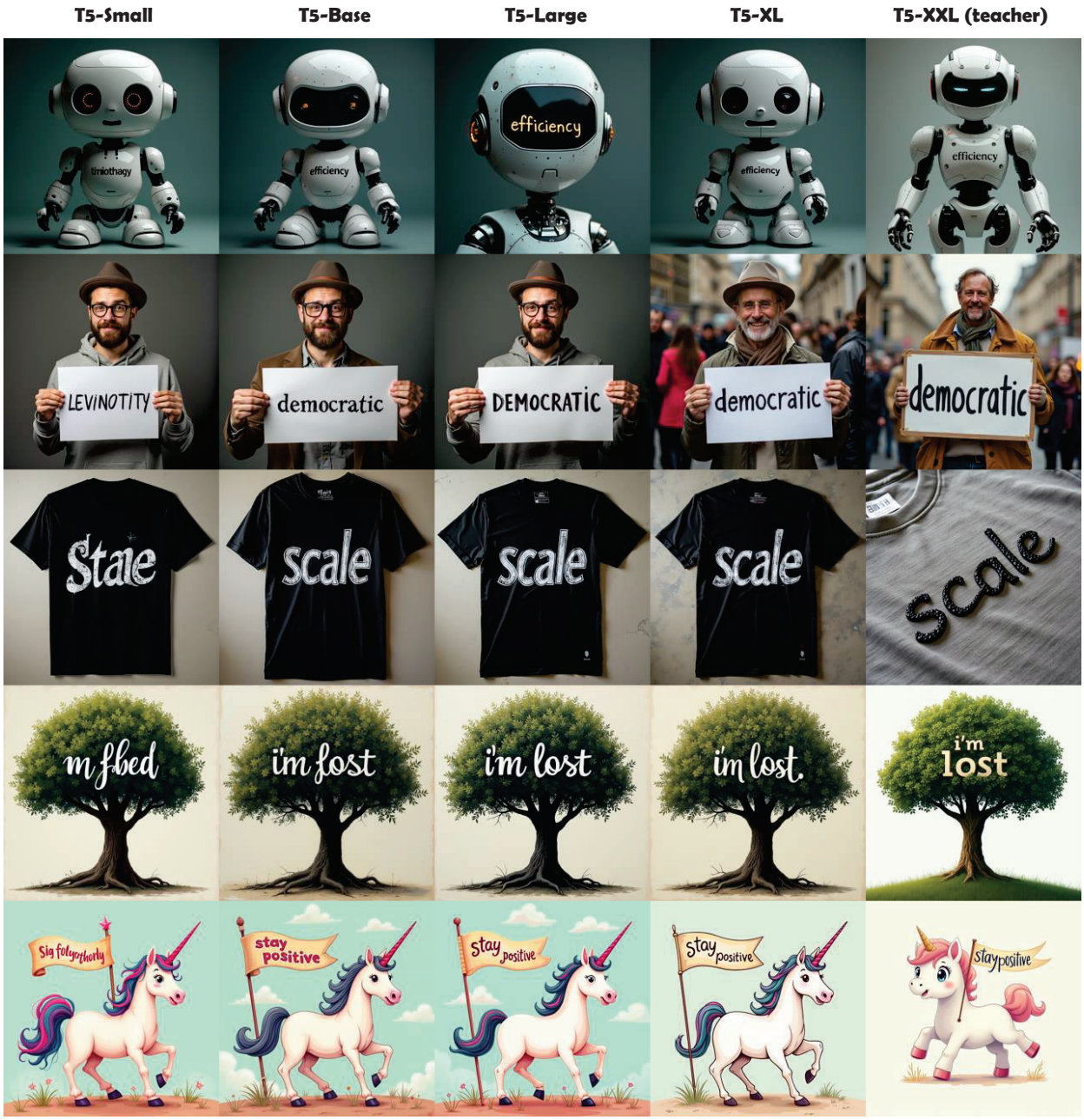

Figure 14. **Text Rendering Comparison.** Models larger than T5-Small can render text on various objects within diverse contexts. (row 1): A robot displaying the word 'efficiency'. (row 2): An optimist holding a sign reading 'democratic'. (row 3): A t-shirt with the inscription 'scale'. (row 4): A tree displaying the word 'i'm lost'. (row 5): A unicorn carrying a banner with words 'stay positive'.

5. Prompts for Generating Images

5.1. Figure 5 in Main Text

1. A stunning Japanese-inspired fantasy painting of a lone samurai, silhouetted against a massive full moon, standing beneath a windswept, crimson-leafed tree. Falling petals swirl around him, creating a melancholic yet serene atmosphere. The dramatic chiaroscuro lighting highlights the dramatic contrast between the cool-toned background of deep blues and grays and the warm reds of the foliage.
2. Ink-splash-style. Extreme closeup of a dapper figure in a stylized, richly detailed black top hat, adorned with decorative golden accents, stands against a white background. The character is a skeleton with very detailed skull and long canines as vampire fangs. He cloaked in a vibrant victorian jacket, featuring intricate golden embellishments and a deep red vest underneath. He wears a large victorian monocle with a yellow-tinted lense and copper frame very reddish. Exquisite details include a shiny silver cross and a blue gem on the chest, harmonizing with splashes of paint in vivid hues of blue, gold, and red that artistically cascade around the figure, blending an impressionistic flair with elements of surrealism. The atmosphere is whimsical and opulent, evoking a sense of grandeur and mystery.
3. An ancient, overgrown temple in a dense jungle, illuminated by the soft light of early morning.
4. A beautiful woman stands gracefully beside her companion, a majestic lion. The lion stands tall and proud, its mane a cascade of alternating black and white gears and intricate cogs, its body radiating a soft, ethereal glow. The woman, radiating grace and elegance, wears a flowing gown of swirling black and white that blends seamlessly with the ethereal landscape. The scene is set against a backdrop of a timeless realm, bathed in the soft glow of a twilight where black and white blend seamlessly.
5. A portrait of a cybernetic geisha, her face a mesmerizing blend of porcelain skin and iridescent circuitry. Her elaborate headdress is adorned with bioluminescent flowers and delicate, glowing wires. Her kimono, a masterpiece of futuristic design, shimmers with holographic patterns that shift and change, revealing glimpses of the complex machinery beneath. Her eyes gaze directly at the viewer with an enigmatic expression.
6. A single, crazy blue and black fighter in the sky. It overwhelms the viewer with its artistic flying skills while trailing a meteor tail. Ace pilot of the Republic who was unrivalled in the 1940s. His second name is: The Magician of the Blue Wings, a genius aviator, one of a kind in 100 years. The warriors who challenged him, were destroyed by him, were overrun by him and scattered became many stars. The Milky Way is said to be the graveyard of such aerialists. 'As we drive our dreams, we fly across the sky and weave our dreams for tomorrow's night.' He told me with few words. 'The fighter who wants peace more than anyone else, who gives up everything, who flies faster than anyone else. Like a song spinning in the night sky, it pioneers the starry skies, scattering fantastic sparkles.
7. An incredibly realistic scene of a white kitten wearing a majestic golden veil, rendered with high attention to detail. The cat's eyes, large and amber, should be given a reflective quality that captures the viewer's gaze. The golden veil should be richly detailed with intricate patterns, cascading elegantly over the cat's head and shoulders, with delicate folds that suggest a soft, luxurious fabric. The veil should be adorned with a diadem featuring a prominent red gemstone, surrounded by golden filigree.
8. A high fantasy castle floating among the clouds at sunset, surrounded by flying mythical creatures.
9. A grand library filled with ancient books and magical artifacts, lit by the warm glow of candlelight.

5.2. Figure 8.a in Main Text

5.2.1. Canny

1. A cyberpunk living room.
2. A living room with British royal style
3. A living room in jungle.
4. A living room with moonlight shedding in.

5.2.2. Depth

1. A white rabbit in forest.
2. A furry rabbit resting on grass of a park.
3. A robot rabbit in a futuristic lab.
4. A brown rabbit on a table with light shining on its fur.

5.2.3. Pose

1. A female travel blogger with messy beach waves.
2. A female adventure photographer with windswept hair.
3. A male travel influencer with tousled mountain curls.
4. A business man with suit in a coffee store.

5.3. Figure 8.b in Main Text

1. Anime ((masterpiece,best quality, detailed)), outdoor, wind lift, souryuu asuka langley, interface headset, red bodysuit, (realistic:1.3).
2. Well-fitting man equipped with hoodie and cap hiding upper face in anime manga style.
3. Anime theme masterpiece,best quality,1girl,solo,looking at viewer, fur (clothing), black hair, black legwear,(electric guitar:1.4), reflection, splash, droplets, rust, sparks, asphalt, ground vehicle, sports car, super car, mechanical,burning, playing instrument, livestream.

4. Anime boy with a dragon companion, standing in a medieval village, with a castle in the background, (heroic:1.3).
5. Anime girl in a futuristic racing suit, riding a high-tech motorcycle through a neon-lit city, (dynamic:1.3).
6. An anime girl in a blue dress and straw hat, with long black hair and flowing curly bangs, in the style of anime, against a background of a coastal street by the sea, on a bright sunny day, with flowers on a windowsill, with a cheerful expression, with detailed design, with a watercolor painting effect, and vibrant colors, Hayao Miyazakis manga, with high resolution and clear details –ar 1:2 –stylize 750 –v 6.1.
7. Pretty cyborg lady, lots of details, sakura flowers, fine art, futuristic setting.
8. Anime boy with headphones, sitting in a cozy room filled with books and plants, working on a computer, (slice of life:1.3).
9. Anime warrior princess with a glowing sword, standing in a mystical forest, surrounded by magical creatures, (epic:1.3).
10. Anime girl with long flowing hair, holding a magical staff, standing on a cliff overlooking a vast ocean, with a sunset in the background, (fantasy:1.3).

5.4. Figure 8.c in Main Text

1. A fantasy forest with glowing mushrooms and mystical creatures.
2. A bustling market in a medieval village with various stalls and people.
3. A space station orbiting a planet with astronauts floating outside.
4. A dramatic stormy sea with a lighthouse and waves crashing against the rocks.
5. A snowy mountain landscape with a cozy cabin and smoke coming from the chimney.

References

[1] Kuluhan Binici, Nam Trung Pham, Tulika Mitra, and Karianto Leman. Preventing catastrophic forgetting and distribution mismatch in knowledge distillation via synthetic data. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 663–671, 2022.

[2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

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

[4] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[5] Shaoyi Huang, Dongkuan Xu, Ian EH Yen, Yijue Wang, Sung-En Chang, Bingbing Li, Shiyang Chen, Mimi Xie, Sanguthevar Rajasekaran, Hang Liu, et al. Sparse progressive distillation: Resolving overfitting under pretrain-and-finetune paradigm. *arXiv preprint arXiv:2110.08190*, 2021.

[6] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024.

[7] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[9] PixArt-alpha. Pixart-alpha/pixart-xl-2-512x512, 2024.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

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

[12] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

[13] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

[14] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.