SynthVLM: Towards High-Quality and Efficient Synthesis of Image-Caption Datasets for Vision-Language Models

Supplementary Material

Zheng Liu[♠], Hao Liang[♠], Bozhou Li[♠], Wentao Xiong[♠], Chong Chen[♡], Conghui He[♣], Wentao Zhang[♠], Bin Cui[♠]

[♠]Peking University [♡]Huawei Cloud BU [♣]Shanghai AI Laboratory

[†]lz030515123@gmail.com, [†]hao.liang@stu.pku.edu.cn, {bin.cui, wentao.zhang}@pku.edu.cn

1 Preliminary

1.1 Diffusion Model

Denoising diffusion probabilistic models (DDPMs) [4, 17, 19] are a class of generative models renowned for their ability to generate extremely high-quality images. The core idea of DDPMs involves modeling the data distribution by gradually adding Gaussian noise to the input image during the forward process and then predicting and removing this noise to reconstruct the image during the backward process.

Given a source image data distribution $x_0 \sim q(x_0)$, Gaussian noise is added over T steps to obtain x_T. The forward process is defined as:

$$q(x_1, \ldots, x_T \mid x_0) := \prod_{t=1}^{T} q(x_t \mid x_{t-1}),$$

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

where β_t controls the variance of the noise added at each step.

The distribution after t steps can be written as:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I),$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$.

The backward process aims to reconstruct the data by learning a series of Gaussian distributions that approximate the forward process:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

where μ_θ and Σ_θ are neural networks parameterized by θ.

While DDPMs have shown promising results, several improvements have been proposed to enhance their efficiency [21, 25] and sample quality [3, 16]. The superior performance of diffusion models has been leveraged in various sub-tasks, including image generation, image translation, inpainting [15, 22].

1.2 Vision Language Models

The integration of visual knowledge into large language models (LLMs) has become a pivotal area of research due to the rapid advancements in LLMs. VLMs combine vision information from vision encoders with LLMs, thus enabling these models to process and interpret visual inputs for various visual tasks [10, 13, 24] with enhanced accuracy and efficiency. Pioneering frameworks like CLIP [18] leverage contrastive learning on expansive image-caption datasets to align modalities, forming the groundwork for cross-modal comprehension. Various adapters [5, 8, 9, 11, 12, 14] are introduced to further integrate different modalities. For example, LLaVA [11, 12] employs a straightforward MLP to inject the vision information into LLMs. Whereas more complex implementations like the Q-Former in BLIP [8, 9] utilize cross-attention to enhance modality integration.

Recent studies [1, 6, 11, 12, 23] aims to boost VLM performance by focusing on the quality of both pre-training and fine-tuning datasets. Models like LLaVA [11, 12] and ShareGPT4V [1] have shown remarkable advancements in understanding and following complex instructions through instruction tuning. Although these improvements help align the vision modality and establish a solid basis for cross-modal comprehension, they require extensive datasets for training and could potentially diminish the model's language capabilities.

2 Implementation Details

2.1 Data Generation

In this section, we detail the hyperparameters and procedures used for data generation.

We employed the Stable Diffusion XL(SDXL) model for image synthesis, following the framework outlined by the original authors [17]. To identify the optimal parameter configuration for our use case, we conducted a grid search strategy aimed at maximizing the CLIPScore for evaluating the semantic alignment between generated images and their corresponding textual descriptions.

Specifically, we randomly sampled 1k captions from our caption pool and used these samples to systematically evaluate different combinations of generation parameters. The grid search allowed us to empirically determine the most effective configuration for producing high-quality, semantically relevant synthetic images.

Based on this optimization process, we configured SDXL with 60 sampling steps. All images were generated at a resolution of 1024×1024 pixels. These configurations consistently yielded superior quality.

2.2 Data Selection

In this section, we describe the strategy and prompts used for data selection. Our goal was to curate a high-quality dataset that aligns closely with our generation objectives.

To achieve this, we employed a two-stage filtering process combining heuristic rules and large language model (LLM)-based evaluation. The specific filtering rules and prompt templates are detailed in Table 1.

For heuristic filtering, we utilized the Data-Juicer framework, which offers a modular and scalable pipeline for rule-based data

Table 1: Metric and Prompt used for Caption Filtering

Caption Filtering

Rule-Based Metrics

- **Alphanumeric Filter:** Tokenization: false, Min ratio: 0.60
- **Character Repetition Filter:** Rep length: 10, Max ratio: 0.09373663
- **Flagged Words Filter:** Language: en, Tokenization: false, Max ratio: 0.0
- **Perplexity Filter:** Language: en, Max perplexity: 5500.0
- **Special Characters Filter:** Min ratio: 0.16534802, Max ratio: 0.42023757
- **Word Repetition Filter:** Language: en, Tokenization: false, Rep length: 10, Max ratio: 0.03085751
- **Image-Text Matching Filter:** HF BLIP: Salesforce/blip-itm-base-coco, Min score: 0.8, Max score: 1.0, Horizontal flip: false, Vertical flip: false, Reduce mode: avg, Any or all: any, Mem required: 1500MB
- **Image-Text Similarity Filter:** HF CLIP: openai/clip-vit-base-patch32, Min score: 0.28

Prompt

Assume you are an expert in the field of AI image generation. Your goal is to select high-descriptive prompts that will enable the successful generation of images. I will provide you with a specific descriptive prompt, and your task is to evaluate it thoroughly. Consider the prompt's level of detail, its logical coherence, and the clarity with which it describes the desired image. It is essential to assess whether the prompt contains sufficient information to guide the diffusion model effectively, ensuring that it can produce an image that meets expectations. You should only respond with Yes or No.

preprocessing. This allowed us to implement filters targeting criteria such as minimum caption length, syntactic completeness, and lexical diversity. Additionally, we removed low-information and repetitive entries to enhance the overall quality of the dataset.

Following this, we conducted LLM-based filtering using LLaMA3-70B-Instruct, a powerful instruction-tuned language model. This model was used to assess the semantic clarity, descriptiveness, and relevance of each caption to ensure alignment with our image generation goals. Captions that met the predefined criteria for specificity, visual richness, and informativeness were retained.

3 Another Advantage: Addressing Data Privacy

Utilizing web-sourced data introduces numerous security and privacy concerns [2, 7]. They may contain personal information or copyrighted materials, posing potential legal and ethical challenges. Moreover, the inclusion of sensitive or inappropriate content within training datasets can instigate ethical issues, thereby compromising the models' integrity and fairness.

Our synthetic approach removes reliance on real-world personal data (e.g., user photos), safeguarding user privacy throughout the data generation process while maintaining model capability.

we compare the synthetic image and the original image in Figure 1. Synthetic data offers significant advantages in protecting data privacy. In Figure 1, synthetic images in (a) show vehicles and tickets without revealing real license plates and ticket information, ensuring privacy protection. In contrast, original images in (b) display actual license plates and ticket information, which can potentially lead to privacy issues.

4 T-SNE visualize of our dataset

In this section, we use t-distributed Stochastic Neighbor Embedding (t-SNE) to compare the distribution of our synthetic dataset with

Privacy ✓ (a) Synthetic Images

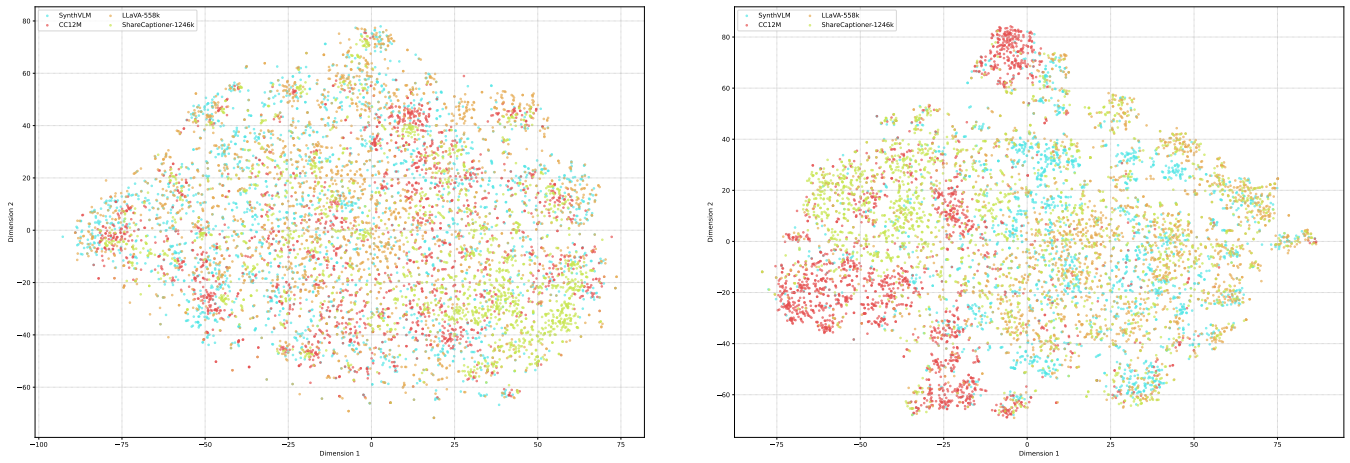
Privacy ✗ (b) Original Images

Figure 1: From (a), it is evident that synthetic images can avoid displaying real license plates and ticket information. In contrast, (b) contains actual license plates and ticket information, which can potentially lead to privacy issues.

several real-world datasets. This comparison aims to evaluate the similarity in semantic and visual space, providing insights into the realism and utility of the generated data.

For the real datasets, we selected LLaVA-558K[11], ShareCaptioner[1], and CC12M[20], which are widely used for vision-language training. As our synthetic dataset, we used SynthVLM-100K, generated using the methods described in earlier sections. From each dataset, we randomly sampled 1k image-caption pairs for analysis.

We performed t-SNE visualization separately on image features and caption embeddings. Feature representations were extracted



(a) Visualization of text embeddings. **(b) Visualization of image embeddings.**

Figure 2: TSNE visualizations of synthetic and real datasets for text and image modalities.

using a pre-trained vision-language model to ensure consistency and comparability across datasets.

As shown in Figure 2, the image and caption distributions of our synthetic dataset are closely aligned with those of the real datasets. This visual overlap indicates that the generated data captures similar semantic and visual characteristics as real-world data, supporting the authenticity and high quality of our generation pipeline.

Furthermore, the observed distributional similarity suggests that models trained on our synthetic data are likely to exhibit strong generalization and performance on real-world tasks. This supports the viability of using synthetic data to supplement or replace real data in various vision-language applications.

5 More examples of our dataset

In this section, we present additional qualitative examples from our synthetic dataset, SynthVLM-100K, to further demonstrate the high quality and diversity of the generated image-caption pairs.

As illustrated in Figures 1 through 4, the samples cover a wide range of visual concepts and exhibit strong semantic alignment between images and captions. These examples highlight the capability of our data generation pipeline to produce visually coherent and semantically rich content across various domains.

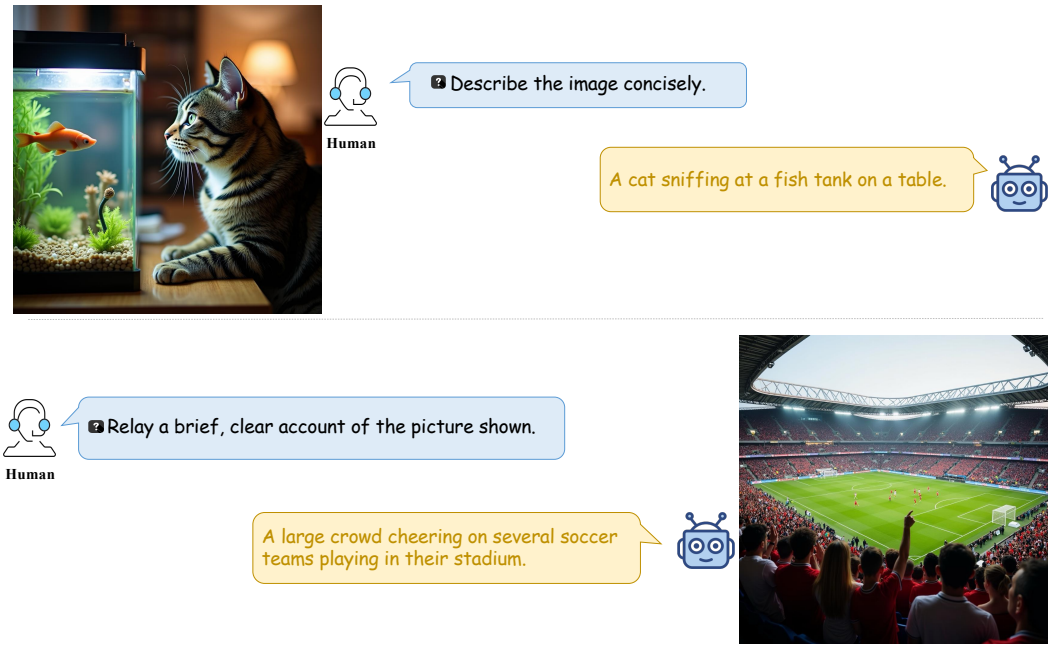


Figure 3: Illustration of our SynthVLM

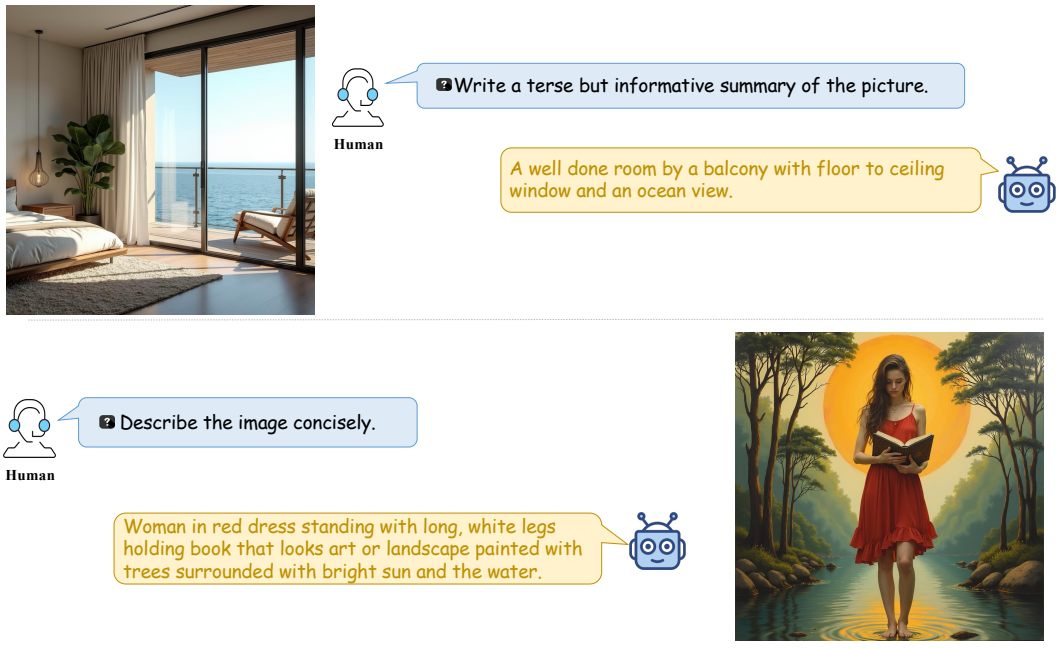


Figure 4: Illustration of our SynthVLM

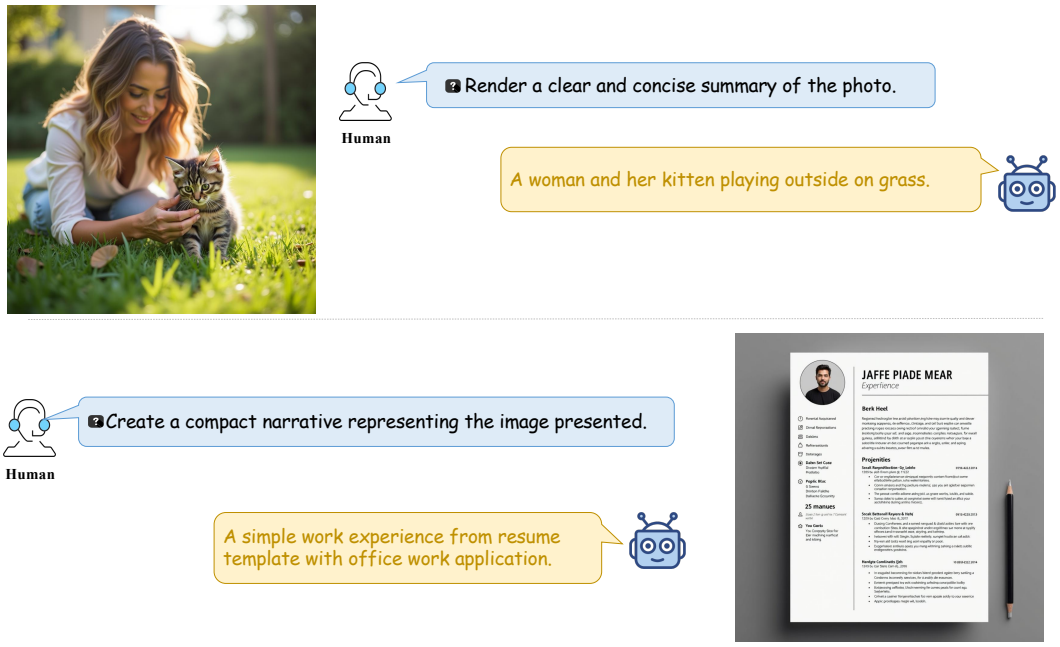


Figure 5: Illustration of our SynthVLM

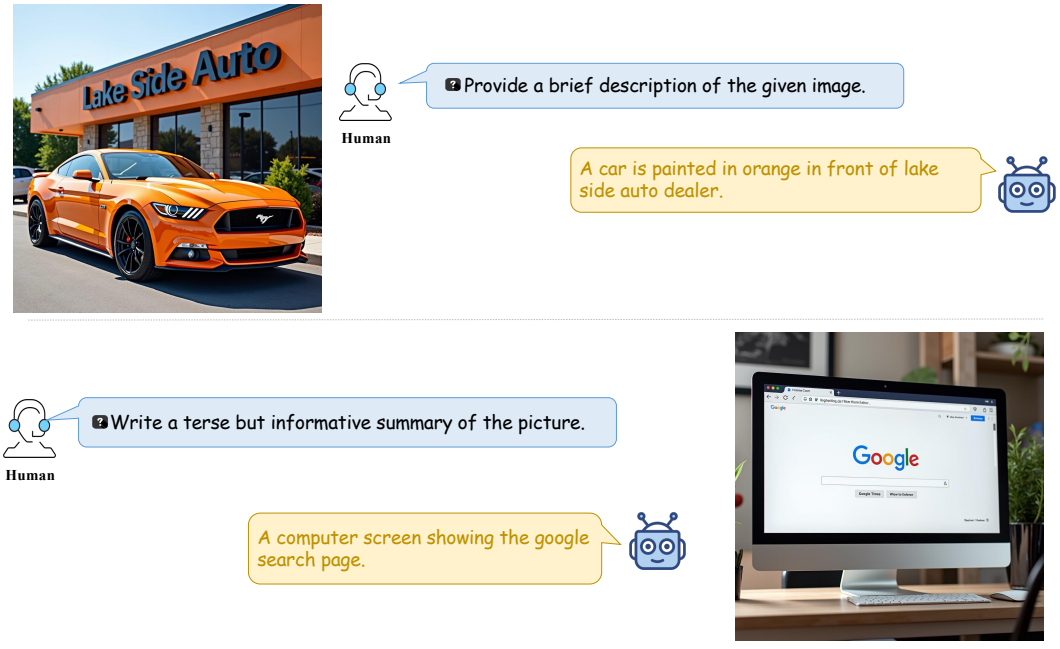


Figure 6: Illustration of our SynthVLM

References

[1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *CoRR* abs/2311.12793 (2023).

[2] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2024. Security and Privacy Challenges of Large Language Models: A Survey. *CoRR* abs/2402.00888 (2024).

[3] Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. 8780–8794.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[5] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2023. Bootstrapping Vision-Language Learning with Decoupled Language Pre-training. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

[6] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *CoRR* abs/2305.03726 (2023).

[7] Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023. Privacy in Large Language Models: Attacks, Defenses and Future Directions. *CoRR* abs/2310.10383 (2023).

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Vol. 162. 12888–12900.

[10] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded Language-Image Pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10955–10965.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

[13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *CoRR* abs/2303.05499 (2023).

[14] Junyu Lu, Ruyi Gan, Dixiang Zhang, Xiaojun Wu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2023. Lyrics: Boosting Fine-grained Language-Vision Alignment and Comprehension via Semantic-aware Visual Objects. *CoRR* abs/2312.05278 (2023).

[15] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. 2022. Conditional Diffusion Probabilistic Model for Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. 7402–7406.

[16] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Vol. 139. 8162–8171.

[17] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR* abs/2307.01952 (2023). https://doi.org/10.48550/ARXIV.2307.01952 arXiv:2307.01952

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. 10674–10685.

[20] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From Pixels to Prose: A Large Dataset of Dense Image Captions. arXiv:2406.10328 [cs.CV] https://arxiv.org/abs/2406.10328

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

[22] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. 2023. Dual Diffusion Implicit Bridges for Image-to-Image Translation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

[23] Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. 2024. Finetuned Multimodal Language Models Are High-Quality Image-Text Data Filters. *CoRR* abs/2403.02677 (2024).

[24] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022. GLIPv2: Unifying Localization and Vision-Language Understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

[25] Jinchao Zhu, Yuxuan Wang, Siyuan Pan, Pengfei Wan, Di Zhang, and Gao Huang. 2024. A-SDM: Accelerating Stable Diffusion through Model Assembly and Feature Inheritance Strategies. *CoRR* abs/2406.00210 (2024).