Enhancing Baidu Multimodal Advertisement with Chinese Text-to-Image Generation via Bilingual Alignment and Caption Synthesis

Kang Zhao*
Baidu Search Ads, Baidu Inc.
Beijing, China
zhaokang01@baidu.com

Xinyu Zhao*
Baidu Search Ads, Baidu Inc.
Beijing, China
zhaoxinyu03@baidu.com

Zhipeng Jin
Baidu Search Ads, Baidu Inc.
Beijing, China
jinzhipeng@baidu.com

Yi Yang†
Baidu Search Ads, Baidu Inc.
Beijing, China
yangyi15@baidu.com

Wen Tao
Baidu Search Ads, Baidu Inc.
Beijing, China
taowen02@baidu.com

Cong Han
Baidu Search Ads, Baidu Inc.
Beijing, China
hancong01@baidu.com

Shuanglong Li
Baidu Search Ads, Baidu Inc.
Beijing, China
lishuanglong@baidu.com

Lin Liu
Baidu Search Ads, Baidu Inc.
Beijing, China
liulin03@baidu.com

ABSTRACT

Recent advances in generative artificial intelligence have revolutionized information retrieval and content generation, opening up new opportunities for the e-commerce industry. In particular, text-to-image generation models offer a novel approach to guiding the image generation process using natural language input, which is inspiring for multimodal search advertising. Traditional multimodal search ads require advertisers to prepare ad creatives, such as ad images, which is time-consuming and requires uniform image specifications and content quality inspection. To this end, we propose a streamlined generation framework for search ad image creatives. First, we prepare a Chinese image caption model with high-quality image-caption pairs to bootstrap training data refinement. With curated high-quality images and synthesized descriptive captions, we then train a Chinese text-to-image generation model, the largest to date, using SDXL and a 10-billion multimodal text encoder. Specifically, we introduce a two-stage bilingual multimodal representation alignment process to seamlessly integrate the text encoder with the generation model. Extensive experiments validate the effectiveness of our framework, including assessments of image captioning and image generation. The implementation of our framework in Baidu Search Ads shows significant revenue increase, For example, beauty industry ads with generated image creatives achieve a 29% higher click-through rate (CTR).

SIGIR '24, July 14–18, 2024, Washington, DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3661350

CCS CONCEPTS

• **Information systems → Multimedia content creation**.

KEYWORDS

Text-to-Image Generation; Multimodal Sponsored Search; Advertisement Image Creatives

ACM Reference Format:
Kang Zhao, Xinyu Zhao, Zhipeng Jin, Yi Yang, Wen Tao, Cong Han, Shuanglong Li and Lin Liu. 2024. Enhancing Baidu Multimodal Advertisement with Chinese Text-to-Image Generation via Bilingual Alignment and Caption Synthesis. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3661350

1 INTRODUCTION

With the widespread of rich media on the Internet, the e-commerce industry increasingly adopts multimodal information mediums, such as engaging visual content, to capture consumer attention. A typical example is the multimodal advertisement. Multimodal advertisement is a type of sponsored search, where the search engines charge advertisers for displaying their ads alongside search results [6]. Multimodal advertisements upgrade traditional text ads to multimedia ad containers with visual ad creatives [25, 27].

Traditional ad image creatives are from advertisers or the database of search ad platforms, exhibit variability in specifications and quality. It is a time-consuming process to design images that not only meet platform requirements but also appeal to users. Meanwhile, search platforms undertake data pre-processing to standardize image sizes, filter low-quality images, and rectify possible mismatched uploaded ad images and text. These challenges lead to the

*Equal contribution.
†Corresponding author.

development of a real-time ad image retrieval system, which represents the varied distribution of image data and realizes cross-modal representation alignment for efficient text-to-image retrieval [26].

Recent advancements in generative AI have facilitated more efficient and intelligent content creation for search and advertising systems. A key innovation in this domain is the text-to-image generation models [23], which offer unparalleled flexibility in image creation by enabling users to guide the image generation process via natural language. Some diffusion-based models, such as Stable Diffusion XL and DaLL-E 3, have demonstrated remarkable generated image authenticity [5, 16]. Additionally, techniques like DreamBooth and ControlNet have enhanced the adaptability and efficiency of image generation models, allowing for the generation of content in specific styles and previously unseen objects [18, 30]. Leveraging the generation model for advertising image creatives presents three benefits. Firstly, it reduces the time and effort required by advertisers, while providing substantial diversity in content creation, including background and entire image generation. Secondly, by training image generation models on high aesthetic quality data, we can significantly enhance the visual appeal of generated ad images, notably in the travel and beauty industry. Thirdly, the standardized output format of generation models simplifies the pre-processing on the part of search advertising platforms.

In this study, we introduce a framework for generating ad image creatives, focusing on enhancing image attractiveness and ensuring consistency with Chinese descriptions. The generated examples are shown in Figure 1. To ensure data quality that is critical for the training generation model, we first curate a set of commercial image-text pairs to train an image captioning model generating descriptive captions in Chinese. Subsequently, we develop the largest Chinese text-to-image generation model to date, building on SDXL and our Chinese text encoder of a 10-billion scale. We implement a two-stage bilingual training, aimed at integrating our Chinese text encoder and the diffusion backbone. This begins with bridging between our text encoder and original English encoder on bilingual texts and image pairs, followed by unified training of text encoder and diffusion model on high-quality images and Chinese captions. After finishing training, we assess the effectiveness of our framework by examining the quality of the generated captions and images, and its impact on online revenue. The key contributions of our work are outlined as follows:

- We have developed the largest Chinese text-to-image generation model to date, incorporating a 10-billion parameter text encoder and the most powerful generation backbone with bilingual alignment. It surpasses its counterparts in producing authentic images closely aligned with textual descriptions, as evidenced by both quantitative metrics and human preference assessments.
- We crafted an image captioning model specifically tailored for Chinese commercial contexts, achieving superior performance over alternative texts and general captioning models in terms of descriptiveness and human preference.
- Deployed within Baidu Search Ads, our generative image creative service has notably increased online revenue, particularly in the beauty industry, where ads featuring our generated image creatives have seen a 29% uplift in click-through rate (CTR).

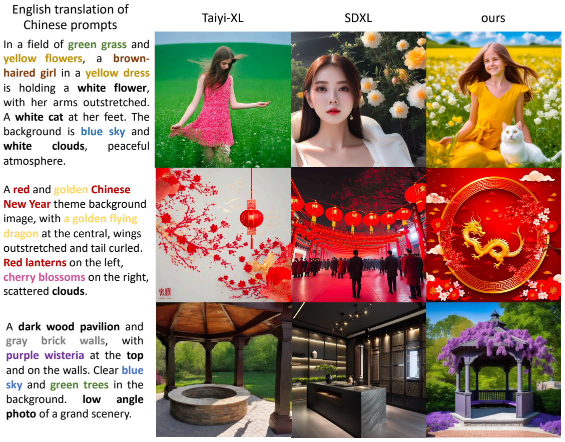
English translation of Chinese prompts
Taiyi-XL
SDXL
ours
In a field of green grass and yellow flowers, a brown-haired girl in a yellow dress is holding a white flower, with her arms outstretched. A white cat at her feet. The background is blue sky and white clouds, peaceful atmosphere.
A red and golden Chinese New Year theme background image, with a golden flying dragon at the central, wings outstretched and tail curled. Red lanterns on the left, cherry blossoms on the right, scattered clouds.
A dark wood pavilion and gray brick walls, with purple wisteria at the top and on the walls. Clear blue sky and green trees in the background. low angle photo of a grand scenery.

Figure 1: Examples of Chinese Text-to-Image Generation. The prompts are English translations. Objects and colors are highlighted.

2 RELATED WORKS

Multimodal Search Ads. Search ad system is by nature a recommendation system, entailing processes of candidate generations and ranking [9]. Multimodal search ads add a multimodal ad creatives retrieval module, designed to pair each text ad with a corresponding image based on real-time queries [21]. Previous studies on multimodal search ads or close scenarios such as fashion and e-commerce have predominantly concentrated on cross-modal representation learning [11, 31, 32]. However, there is scant research regarding the application of text-to-image generation within multimodal search ads. Baidu as a prominent search engine in China, has adopted the latest advancements in multimodal research within its search ad system, showcasing significant progress [21, 28, 29].

Text-to-image Generation. The past few years have seen significant advancements in text-to-image generation technologies, with the introduction of various models such as autoregressive models [17, 24], GANs [2, 10], and VAEs [14]. The dominant architecture now is the diffusion model, with some pilot studies incorporating Transformer in image generation [4, 7, 15]. A key insight from recent studies is the crucial role of high-quality image-text datasets in training effective image-generation models. For instance, DALL-E 3 demonstrates that image captioning improves the prompt following ability of image generation model [5]. Given the predominance of English-centric image-text data, the exploration of multilingual text-to-image generation presents a valuable research avenue. Previous works build upon existing generation backbone by integrating encoders that support additional languages: [8] adapts a multilingual text encoder to separately align language and image representations; [13] weighs different language encoders with an Ensemble Adapter; [22] conducts bilingual continuous pre-training of CLIP and multi-resolution denoising training.

3 METHODOLOGY

We propose a pipeline of Chinese text-to-image generation, including dataset creation, model architecture, and training methodologies, shown in Figure 2. This section is divided into two parts: 1) Data preparation: we propose a multimodal large language model to produce descriptive captions for the generation training dataset.

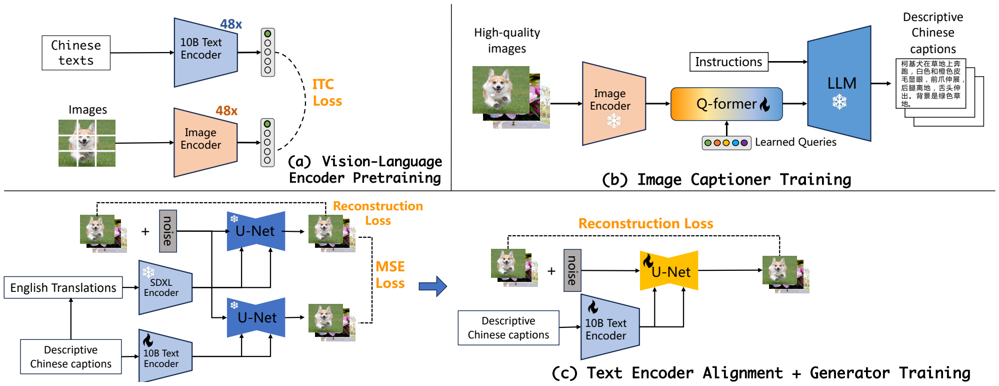
Chinese texts
10B Text Encoder
48x
ITC Loss
Images
Image Encoder
48x
(a) Vision-Language Encoder Pretraining
High-quality images
Image Encoder
Instructions
Q-former
LLM
Learned Queries
Descriptive Chinese captions
柯基犬在草地上奔跑，白色和橙色皮毛显眼，前爪伸展，后腿离地，舌头伸出。背景是绿色草地。
(b) Image Captioner Training
+
noise
U-Net
Reconstruction Loss
English Translations
SDXL Encoder
U-Net
MSE Loss
Descriptive Chinese captions
10B Text Encoder
Reconstruction Loss
+
noise
U-Net
Descriptive Chinese captions
10B Text Encoder
(c) Text Encoder Alignment + Generator Training

Figure 2: Overview of our ad image creatives generation framework. (a) Training a vision-language foundation model on advertising data by contrastive learning. (b) Generating Chinese captions with a multimodal LLM fine-tuned on image-conditioned soft prompts. (c) Aligning Chinese text encoders with SDXL through a two-step training, first only train our text encoders, then proceed to unfreeze UNet.

2) Image generation model: based on the latent diffusion model, we incorporate a Chinese text encoder, facilitating the reconciliation of disparities in Chinese and English text-to-image generation.

3.1 Dataset Construction

It is widely acknowledged that open-web datasets, commonly used for training text-to-image models, exhibit certain deficiencies. For instance, LAION [19] relies on alternative HTML tags (alt text) that typically cover limited facets of images, neglecting background details and object interactions. Furthermore, some alt text is inaccurate and contextually irrelevant to the corresponding image. The problems lead to the necessity of data refinement for text-to-image generation, particularly for Chinese texts, given that most alt texts are in English. Thus, we prioritize the creation of a training dataset enriched with high-quality Chinese image captions.

We first establish criteria for high-quality image captions, emphasizing the comprehensiveness and accuracy of the information about the background, primary subjects, and their interrelations within the image. According to the criteria, we engage human annotators to manually generate captions for 10K images from our commercial image dataset. Then, we fine-tune Qwen-7B-VL [3], a multimodal large language model, with soft prompt learning [12] on the annotations, following specific instructions: "Provide detailed descriptions of the image in Chinese limited to 100 words, covering the background (including scene and style), primary subjects (specifying type, quantity, color, etc.), and the relationships among these subjects. Avoid any subjective interpretation." Meanwhile, we use the image encoder from our vision-language foundation model (Figure 2(a)) to infer image features. The image features and learnable queries are fed into a cross-attention module, Q-former, to project image features into language representations, which is then fed into a large language model (LLM) along with the instruction. During fine-tuning, we update only the Q-former. Then we apply the captioning model for large-scale, high-quality aesthetic images, including 1B commercial images. We further construct an English version of the Chinese captions using Baidu translation service*, as bilingual data for the following encoder alignment.

3.2 Bilingual Alignment of Text Encoders

Our generation model adopts the text-conditioned latent diffusion model, including a Variational Auto-Encoder (VAE), a text encoder, and UNet, all initialized from SDXL. However, SDXL is mainly trained on English datasets thus performing poorly on Chinese prompts. And its text encoders struggle with long texts exceeding 77 token limits. To tackle these problems, we substitute the original text encoder with our 10B-parameter language model from our vision-language foundation model. As shown in Figure2(a), the vision-language model is pre-trained via image-text contrastive learning (ITC) on Chinese image-text pairs. This adjustment enhances the generation model in semantic comprehension of extensive Chinese texts, which also makes it the largest (13B-parameter) among Chinese text-to-image generation models.

The major training objective of the generation model follows the design of the latent diffusion model. Given a text input y, the text-to-image diffusion models learn conditional distributions of $p(z \mid y)$, where a conditional denoising auto-encoder $\epsilon_\theta\left(z_t, t, y\right); t \in \{1, \ldots, T\}$ is used to model the reverse process of a fixed Markov Chain of length T. Let τ_θ denote the text encoder, $\mathcal{E}$ denote a AE for mapping image to latent features, and ϵ_θ denote the time-conditional UNet, the reconstruction loss can be formulated as:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right] \quad (1)$$

The text encoder collaborates with UNet to exert conditional control over the image generation process. A misalignment arises due to the incompatibility between the initial UNet and the Chinese text encoder. Considering the advanced English text-to-image capabilities of SDXL, we initially freeze the entire UNet and VAE and exclusively train our text encoder. We add an auxiliary alignment loss L_{MSE} to L_{LDM} in this training stage. Especially, the previously generated Chinese captions and their English translations are fed into our Chinese text encoder τ_θ and SDXL text encoders τ'_θ respectively. A mean square error (MSE) loss is then calculated based on the noise vectors generated under the conditions by both encoders. This alignment loss facilitates the transfer of SDXL text encoder capabilities to the Chinese text encoder, as is shown in Figure 2(c).

$$L_{MSE} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon_\theta(z_t, t, \tau'_\theta(y)) - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \quad (2)$$

[*] https://fanyi.baidu.com/

Once the Chinese text encoder is adequately aligned, SDXL is discarded and our UNet is unlocked and concurrently trained. Mirroring the approach of SDXL, this training phase is executed across multiple image resolutions$\in \{256, 512, 1024\}$.

4 EXPERIMENTS

4.1 Results of Image Captioning

To validate our image captioning method, we assess caption metrics and conduct a manual evaluation on 300 images alongside their annotated captions from the test dataset. This analysis is benchmarked against three baselines: alt text, BLIP2 with Chinese translation [12], and Qwen-7b-VL [3].

For caption evaluation, CIDER [20] and SPICE [1] are applied to assess the alignment of generated captions with actual contents. As shown in Table 1, our fine-tuned captioning model markedly surpasses comparative baseline models. In human evaluation, annotators evaluate the comprehensiveness and accuracy of generated captions, assigning a 3-tier ratings score (0 for poor-quality, 1 for satisfactory, and 2 for high-quality captions). The human quality ratings are detailed in Table 2, confirming that captions generated by our model exhibit a marked quality improvement.

Table 1: Quantitative evaluation of synthesized Caption Quality.

Metric	alt text	BLIP2	Qwen-7b-VL	Ours
CIDER ↑	20.5	110.3	126.9	**144.6**
SPICE ↑	3.2	12.5	15.2	**20.7**

4.2 Results of Text-to-Image Generation

The assessment of our text-to-image generation encompasses both quantitative metrics and qualitative evaluations, focusing primarily on performance with complex Chinese captions. We selected 300 Chinese descriptive text prompts from the test set, comparing the generation outcomes of our model with the prominent Chinese text-to-image model, Taiyi-XL [22] and SDXL with Chinese inputs [16].

Table 3 presents quantitative metrics including FID, CLIP score, and aesthetic score: FID quantifies the semantic content distribution similarity between generated images and original images; the CLIP score evaluates the alignment between the prompt and its generated image; and the aesthetic score evaluates the visual appeal of the generated images. It demonstrates that our model consistently produces the highest quality images. The results validate the overall superior performance of our model in semantic relevance to the original image, text prompt, and visual appeal of generated images.

In the human evaluation of generated content, we compare our model to SDXL in terms of text adherence and visual appeal. Annotators are asked to assign scores from a set of three options 0, 1, 2 similar to Section 4.1. The findings, presented in Table 4, reveal that images produced by our model exhibit greater consistency with text prompts and superior attractiveness compared to those generated by SDXL. The cases in Figure 1 also show that our model is capable of generating images with higher element completeness, attribute accuracy, and visual appeal.

4.3 Ablation Studies

The impact of high-quality long captions, the utilization of a large text encoder, and the 2-stage alignment training strategy are analyzed using the CLIP score as the primary metric, with results detailed in Table 5. A comparative analysis reveals that models trained with descriptive captions exhibit enhanced text fidelity over those trained on alt text. In terms of the text encoder, substituting our 10B text encoder with a smaller CNCLIP text encoder results in performance decline. Moreover, the absence of bilingual alignment significantly reduces the CLIP score, underscoring its importance.

Table 2: Human evaluation of synthesized captions.

Method	Comprehensiveness			Accuracy		
	0	1	2	0	1	2
altext	65.3	19.4	15.6	69.9	15.8	14.3
BLIP2	45.9	34.7	19.4	47.4	21.9	30.6
Qwen-7b-VL	6.2	34.4	60.4	25.3	38.5	37.2
Ours	7.1	30.6	62.2	8.7	29.1	62.2

Table 3: Results of Text-to-Image Generation Metrics.

Method	FID ↓	CLIP Score ↑	Aes. Score ↑
Taiyi-XL	72.5	23.4	5.77
SDXL	76.2	21.6	5.82
Ours	**69.8**	**24.5**	**5.89**

Table 4: Human evaluation of generated image quality from text following and appealing perspectives.

Method	Text adherence			Visual appeal		
	0	1	2	0	1	2
SDXL	2.7	40.2	57.1	15.7	33.2	51.2
Ours	2.2	20.2	77.6	15.6	25.2	59.2

Table 5: Ablation studies of synthesis captions, different text encoders, and text encoder alignment (align.).

Method	CLIP Score ↑
Ours	**24.5**
alt text	19.7
CNCLIP	22.1
w/o align.	20.6

4.4 Online A/B Test

Our text-to-image generation model offers high-quality creatives for search advertising. To assess its efficacy, we implemented the model within the beauty industry of the Baidu search ads system. A two-week A/B testing period revealed that the relevant and visually appealing images generated by our model contributed to a significant increase in click-through rate by 29.1%.

5 CONCLUSION

Our study introduces a streamlined framework for generating ad image creatives, comprising: a captioning model generating high-quality Chinese descriptions, and a text-to-image model utilizing synthesized data and a bilingual encoder alignment. We assess our framework based on data quality, generation efficacy and commercial values, demonstrating superiority over conventional models in ad-related tasks.

SPEAKERS BIO

Kang Zhao is an algorithm engineer at Baidu Search Ads. His research interests include text-to-image generation and multimodal content understanding. Zhao currently works on the construction of ad image generation model and its applications.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. *ArXiv* abs/1607.08822 (2016).

[2] Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *ArXiv* abs/2308.12966 (2023).

[4] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2022. All are Worth Words: A ViT Backbone for Diffusion Models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 22669–22679.

[5] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. [n. d.]. Improving Image Generation with Better Captions.

[6] Christian Borgs, Jennifer T. Chayes, Nicole Immorlica, Kamal Kumar Jain, Omid Etesami, and Mohammad Mahdian. 2007. Dynamics of bid optimization in online advertisement auctions. In *The Web Conference*.

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. *ArXiv* abs/2301.00704 (2023).

[8] Zhongzhi Chen, Guangyi Liu, Bo Zhang, Fulong Ye, Qinghong Yang, and Ledell Yu Wu. 2022. AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities. *ArXiv* abs/2211.06679 (2022).

[9] Paul Covington, Jay K. Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems* (2016).

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[11] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. 2023. Learning Instance-Level Representation for Large-Scale Multi-Modal Pretraining in E-Commerce. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 11060–11069.

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*.

[13] Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulic, and Anna Korhonen. 2023. Translation-Enhanced Multilingual Text-to-Image Generation. In *Annual Meeting of the Association for Computational Linguistics*.

[14] Romain Lopez, Pierre Boyeau, Nir Yosef, Michael I. Jordan, and Jeffrey Regier. 2020. AUTO-ENCODING VARIATIONAL BAYES.

[15] William S. Peebles and Saining Xie. 2022. Scalable Diffusion Models with Transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2022), 4172–4182.

[16] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *ArXiv* abs/2307.01952 (2023).

[17] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *ArXiv* abs/2102.12092 (2021).

[18] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 22500–22510.

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *ArXiv* abs/2210.08402 (2022).

[20] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. CIDEr: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), 4566–4575.

[21] Zhoufutu Wen, Xinyu Zhao, Zhipeng Jin, Yi Yang, Wei Jia, Xiaodong Chen, Shuanglong Li, and Lin Liu. 2023. Enhancing Dynamic Image Advertising with Vision-Language Pre-training. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2023).

[22] Xiaojun Wu, Di Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2024. Taiyi-Diffusion-XL: Advancing Bilingual Text-to-Image Generation with Large Vision-Language Model Support. *ArXiv* abs/2401.14688 (2024).

[23] Ling Yang, Zhilong Zhang, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. 2022. Diffusion Models: A Comprehensive Survey of Methods and Applications. *Comput. Surveys* 56 (2022), 1 – 39.

[24] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Trans. Mach. Learn. Res.* 2022 (2022).

[25] Tan Yu, Zhipeng Jin, Jie Liu, Yi Yang, Hongliang Fei, and Ping Li. 2022. Boost CTR Prediction for New Advertisements via Modeling Visual Content. *2022 IEEE International Conference on Big Data (Big Data)* (2022), 2140–2149. https://api.semanticscholar.org/CorpusID:252519593

[26] Tan Yu, Jie Liu, Zhipeng Jin, Yi Yang, Hongliang Fei, and Ping Li. 2022. Multi-scale Multi-modal Dictionary BERT For Effective Text-image Retrieval in Multimedia Advertising. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (2022).

[27] Tan Yu, Jie Liu, Yi Yang, Yi Li, Hongliang Fei, and Ping Li. 2022. EGM: Enhanced Graph-based Model for Large-scale Video Advertisement Search. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022).

[28] Tan Yu, Xuemeng Yang, Yan Jiang, Hongfang Zhang, Weijie Zhao, and Ping Li. 2021. TIRA in Baidu Image Advertising. *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (2021), 2207–2212.

[29] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021. Heterogeneous Attention Network for Effective and Efficient Cross-modal Retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021).

[30] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 3813–3824.

[31] Xiaoyang Zheng, Fuyu Lv, Zilong Wang, Qingwen Liu, and Xiaoyi Zeng. 2023. Delving into E-Commerce Product Retrieval with Vision-Language Pre-training. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2023).

[32] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Hao Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-BERT: Vision-Language Pre-training on Fashion Domain. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 12642–12652.