Large Vison-Language Foundation Model in Baidu AIGC Image Advertising

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

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

Yafei Li
Baidu Inc.
Beijing, China
liyafei03@baidu.com

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

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

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

Lin Liu
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. Alignment learning between small models and parallel corpora cannot meet current needs. The success of ChatGPT demonstrates that large models need to first establish a fundamental understanding, and then utilize high-quality corpora for generation. Having a large model foundation is indispensable. In this paper, we establish a fundamental 10B multimodal model foundation for multimodal generation tasks and propose a scene-based alignment learning approach called conditional sample supervised fine-tuning for downstream generation tasks. Meanwhile, diffusion models are known to be vulnerable to outliers in training data. To address this, we utilize an alternative diffusion loss function that preserves the high quality of generated data like the original squared L2 loss while being robust to outliers.In practical test sets, the multimodal foundation fully demonstrates its alignment and comprehension abilities for graphic and textual content. Additionally, conditional fine-tuning and the design of the loss function significantly enhance the quality of generated content. The quality rate of images has increased by 34.3 percentage points, and prompt control has improved by 19.8 percentage points. The application of our framework in Baidu Search Ads has led to significant revenue growth. For instance, ads with generated image creatives have achieved a 29% higher click-through rate (CTR), resulting in a daily consumption of 3 million yuan.

*Equal contribution.
†Corresponding author.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1245-6/25/08
https://doi.org/10.1145/3690624.3709401

CCS Concepts

• **Information systems** → **Image search**; **Multimedia content creation**.

Keywords

Cross-modal Retrieval;Text-to-Image Generation; Multimodal Sponsored Search; Advertisement Image Creatives

ACM Reference Format:
Zhipeng Jin, Wen Tao, Yafei Li, Yi Yang, Cong Han, Shuanglong Li, and Lin Liu. 2025. Large Vison-Language Foundation Model in Baidu AIGC Image Advertising. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3690624.3709401

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 [1]. Multimodal advertisements upgrade traditional text ads to multimedia ad containers with visual ad creatives [2, 3].

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 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 [4].

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 [5], 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 [6, 7]. 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 [8, 9]. 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.

Nevertheless, three principal challenges currently impede the development of Chinese text-to-image generation. Firstly, the prompt's capacity to follow is relatively limited, and it is often unable to generate images that meet user input requirements for lengthy texts. It is hypothesised that the primary cause is the limited capacity of the text encoder model, which is unable to accurately comprehend lengthy texts. Secondly, the current mainstream text-to-image models employ a multi-stage training approach, typically comprising two stages: pre-training on a large number of low-resolution images and post-training on a limited number of high-resolution datasets. The limitation of this approach is that the small amount of data trained in the final stage results in a reduction in the model's generalisation ability and an increased risk of overfitting. In terms of objective optimisation, the most commonly used method is the MSE L2 loss, which has the disadvantage of being highly sensitive to outliers and prone to falling into local optima.

In order to address these challenging issues, a framework for the generation of creative advertising images has been introduced, with a particular focus on the enhancement of image appeal and the assurance of consistency with Chinese descriptions. Firstly, we draw inspiration from the training and fine-tuning of large language models, where the upper limit of fine-tuning effectiveness depends on the capabilities of the underlying large model. It is therefore evident that in order to further enhance text control, a text encoder with a larger capacity, stronger capability and fully aligned with images is required. Consequently, a text image pre-training model with parameter levels exceeding 10 billion was initially trained using a substantial corpus of text-image pairs, and the text encoder was subsequently integrated into the text image model. Subsequently, the CFT (conditional fine-tuning) method was employed for sample organisation, whereby high-quality and medium-quality data were trained in a single stage, thus ensuring the quality of image generation without compromising the model's generalisation ability. Ultimately, Huber loss was employed to address the challenge of model training instability resulting from sample outliers. The principal contributions of this research are outlined below:

- A large pre-trained model with parameters exceeding 10b was trained, thereby achieving a new level of text and image representation ability in the Chinese text-image retrieval field.
- A 10-billion parameter text encoder was incorporated into the text2image model, and the CFT method and Huber loss were introduced to enhance the text2image model's generalisation ability, thereby surpassing its counterparts in the generation of realistic images that closely match the text description. This has been demonstrated by both quantitative metrics and human preference assessments.
- The deployment of our generative image creative service within Baidu Search Ads has notably increased online revenue, particularly in the beauty industry, where ads featuring our generated image creatives have demonstrated a 29% uplift in click-through rate (CTR).

2 Related Works

Multimodal Search Ads. Search ad system is by nature a recommendation system, entailing processes of candidate generations and ranking [10]. 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 [11]. Previous studies on multimodal search ads or close scenarios such as fashion and e-commerce have predominantly concentrated on cross-modal representation learning [12–14]. 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 [11, 15–17].

Vision-Language Pre-training. The Vision-Language Pre-trained (VLP) model learns to model visual and textual information and their interaction on large scale image-text pairs. Early VLP models follow a single-encoder structure, which handles different modalities inputs with a single transformer [18–21]. The dual-encoder model [22, 23], is to train two separate encoders for contrastive learning. Later studies combine the two structures [24, 25]. In recent years, VLP models with decoder have excelled in generative tasks [26–28]. Recent research further combines large language models to build large multimodal models [29–32]. Most VLP models are trained on English datasets, while some other studies focus on building Chinese VLP models and image-text datasets [33–37]. In sum, VLP has evolved from simply following the language model to constructing more complex training patterns to maximize the capability of multimodal understanding and generation. In particular, the downstream application of VLP in this study are in text-to-image retrieval and representation learning and efficient retrieval are the most important concern, where the de-facto state-of-the-art solution is dual-encoder model.

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 [38, 39], GANs [40, 41], and VAEs [42]. The dominant architecture now is the diffusion model, with some pilot studies incorporating Transformer in image generation [43–45]. A key insight from recent studies is the crucial role of high-quality image-text datasets in training effective image-generation models. For instance,

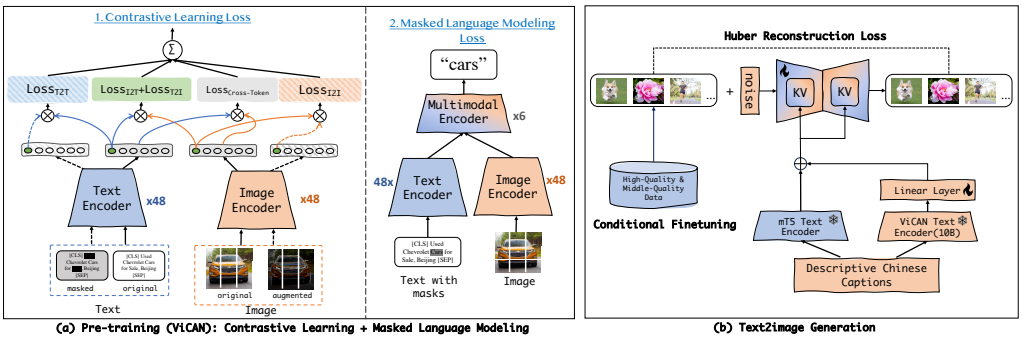

Figure 1: Overview of the two stage-learning of the text2image generation. (a) Vision-Language Pre-training: We train dual encoders jointly using cross-view and cross-token contrastive losses between image and text (Loss_{I2T} and Loss_{T2I}), image and augmented image (Loss_{I2I}), text and masked text (Loss_{T2T}), and image patches and text tokens ($\text{Loss}_{\text{Cross-token}}$). Additionally, we compute a Masked Language Modeling loss using a cross-modal encoder. (b) Text2image Generation: We integrate the fully text-image aligned 10b text encoder with the text2image model, and introduce CFT and Huber loss to improve training stability and generalization ability.

DALL-E 3 demonstrates that image captioning improves the prompt following ability of image generation model [6]. 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: [46] adapts a multilingual text encoder to separately align language and image representations; [47] weighs different language encoders with an Ensemble Adapter; [48] conducts bilingual continuous pre-training of CLIP and multi-resolution denoising training.

3 Methodology

In order to improve the overall performance of the Chinese text generated image model in Baidu's commercial scenario, we conducted two-stage learning: vision-language pretraining and text2image generation. We found that pretraining a large vision-language model is crucial for the learning of text2image models. Therefore, we first implemented a VLP basic model based on CLIP (referred to as ViCAN), which is used for general multimodal representation learning to improve the general text and visual alignment ability. The basic model is initialized with single-modal pre-trained models. To promote coarse to fine image text alignment and cross modal interaction, we enhance traditional image text contrastive learning through cross view and cross label contrastive loss, as well as auxiliary visual enhancement language modeling loss. In the stage of the text2image generation, we adopted a dual text encoder structure and integrated the pretrained 10B parameter text encoder into the diffusion process. Due to the alignment of visual semantics with the text encoder, it can better interact with the image diffusion process. In addition, we have introduced the CFT mechanism, which ensures high-quality data fine-tuning through data quality classification, while also introducing a large amount of medium quality data to ensure model generalization performance. Finally, we introduced Huber loss to penalize the problem of excessive L2 loss errors caused by abnormal samples, making the model more robust. See Figure 1 for an overview of the framework.

3.1 Vision-language Pre-training

Preliminaries. Image-text contrastive learning (ITC), which models the alignment of vision and language features, has been popularized by CLIP and has proven to be one of the most effective pre-training tasks for learning vision-language representations. The key idea of ITC approach is to encourage the vision encoder and the text encoder to produce similar representations from corresponding image-text pairs (positive) and, conversely, to maximize the feature distance between non-corresponding pairs (negative). The output features from two encoders are normalized to compute embedding similarity by:

$$s(x, y) = (\boldsymbol{h}_x)^\top (\boldsymbol{h}_y) \tag{1}$$

where $\boldsymbol{h}$ is the output embeddings, and x and y denote samples from different modalities image I and text T. Thus, $s(x, y)$ represents the similarity between x and y measured by the dot product of their hidden embeddings. Then a softmax-normalized similarity score for each paired image x_i and text y_i is computed as:

$$S_{x2y} = \frac{\exp(s\,(x_i, y_i)\,/\tau)}{\sum_{j=1}^{N} \exp(s\,(x_i, y_j)\,/\tau)} \tag{2}$$

where i and j denote data index and N denotes the batch-size. τ is a learnable temperature parameter. Taking all the pairs from the same batch into account and to maximize similarity scores of N positive pairs while minimizing those of $N^2 - N$ negative pairs, the InfoNCE loss is calculated as:

$$\mathcal{L}_{x2y} = -\frac{1}{N} \sum_{i=1}^{N} S_{x2y} \tag{3}$$

Cross-View Learning. We further enhance ITC by incorporating single-modal contrastive losses to promote the discrimination of similar images or texts, which is interpreted as learning both inter-modal and intra-modal correlations to capture diverse semantic information [49]. To implement cross-view contrastive learning, we extend ITC loss into four components: image-to-image (I2I) loss, text-to-text (T2T) loss, and original text-to-image (T2I) and image-to-text (I2T) losses, which are formulated as:

$$\mathcal{L}_{\text{cross-view}} = \mathcal{L}_{\text{I2I}} + \mathcal{L}_{\text{T2T}} + \mathcal{L}_{\text{I2T}} + \mathcal{L}_{\text{T2I}} \tag{4}$$

To construct input samples for cross-view learning, we apply RandomAugment [50] to images and random masking to texts in each mini-batch. The generated images and texts, paired with their original counterparts, are positive samples for I2I and T2T contrastive learning.

Cross-Token Learning. The basic ITC and cross-view contrastive learning are at instance-level. There are studies [24, 51] showing the need to learn more fine-grained content understanding. This finding also applies to image advertising systems, which need to capture keywords in noisy user queries. Therefore, we employ cross-token contrastive learning, regarding image patches as visual tokens, to enable the model to capture fine-grained token-level semantics and build more complex interactions between text and image data. We introduce a cross-token contrastive loss, where the instance-level ITC loss is incorporated by the weighted token-level similarity scores. Specifically, n and m denote the number of patches of the image and words of the text embeddings. We first compute the fine-grained similarities across all the tokens using Equation (1), which denotes as $s(p, w) \in \mathbb{R}^{n \times m}$ and p and w denotes image patches and text words. Then we calculate an aggregation attention map using token similarities. The normalized token-level image and text similarity vectors are first computed as:

$$S_{i2w} = \sum_{i=1}^{n} \frac{\exp(s(p,w)_{(i,*)}/\tau)}{\sum_{j=1}^{n} \exp(s(p,w)_{(j,*)}/\tau)} s(p,w)_{(i,*)} \tag{5}$$

$$S_{t2p} = \sum_{i=1}^{m} \frac{\exp(s(p,w)_{(*,i)}/\tau)}{\sum_{j=1}^{m} \exp(s(p,w)_{(*,j)}/\tau)} s(p,w)_{(*,i)} \tag{6}$$

where the token-level image similarity vectors $S_{i2w} \in \mathbb{R}^{1 \times m}$ represents the similarity score between the image and m words of the text, and the token-level text similarity vector $S_{t2p} \in \mathbb{R}^{n \times 1}$ represents the similarity score between the text and n patches of the image. Then the fine-grained image and text similarity scores are computed by normalizing the similarity vectors:

$$S_{\mathrm{I}} = \sum_{i=1}^{m} \frac{\exp(S_{i2w(1,i)}/\tau)}{\sum_{j=1}^{m} \exp(S_{i2w(1,j)}/\tau)} S_{i2w(1,i)} \tag{7}$$

$$S_{\mathrm{T}} = \sum_{i=1}^{n} \frac{\exp(S_{t2p(i,1)}/\tau)}{\sum_{j=1}^{n} \exp(S_{t2p(j,1)}/\tau)} S_{t2p(i,1)} \tag{8}$$

Then the cross-token image-text similarity score is the average of the above two scores and the cross-token contrastive learning loss $\mathcal{L}_{\text{cross-token}}$ is obtained by aggregating the cross-token image-text similarity:

$$\mathcal{L}_{\text{cross-token}} = -\frac{1}{N} \sum_{i=1}^{N} (S_{\mathrm{I}} + S_{\mathrm{T}})/2 \tag{9}$$

The final contrastive loss $\mathcal{L}_{\mathrm{CL}}$ is the combination of cross-view and cross-token contrastive losses:

$$\mathcal{L}_{\mathrm{CL}} = \mathcal{L}_{\text{cross-view}} + \mathcal{L}_{\text{cross-token}} \tag{10}$$

We add Masked Language Modeling (MLM) loss to improve the ability to enhance the interaction between vision and language features. The image features are fused in a multimodal encoder taking the output of the text encoder. The MLM loss is designed to minimize the predictions of the masked tokens with the 15% dropped tokens of input text. Let $\hat{t}$ denotes masked text input, $\boldsymbol{p}^{\mathbf{mask}}(i, \hat{t})$ denotes the output probability, and H be the cross entropy function, the MLM loss is computed as:

$$\mathcal{L}_{\mathrm{MLM}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{H}(\boldsymbol{y}^{\mathbf{mask}}, \boldsymbol{p}^{\mathbf{mask}}(i, \hat{t})) \tag{11}$$

The final loss function includes the contrastive loss and MLM loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CL}} + \mathcal{L}_{\mathrm{MLM}} \tag{12}$$

Model Architecture. We use a dual-encoder backbone for our vision-language model ViCAN because of its flexibility, allowing the vision and text encoders to be readily adapted to a variety of applications, either separately or in combination. Since training a large vision-language model from scratch is too expensive, we initialize our model from pre-trained weights. Specifically, the vision encoder is based on the pre-trained ViT-bigG with 1.9B parameters from OpenCLIP [52], while the text encoder is based on ERNIE-3.0 [53] with over 10B parameters.

A substantial corpus of Chinese text and image data was procured from the website and subjected to rigorous cleansing, resulting in a parallel corpus of text and image data at the 5B level. This data was leveraged to train ViCAN model with enhanced universal text and image alignment capabilities.

3.2 Text2image Generation

Preliminaries. 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:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right] \tag{13}$$

Dual Text Encoder Injection. Efficient text encoders are crucial in text to image generation as they require accurate understanding and encoding of input text prompts to generate corresponding images. The Matryoshka diffusion model, Imagen, MUSE, and Pixart-α only use T5 to enhance understanding of input text prompts. More and more work is adopting multiple text encoders to further enhance the comprehension of text. Therefore, we chose a dual text-encoder structure, using mT5 and ViCAN respectively. mT5[54] is a variant based on the T5 model, trained using corpus from 101 languages, and has good understanding ability in Chinese scenarios. As introduced in section 3.1, the ViCAN model is a pre-trained large model for images and text that we trained using large-scale parallel corpora, achieving excellent alignment between images and Chinese semantics. We integrate the text encoder with a parameter level of 10b into the text2image generation model. As shown in Figure 1, we fuse the ViCAN model with mT5 using residual linking and insert it into the diffusion network:

$$h_{merge} = h_{mT5} + LinearLayer(h_{ViCAN}) \tag{14}$$

where h_{mT5}, h_{ViCAN} denote as the last hidden states of mT5 and ViCAN respectively. To seamlessly integrate this modification within the existing Diffusion Transformer architecture, without necessitating a comprehensive retraining, we initialize $LinearLayer$ with

zero-filled weights. This approach minimizes the impact on the distribution of other structural parameters of the model during the upgrade from a single encoder to a dual encoder, while the new 10b parameters text encoder injection achieves sufficient alignment with the image during the pre-training stage, further improving the accuracy and diversity of the text to image generation process.

Conditional Fine-Tuning. Conditional Fine-Tuning[55] is widely employed in downstream fine-tuning tasks of Large Language Models (LLMs), enhancing model training effectiveness by incorporating signals such as data source and quality into the fine-tuning dataset. Conditional fine-tuning achieves selective learning from training data, enabling the acquisition of prior knowledge beneficial for downstream training tasks while avoiding the learning of irrelevant statistical data within the training set. This selective learning approach results in less forgetting and improved stability during the training process, ultimately enhancing their overall performance. Conditional fine-tuning enables selective learning of useful information, where knowledge pertinent to downstream tasks can be learned thoroughly while minimizing the learning of statistical data within the training set. This significantly reduces modifications to the pre-trained model during fine-tuning. Such selective learning achieves better stability and less forgetting in both one-time fine-tuning (transfer learning) and multiple fine-tuning (continual learning) scenarios, making it a superior alternative to traditional SFT in the continual learning of text-to-image models.

In the field of text-to-image generation, images of moderate-quality greatly assist the model in understanding semantic concepts but can adversely affect the generated images quality. To fully leverage the moderate-quality data and allow the model to thoroughly learn the semantic information of image captions, we augment the moderate-quality data with contextual descriptions of image quality. Specifically, images with an aesthetic score above 5.8 and a clip-score above 0.49 are classified as high-quality data, while those with an aesthetic score above 5.0 and a clip-score above 0.43 are considered moderate-quality data. By employing the CFT strategy, we incorporate contextual information into the captions to indicate the quality of the image using an "explain away" approach, as illustrated in Figure 2, enabling the model to selectively learn from datasets with different distributions based on the context.

Huber Loss. One-dimensional Huber loss[56] is defined as:

$$h_\delta(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \le \delta \\ \delta(|x| - \frac{1}{2}\delta) & \text{for } |x| > \delta \end{cases} \tag{15}$$

for a positive δ, Its multi-dimensional version is just a coordinate-wise sum of losses $h_{\delta_j}(x_j)$ of the corresponding components x^j of the input vector $x \in \mathbb{R}^n$. Parameters δ_j for $j = 1, \ldots, n$ can be different. When computed on the difference between the true value and its prediction by some statistical model, this function penalizes the small errors like Mean Square Error(MSE) loss and the large errors like Mean Absolute Error (MAE) loss. Thus, Huber loss penalizes the large errors caused by outliers less than MSE loss which makes it attractive for robust statistical methods[57]. It is easy to see that derivative of Huber loss is continuous, but not differentiable in points$|x| = \delta$. Pseudo-Huber loss is a more smooth function also behaving like MSE in one dimensional case in the neighbourhood of zero and like MAE in the neighbourhood of infinity:

$$H_\delta(x) = \delta^2\left(\sqrt{1 + \frac{x^2}{\delta^2}} - 1\right) \tag{16}$$

This function is defined for positive values of the parameter δ controlling its tolerance to errors with a largeL_2norm. It is extended to multi-dimensional input in the coordinate-wise manner as the standard Huber loss $h_\delta(x)$[58].

We observed that the aesthetic quality of the generated images only marginally improved after incorporating moderate-quality data through the CFT method, due to the inherent presence of low aesthetically pleasing data within the moderate-quality dataset. So we incorporate the Pseudo-Huber loss during the fine-tuning phase of the model to address the issue of outliers in the training data. In our experiments, we set the penalty coefficient δ for the Pseudo-Huber loss to 0.1 and 0.01, and the results indicated that 0.1 is a suitable value.

4 Experiments

4.1 Effectiveness of Vision-Language Pre-training

Datasets and Pre-training Implementations. The pre-training dataset is based on the curated image—text data from the system log of our online search and advertising system. The raw data is dependent on the entire dynamic advertising system, which is not only specifically designed for pairing relevant images and texts, but also considers their commercial value, such as CTR. As a result, the image-text pairs may not be fully correspondent. Additionally, the raw data comprises user queries and advertiser keywords, which contain a significant amount of noise, *e.g.*, interrogative phrases, numbers, and dates. Following [28], we perform image captioning and image-text pair filtering on several metrics. Initially, we collect all image-text pairs that have been clicked at least once in the previous year from our advertising system. First, those low-resolution images, and those with more than 20 correspondents are filtered, assuming that their information are too general or vague. Subsequently, the multimodal large language model Ernie-Bot [1] developed by Baidu, is utilized for image captioning due to its powerful understanding and generation abilities. Two captions are generated for each image. These are then re-ranked based on the similarity between the image and text, along with the original paired texts, to keep only two textual counterparts for each image. The remaining image-text pairs are then processed by our original content moderation model to exclude any NSFW content. Ultimately, we obtain a high-quality 5 billion image-text pairs for training ViCAN consuming 2048 NVIDIA A100 days. The datasets are collected from our internal systems with property protection, thus not allowed to be publicly released.

To build our ViCAN model, ViT-BigG from OpenClip and ERNIE-3.0 are initialized as vision and text encoders respectively and their details are shown in Table 1. The input image resolution is 224 and the patch-size is 14. The maximum text sequence length is set to 64 and the Chinese text tokenizer from ERNIE-3.0 is employed.

[1] https://yiyan.baidu.com/

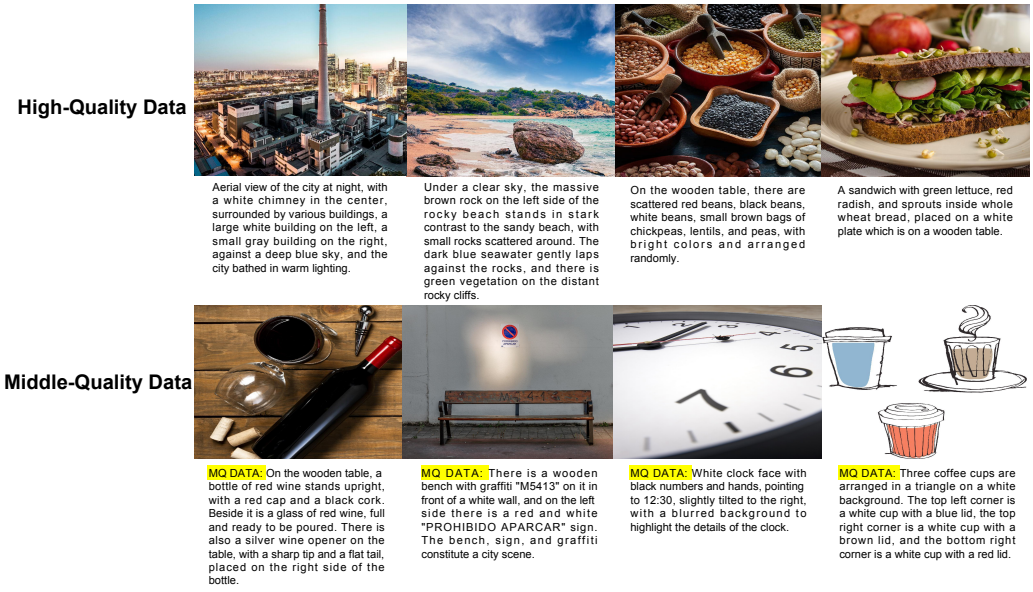

Figure 2: Illustration of conditional finetuning a text-to-image model on different quality dataset. Compared to standard finetuning, conditional finetuning prepends a context to each captions for middle quality data and only learns information conditioned on the context.

Table 1: model settings. *arch*: the initialized model, *l*: the number of transformer layers, *h*: hidden dimension, *attn*: the number of attention heads, *emb*: embedding dimension, *size*: total number of parameters.

Model	Encoder	*arch*	*size*	*l*	*h*	*attn*	*emb*
ViCAN	Vision	ViT-BigG	1.9B	48	1664	104	1024
	Text	ERNIE-3.0	10B	48	4096	64	1024

Table 2: Zero-shot performance on text-to-image retrieval on public benchmarks (MSCOCO-CN, Flickr30-CN, MUGE) and private business datasets (e-commerce, search, advertising).

Method	e-commerce			search			advertising		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
R2D2-ViT-L	27.0	53.0	62.7	17.0	38.8	48.8	17.2	38.9	49.6
$\text{Taiyi}_{\text{large}}$	42.8	68.7	76.1	22.2	47.2	56.8	23.8	51.6	62.9
$\text{CN-CLIP}_{\text{base}}$	58.9	79.7	84.8	23.6	48.9	59.0	27.2	55.9	67.5
$\text{CN-CLIP}_{\text{huge}}$	66.2	83.7	87.4	27.2	53.6	63.1	28.9	57.7	68.5
ViCAN	**69.9**	**85.8**	**90.3**	**42.4**	**71.9**	**80.8**	**37.8**	**66.5**	**75.3**
Method	MSCOCO-CN			Flickr30-CN			MUGE		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
R2D2-ViT-L	41.4	71.0	83.0	8.8	42.6	54.7	60.1	82.9	89.4
$\text{Taiyi}_{\text{large}}$	50.4	79.7	89.7	9.7	53.2	65.4	44.3	69.2	78.1
$\text{CN-CLIP}_{\text{base}}$	56.0	83.4	90.8	62.7	86.9	92.8	52.1	76.7	84.4
$\text{CN-CLIP}_{\text{huge}}$	**69.2**	**89.9**	**96.1**	**71.2**	**91.4**	**95.5**	63.0	84.1	89.2
ViCAN	64.5	88.2	94.4	51.4	82.6	93.3	**69.3**	**87.9**	**92.1**

The learning rate is warmed up to 5e-4 in the first 5000 steps and decayed to 1e-5 following a cosine schedule. The training batch-size per GPU is scaled to 352.

Results of ViCAN. The pre-training results are primarily evaluated on Chinese text-to-image retrieval benchmark and are compared with state-of-the-art Chinese vision-language pre-trained models.

Scaling VLP model achieves better representation across public and in-house datasets. We gathered evaluation data for Chinese text-to-image retrieval from public datasets, such as MSCOCO-CN [59], Flickr30k-CN [60], and MUGE [61]. In addition, we also collected industrial business data related to e-commerce, search, and advertising scenarios from our online system. The text-to-image retrieval Recall@K metrics were calculated for evaluation, where $K \in \{1, 5, 10\}$, measuring the percentage of ground-truth data in the top-K recall results. We compare ViCAN to similar-sized public VLP models CN-CLIP, Taiyi, and R2D2, as general baselines for visual-linguistic pre-training. As presented in Table 2, ViCAN outperforms general VLP baselines on business datasets by a significant margin. Additionally, it demonstrates competitive performance on public benchmarks, comparable to $\text{CN-CLIP}_{\text{huge}}$.

Table 3: Zero-shot image classification on ELEVATER and ImageNet.

Model	CIFAR10	CIFAR100	DTD	EuroSAT	FER	FGVC
$\text{CN-CLIP}_{\text{best}}$	96.0	79.7	51.2	52.0	55.1	26.2
ViCAN	**97.8**	**81.3**	**53.4**	**58.8**	**61.2**	**29.0**
Model	KITTI	MNIST	PC	VOC	ImageNet-CN	ImageNet-EN
$\text{CN-CLIP}_{\text{best}}$	49.9	79.4	63.5	84.9	59.6	32.5
ViCAN	**52.6**	**80.2**	**63.9**	**85.4**	**75.4**	**42.8**

Additionally, we experiment with bilingual image classification to show the generalization to other vision-language tasks. We use ELEVATER and bilingual ImageNet benchmarks, by constructing a

prompt with label text as the input of the text encoder. As shown in Table 3, where our model outperforms in all subsets.

Table 4: Ablation studies of multi-task learning. The methods ITC, cross-view, MLM and cross-token denotes image-text contrastive learning, cross-view contrastive loss, masked language modeling and cross-token contrastive loss respectively.

Loss	R@10(%)			
	e-commerce	search	MSCOCO-CN	Flickr30k-CN
ITC	89.8	80.2	93.0	91.6
+cross-view	90.0	80.3	94.1	92.0
+MLM	90.1	80.6	93.3	93.1
+cross-token	90.2	80.2	93.2	91.8
total	**90.3**	**80.8**	**94.4**	**93.3**

Cross-view and Cross-token Contrastive Learning consistently improve VLP. To investigate the contribution of different learning objectives, we conduct ablation studies by training DIVERSE_{12B} with different subsets of the loss function. We measure each setting of loss function by the text-to-image retrieval Recall@10 on 2 business datasets (e-commerce, search) and 2 public benchmarks (MSCOCO-CN, Flickr30k-CN). As shown in Table 4, the incorporation of cross-view learning losses, has been found to contribute to significant improvements to model performance. The token-level learning, with its advantages of more fine-grained modeling of text and image matching, also shows advantages over the method only using ITC. In summary, applying cross-view and cross-token contrastive learning strategy enhances vision-language alignment and substantially improves the model performance.

4.2 Effectiveness of Text2image Generation

Datasets and Training Implementations. It is widely acknowledged that open-web datasets, commonly used for training text-to-image models, exhibit certain deficiencies. For instance, LAION [62] 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. Therefore approximately 60 million 1024-resolution image data from commercial scenarios were obtained and used as input to the Qwen-VL [63] model, which was employed to generate Chinese captions for the images. The images were evaluated using an aesthetics model, with approximately 31.6% classified as high-quality and 68.4% as medium-quality. In the training of the text2image generation model, a one-stage training process was conducted on the entire dataset. The dual text encoder was maintained as a fixed component, while the UNet network and the zero initialization linear layer were updated. The Adam optimiser is employed, with an initial learning rate of 1e-5 and a cosine warm-up schedule. DeepSpeed Zero is employed for the purpose of accelerating the training process.

Results of Text2image Generation. To calculate the performance of our works, we use the VICAN model to calculate the

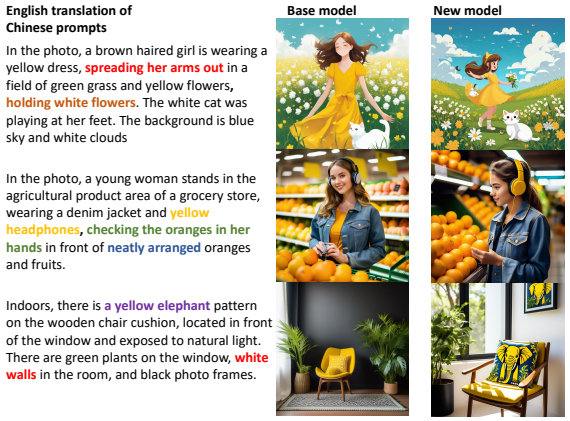

Figure 3: Examples of Chinese Text-to-Image Generation. The prompts are English translations. Objects and colors are highlighted.

ClipScore between images and texts, and employ an no reference aesthetic scoring model to quantify the exquisite level of the images. A higher score indicates better adherence to the instructions and better beauty level for generating the image. As show in Table 5, We first compared the effect of introducing dual encoders. The incorporation of a text encoder with 10b parameters has led to a notable enhancement in clip score, reaching 10.1%. This outcome suggests that the fundamental image text pre-training model has attained a substantial improvement over the text2image model. Then, we conducted a comparative analysis of different training methods between traditional SFT and CFT. Benefiting from the selective learning capability of CFT, we fully leveraged medium-quality data, with high-quality training data and medium-quality training data accounting for 31.6% and 68.4% respectively. Our model demonstrated strong performance, achieving a ClipScore improvement of +13.6% (0.4869 vs 0.4284) and an Aesthetics score increase of +2% (5.803 vs 5.671). On the other hand, we replaced MSE Loss with Huber loss to reduce the impact of outlier data, making the training process more stable and enhancing the robustness of the model. This further improved the fineness of generated images, resulting in a ClipScore increase of +0.4% (0.4889 vs 0.4869) and an Aesthetics score elevation of +2.4% (5.944 vs 5.803).

In the human evaluation of generated content, we compare our new model to the base model(single text encoder, no CFT, no huber loss) in terms of text adherence and visual appeal. Our human evaluation dataset includes 1,000 prompts generated by our captioning model from Baidu ads across various trades. We used a simultaneous evaluation method with five evaluators, assessing two dimensions: text adherence and image visual appeal, each scored on a scale of 0 to 2 (0 for poor-quality, 1 for satisfactory, and 2 for high-quality). Evaluators conducted blind reviews, and a score was considered valid if more than three evaluators agreed. The findings, presented in Table 6, reveal that images produced by our model exhibit greater consistency with text prompts and superior attractiveness compared to those generated by base model. The cases in Figure 3 also show that our model is capable of generating

Table 5: For the experimental comparison data, we conducted ablation experiments by continuously adding new optimization methods to the baseline model. The results demonstrate that the combination of VLP, CFT and Huber loss is very effective.

Model	Text Encoder	Training Data	Loss	ClipScore	Aesthetics
Base(SFT)	mT5	HQ Data	MSE_Loss	0.3888	5.621
+ Dual Encoder	mT5 + ViCAN	HQ Data	MSE_Loss	0.4284	5.671
+ Dual Encoder +CFT	mT5 + ViCAN	HQ + MQ Data	MSE_Loss	0.4869	5.803
ALL	mT5 + ViCAN	HQ + MQ Data	Huber_Loss	**0.4889**	**5.944**

Table 6: Human evaluation of generated image quality from text following and appealing perspectives.

Method	Text adherence			Visual appeal		
	0	1	2	0	1	2
base model	17.0	53.8	29.2	33.9	62.3	3.8
new model	7.7	43.3	49.0	5.0	56.9	38.1

images with higher element completeness, attribute accuracy, and visual appeal.

4.3 Online A/B Test

Table 7: Online A/B Test Performance by Image Type

Image Type	Ad Show Num	Ad Click Num	Ad CTR
Self-uploaded	217,554	8,341	3.83%
AIGC	307,338	15,214	4.95% **(+29.1%)**

Table 8: Ad Performance Comparison

Group	Shows	Clicks	CTR / Rev
Control	2.73M	126K	4.63% / 2.05M
Experimental	2.73M	128K	4.67% / 2.11M (+2.9%)

Our text-to-image generation model offers high-quality creatives for search advertising. To assess its efficacy, we identified a subset of clients from the medical aesthetics sector to whom we presented high-quality portrait image candidates generated by our model. We utilized 10% of the traffic for experimental observation. In the experimental group, the advertisements included both self-uploaded images and AIGC-generated images, while the control group exclusively featured self-uploaded images. Initially, we compared the self-uploaded images and the AIGC-generated images within the experimental group. Our findings revealed that AIGC images achieved a 29.1% higher click-through rate compared to the self-uploaded images, as illustrated in the Table 7.

Additionally, we compared the overall performance of the experimental group and control group for this subset of clients. As illustrated in Table 8, the higher click-through rate of AIGC images resulted in a 0.93% increase in the click-through rate of the experimental group compared to the control group. Furthermore, there was a corresponding increase of 2.91% in overall commercial revenue. This further validates the commercial value of our approach for Baidu's image-text advertising.

5 Deployment Details

During model deployment, the inference time associated with text-to-image generation often serves as a bottleneck. Consequently, we adopted an offline deployment strategy. We collect the advertising text information submitted by clients on a daily basis to generate images. These images undergo a review system to filter out low-quality and high-risk content before being stored in an online key-value (KV) database. In this database, the key corresponds to the client ID, while the value represents the AIGC-generated candidate images. Through this approach, we effectively integrated the AIGC images into the online system. When a query request matches a relevant client, the corresponding candidate AIGC images are returned, and the system employs a funnel ranking mechanism to select the optimal images for display in the front-end advertising slots.

6 Conclusions

The objective of this paper is to present a novel framework for text-to-image generation. The framework is comprised of two distinct components. The initial stage of the process entails training a text encoder with a parameter count of up to 10B. This utilises a vast and universal corpus of image and text data with the objective of achieving basic semantic alignment between text and images. In the second part of the process, the 10B text encoder is introduced into the text2image diffusion process through residual methods. The CFT mechanism is then used to grade the samples, eliminating the need for multi-stage fine-tuning. Furthermore, Huber loss is introduced to mitigate the impact of outliers and enhance the model's resilience. Our approach has demonstrated a 34.3 percentage point improvement in image quality and a 19.8 percentage point improvement in prompt following compared to the base model. The image advertisement generated based on the new model was showcased on the Baidu Search platform, and the click-through rate exhibited a 29% increase, providing additional evidence of the model's efficacy.

References

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

[2] Tan Yu, Jie Liu, Yi Yang, Yi Li, Hongliang Fei, and Ping Li. 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.

[3] Tan Yu, Zhipeng Jin, Jie Liu, Yi Yang, Hongliang Fei, and Ping Li. Boost ctr prediction for new advertisements via modeling visual content. *2022 IEEE International Conference on Big Data (Big Data)*, pages 2140–2149, 2022.

[4] Tan Yu, Jie Liu, Zhipeng Jin, Yi Yang, Hongliang Fei, and Ping Li. 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.

[5] Ling Yang, Zhilong Zhang, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56:1 – 39, 2022.

[6] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

[7] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023.

[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2022.

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023.

[10] Paul Covington, Jay K. Adams, and Emre Sargin. Deep neural networks for youtube recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

[11] Zhoufutu Wen, Xinyu Zhao, Zhipeng Jin, Yi Yang, Wei Jia, Xiaodong Chen, Shuanglong Li, and Lin Liu. 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.

[12] Xiaoyang Zheng, Fuyu Lv, Zilong Wang, Qingwen Liu, and Xiaoyi Zeng. 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.

[13] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Learning instance-level representation for large-scale multi-modal pretraining in e-commerce. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11060–11069, 2023.

[14] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Hao Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12642–12652, 2021.

[15] Tan Yu, Xuemeng Yang, Yan Jiang, Hongfang Zhang, Weijie Zhao, and Ping Li. Tira in baidu image advertising. *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 2207–2212, 2021.

[16] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 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.

[17] Kang Zhao, Xinyu Zhao, Zhipeng Jin, Yi Yang, Wen Tao, Cong Han, Shuanglong Li, and Lin Liu. 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*, pages 2855–2859, 2024.

[18] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2019.

[19] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020.

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021.

[21] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI Conference on Artificial Intelligence*, 2019.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

[24] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems*, 2021.

[25] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv*, abs/2111.02358, 2021.

[26] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2021.

[27] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *ArXiv*, abs/2205.01917, 2022.

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023.

[30] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390, 2023.

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

[32] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592, 2023.

[33] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *ArXiv*, abs/2211.01335, 2022.

[34] Chunyu Xie, Hengyi Cai, Jianfei Song, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, Xiangyang Ji, and Yafeng Deng. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework. *ArXiv*, abs/2205.03860, 2022.

[35] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. Wukong 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *ArXiv*, abs/2202.06767, 2022.

[36] Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, Guanhui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. In *Neural Information Processing Systems*, 2022.

[37] Zhongzhi Chen, Guangyi Liu, Bo Zhang, Fulong Ye, Qinghong Yang, and Ledell Yu Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *ArXiv*, abs/2211.06679, 2022.

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021.

[39] 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. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022.

[40] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

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

[42] Romain Lopez, Pierre Boyeau, Nir Yosef, Michael I. Jordan, and Jeffrey Regier. Auto-encoding variational bayes. 2020.

[43] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22669–22679, 2022.

[44] William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2022.

[45] 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. Muse: Text-to-image generation via masked generative transformers. *ArXiv*, abs/2301.00704, 2023.

[46] Zhongzhi Chen, Guangyi Liu, Bo Zhang, Fulong Ye, Qinghong Yang, and Ledell Yu Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *ArXiv*, abs/2211.06679, 2022.

[47] Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulic, and Anna Korhonen. Translation-enhanced multilingual text-to-image generation. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

[48] Xiaojun Wu, Di Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. Taiyi-diffusion-xl: Advancing bilingual text-to-image generation with large vision-language model support. *ArXiv*, abs/2401.14688, 2024.

[49] Bin Shan, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil 2.0: Multi-view contrastive learning for image-text pre-training. *ArXiv*, abs/2209.15270, 2022.

[50] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2019.

[51] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 638–647, New York, NY, USA, 2022. Association for Computing Machinery.

[52] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2022.

[53] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Ouyang Xuan, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *ArXiv*, abs/2107.02137, 2021.

[54] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv e-prints*, page arXiv:2010.11934, October 2020.

[55] Xiao Zhang, Miao Li, and Ji Wu. Conditional language learning with context. *arXiv preprint arXiv:2406.01976*, 2024.

[56] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[57] Art B Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443(7):59–72, 2007.

[58] Artem Khrapov, Vadim Popov, Tasnima Sadekova, Assel Yermekova, and Mikhail Kudinov. Improving diffusion models's data-corruption resistance using scheduled pseudo-huber loss. *arXiv preprint arXiv:2403.16728*, 2024.

[59] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21:2347–2360, 2018.

[60] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. *Proceedings of the 25th ACM international conference on Multimedia*, 2017.

[61] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. M6: A chinese multimodal pretrainer. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3251–3261, 2021.

[62] 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. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022.

[63] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv e-prints*, page arXiv:2308.12966, August 2023.