RWKV-CLIP: A Robust Vision-Language Representation Learner

Anonymous ACL submission

Abstract

Contrastive Language-Image Pre-training (CLIP) has significantly improved performance in various vision-language tasks by expanding the dataset with image-text pairs obtained from websites. This paper further explores CLIP from the perspectives of data and model architecture. To address the prevalence of noisy data and enhance the quality of large-scale image-text data crawled from the internet, we introduce a diverse description generation framework that can leverage Large Language Models (LLMs) to synthesize and refine content from web-based texts, synthetic captions, and detection tags. Furthermore, we propose RWKV-CLIP, the first RWKV-driven vision-language representation learning model that combines the effective parallel training of transformers with the efficient inference of RNNs. Comprehensive experiments across various model scales and pre-training datasets demonstrate that RWKV-CLIP is a robust and efficient vision-language representation learner; it achieves state-of-the-art performance in several downstream tasks, including linear probe, zero-shot classification, and zero-shot image-text retrieval. To promote the reproducibility of results, we will release pre-processed data, training code, and pre-trained model weights.

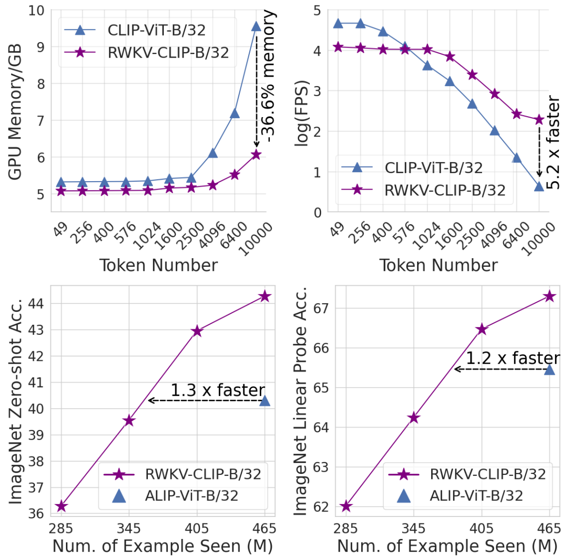

Figure 1: RWKV-CLIP combines the effective parallel training of transformers with the efficient inference of RNNs, achieving better efficiency and accuracy than the baseline methods (e.g., CLIP and ALIP).

1 Introduction

The proliferation of mobile networks and social platforms has dramatically accelerated the large-scale production of image-text pairs. This unprecedented abundance of data has established the foundation for vision-language pre-training. Contrastive Language-Image Pre-training (CLIP) employs two distinct unimodal encoders for images and text, utilizing a contrastive loss, a highly effective mechanism for representation learning. Having been pre-trained on extensive image-text pairs collected from the internet, CLIP demonstrates strong transferability and has been widely applied across various domains (Zhou et al., 2023; Yao et al., 2023).

Many large-scale image-text datasets collected from the internet have been released in recent years. LAION400M (Schuhmann et al., 2021) is created for research purposes, and it contains 400 million image-text pairs curated using the CLIP model. LAION5B (Schuhmann et al., 2022), which consists of 5.85 billion CLIP-filtered image-text pairs, successfully replicates and fine-tunes basic models such as CLIP. However, using the CLIP model to filter web-based image-text pairs still retains a considerable presence of noisy data. To improve data quality, DataComp (Gadre et al., 2024) employs various strategies such as basic filtering, CLIP score filtering, and text&image-based filtering. However, inherent characteristics of internet data, such as abstract text representations and semantic discrepancies between text and images, remain significant obstacles.

In recent years, the Transformer (Vaswani et al., 2017) model has been extensively applied in large-scale representation learning, yielding significant performance improvements across multiple downstream tasks (Acosta et al., 2022; Kirillov et al., 2023; Wang et al., 2023b), including image classification (Dosovitskiy et al., 2020; Wang et al., 2023a), text generation (Brown et al., 2020), and speech recognition (Radford et al., 2023). Despite these achievements, the quadratic computational complexity inherent in the Transformer limits its capacity to effectively process high-resolution images and long sequences, posing a substantial challenge to its broader applicability across varied domains.

In this paper, we design a framework for generating diverse descriptions. Following ALIP (Yang et al., 2023), we first use the OFA (Wang et al., 2022) model to generate synthetic descriptions consistent with image content. However, constrained by the training data, OFA can only partially identify coarse-grained object categories. Therefore, we introduce an open-set image tagging model RAM++ (Huang et al., 2023) to capture more detailed and precise semantic information from images. By leveraging LLMs, we synthesize and refine information from web-based texts, synthetic captions, and detection tags. Additionally, inspired by RWKV (Peng et al., 2024) and Vision-RWKV (Duan et al., 2024), we propose RWKV-CLIP, the first RWKV-driven vision-language representation learning model. As shown in Fig. 1, the proposed RWKV-CLIP combines the effective parallel training of Transformers with the efficient inference of RNNs. Extensive experiments across various model scales and pre-training datasets demonstrate that RWKV-CLIP is a robust and efficient vision-language representation learner. The main contributions of this paper are summarized as follows:

- We introduce a diverse description generation framework, which can leverage LLMs to synthesize and refine information from web-based texts, synthetic captions, and detection tags to produce more accurate and semantically enriched descriptions.
- We propose the RWKV-CLIP, the first RWKV-driven vision-language representation learning model, which combines the parallel training effectiveness of Transformers with the inference efficiency of RNNs.
- We demonstrate the robustness and effectiveness of RWKV-CLIP as a vision-language representation learner through extensive experiments across various model scales and pre-training datasets.

2 Related Work

2.1 Vision-Language Representation Learning

As the milestone in vision-language representation learning, CLIP (Radford et al., 2021) has garnered unparalleled interest due to its remarkable zero-shot recognition capability and outstanding transfer performance. Subsequently, a significant amount of enhancement works based on CLIP have been proposed. SLIP (Mu et al., 2022) combines self-supervised learning with CLIP pre-training to achieve significant performance improvements. De-CLIP (Li et al., 2022b) employs multi-view supervision across modalities and nearest-neighbor supervision from similar pairs to enhance representation learning efficiency. FILIP (Yao et al., 2022) refines contrastive loss to learn fine-grained representations for image patches and sentence words. UniCLIP (Lee et al., 2022) boosts data efficiency by integrating contrastive loss across multiple domains into a single universal space. HiCLIP (Geng et al., 2023) enhances cross-modal alignment by incorporating hierarchy-aware attention into CLIP's visual and language branches. ALIP (Yang et al., 2023) introduces a gating mechanism to reduce the influence of noisy pairs using synthetic data. Different from the above methods, this paper further explores the data and model architecture, proposing a diverse description generation framework and introducing RWKV-CLIP, the first RWKV-driven vision-language representation model.

2.2 Text Augumentation

With the success of LLMs in Natural Language Processing (NLP), there is growing interest in leveraging LLMs to enhance text descriptions in large-scale image-text pairs. LaCLIP (Fan et al., 2023) explores different strategies to generate rewrite examples and uses the in-context learning ability of LLMs to rewrite text within image-text datasets. However, the hallucination issue of LLMs and reliance on limited samples to guide the rewriting process can still introduce significant noise. To address this, CapsFusion (Yu et al., 2024) generates synthetic captions for each image and utilizes ChatGPT to merge raw texts and synthetic captions, creating a dataset with one million instruc-

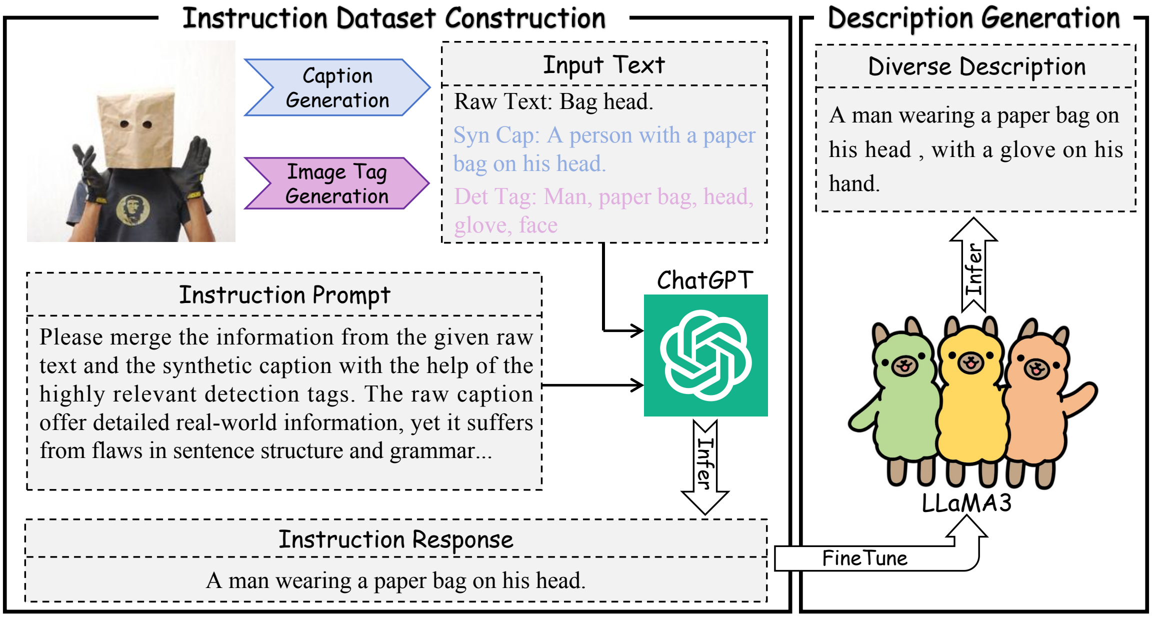

Figure 2: The architecture of our proposed diverse description generation framework.

tions for LLaMA fine-tuning. Despite this, caption generation models such as OFA (Wang et al., 2022) and BLIP (Li et al., 2022a) are limited by their training data and can only identify a restricted set of coarse-grained object categories. In this paper, we introduce the open-set image tagging model RAM++ (Huang et al., 2023) to assign semantic detection tags to each image. Beneficial from detection tags, more semantic information can be introduced from images, which in turn further constrains LLMs and mitigates hallucinations.

2.3 Receptance Weighted Key Value

RWKV (Peng et al., 2023) is first proposed in NLP, it addresses memory bottleneck and quadratic scaling in Transformers through efficient linear scaling while retaining expressive characteristics like parallelized training and robust scalability. Recently, Vision-RWKV (Duan et al., 2024) successfully transferred the RWKV from NLP to vision tasks, outperforming ViT in image classification with faster processing and reduced memory consumption for high-resolution inputs. PointRWKV (He et al., 2024) demonstrates leading performance across various downstream tasks, surpassing Transformer- and Mamba-based counterparts in efficiency and computational complexity. Furthermore, Diffusion-RWKV (Fei et al., 2024) adapts RWKV for diffusion models in image generation tasks, achieving competitive or superior performance compared to existing CNN or Transformer-based diffusion models. However, these methods have only validated RWKV in specific downstream tasks, and the potential of RWKVs to replace ViTs in vision-language representation learning remains unverified.

3 Method

In this section, we first introduce a diverse description generation framework that leverages the capabilities of large language models to integrate information from web-based texts, synthetic captions, and detection tags. Subsequently, we provide a detailed exposition of RWKV-CLIP.

3.1 Diverse Description Generation

The architecture of our proposed diverse description generation framework is illustrated in Fig. 2. To mitigate the effects of mismatched image-text pairs, following ALIP (Yang et al., 2023), we first adopt the OFA$_{base}$ model to generate a synthetic caption for each image. The synthetic captions exhibit a high degree of semantic alignment with the image, facilitating alignment across different modal feature spaces. However, constrained by the training data, OFA$_{base}$ can recognize a limited number of object categories and tends to produce captions with a simplistic sentence structure. To capture finer-grained semantic information within images, we incorporate the open-set image tagging models RAM++ (Huang et al., 2023) to extract object detection tags for each image.

Following CapsFusion (Yu et al., 2024) to assess our approach's viability, we initially leverage ChatGPT to combine information from raw texts,

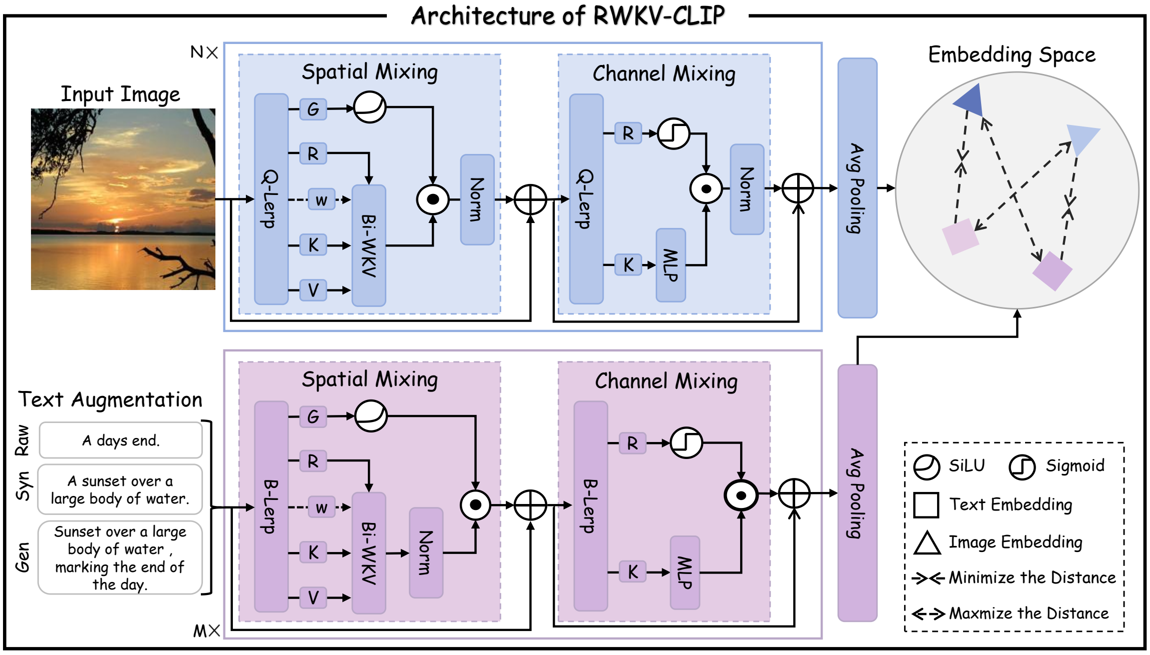

Figure 3: The architecture of RWKV-CLIP, which consists of $M\times$ and $N\times$ RKWV-driven blocks followed by an average pooling layer.

synthetic captions, and detection tags. However, the time and computational effort involved is prohibitive. Therefore, we constructed an instruction dataset based on ChatGPT interactions and fine-tuned the open-source LLaMA3 with this dataset. After that, we leverage the fine-tuned LLaMA3 model (Touvron et al., 2023) for large-scale inference. Specifically, we select 70K image-text pairs from YFCC15M with more than 10 detection tags. Then, we input the raw texts, synthetic captions, and detection tags of these data into ChatGPT to get instruction responses. The details of the instruction prompt are provided in the supplementary material.

After obtaining the instruction dataset, we utilize the LLaMA Factory (Zheng et al., 2024) to finetune the LLaMA3-8B and leverage vLLM (Kwon et al., 2023) to accelerate large-scale inference.

3.2 RWKV-CLIP

In this section, we propose RWKV-CLIP, a robust and efficient RWKV-driven vision-language representation learner. Inspired by CLIP (Radford et al., 2021) and Vision-RWKV (Duan et al., 2024), RWKV-CLIP adopts a dual-tower architecture with a block-stacked encoder design like the Transformer (Vaswani et al., 2017), where each block consists of a spatial mixing and a channel mixing module. The overview architecture of our proposed RWKV-CLIP is shown in Fig. 3.

Input Augmentation. Based on our proposed diverse description generation framework, we can obtain three types of text: raw text T_r, synthetic caption T_s, and generated description T_g. To improve the robustness of the model, we randomly select a text from $[T_r, T_s, T_g]$ as the augmentation for text inputs:

$$\text{aug}(T) = \text{Sample}([T_r, T_s, T_g]). \tag{1}$$

Meanwhile, the input image $I \in \mathbb{R}^{H\times W\times 3}$ is transformed into HW/p^2 patches, where p is the patch size.

Spatial Mixing. The input text $\text{aug}(T)$ and image I are passed through the spatial mixing module, which acts as an attention mechanism and performs global attention computation of linear complexity. Specifically, the input data is shifted and entered into four parallel linear layers to obtain multi-head vectors $G_x^s, R_x^s, K_x^s, V_x^s$:

$$\psi_x^s = \text{Lerp}_\psi(x) \cdot w_\psi^s, \quad \psi \in \{G, R, K, V\}, \tag{2}$$

where Lerp is the linear interpolation (Peng et al., 2024). In this paper, we adopt Q-Lerp and B-Lerp for image and text encoders respectively. The Q-Lerp can be formulated as:

$$\begin{aligned} &\text{Q-Lerp}_\Psi(I) = I + (1-\eta_\Psi)\cdot I^\star, \\ &I^\star = \text{Concat}(I_1, I_2, I_3, I_4). \end{aligned} \tag{3}$$

The B-Lerp can be presented as:

$$\begin{aligned} &\text{B-Lerp}_\Psi(T) = T + (1-\eta_\Psi)\cdot T^\star, \\ &T^\star = \text{Concat}(T_1, T_2), \end{aligned} \tag{4}$$

where $\Psi \in \{G, R, K, V, w\}$, η_Ψ denotes learnable vectors, $I^\star$ is the quad-directional shift vector in the image, *i.e.*, $I_1 = x[h-1, w, 0 : C/4], I_2 = x[h+1, w, C/4 : C/2], I_3 = x[h, w-1, C/2 : 3C/4], I_4 = x[h, w+1, 3C/4 : C]$, $T^\star$ is the bi-directional shift vector in the text *i.e.*, $T_1 = [w-1, 0 : C/2], T_2 = [w+1, C/2 : C]$, where h, w, C present the number of height, width, and channel. These shift functions enhance feature interaction at the channel level, enabling a focus on neighboring tokens. Specifically, the bi-directional shift ensures forward and backward interaction of text tokens without increasing additional FLOPs. To avoid a fixed learned vector, a new time-varying decay w_x is calculated as follows:

$$\begin{aligned} \phi(x) &= \lambda + \tanh(x \cdot M_i) \cdot M_j, \\ \hat{w}_x^s &= x + (1 - \phi(\text{Lerp}_w(x))) \cdot x^\star, \\ \tilde{w}_x^s &= \phi(\hat{w}_x^s), w_x^s = \exp\left(-\exp(\tilde{w}_x^s)\right), \end{aligned} \tag{5}$$

where $x \in \{I, T\}$, λ is a learnable vector, M_i, M_j are learnable weight matrices. The function ϕ is used to obtain learned vectors by inexpensively augmenting inputs with additional offsets. $\hat{w}_x^s$ and $\tilde{w}_x^s$ are middle values of w_x^s during the calculation process. This process allows each channel of w_x to vary based on a mix of the current and prior tokens $x^\star$.

Subsequently, $w_x^s, R_x^s, K_x^s, V_x^s$ are used to compute the global attention result wkv_t via a linear complexity bidirectional attention mechanism. This result is then multiplied by $\sigma(G_x^s)$, functioning as a gate mechanism to control the output O_x^s:

$$\begin{aligned} wkv_t &= \text{Bi-WKV}_t(w_x^s, R_x^s, K_x^s, V_x^s), \\ O_x^s &= \text{Concat}\left(\sigma(G_x^s) \odot \text{LN}(wkv_t)\right) \cdot w_o^s, \end{aligned} \tag{6}$$

where $\sigma(\cdot)$ denotes the SiLU function (Elfwing et al., 2018), and $\odot$ means element-wise multiplication, LN is the layer norm and the Bi-WKV (Duan et al., 2024; Peng et al., 2024) can be formulated as:

$$\begin{aligned} \text{Bi-WKV}_t = R_{s,t} \cdot (&\text{diag}(u) \cdot K_{s,t}^\top \cdot V_{s,t} \\ &+ \sum_{i=0}^{t-1} \text{diag}(\epsilon_{i,j}) \cdot K_{s,i}^\top \cdot V_{s,i} \\ &+ \sum_{i=t+1}^{T-1} \text{diag}(\epsilon_{i,j}) \cdot K_{s,i}^\top \cdot V_{s,i}), \end{aligned} \tag{7}$$

where u is a per-channel learned boost and $\epsilon_{i,j} = \odot_{j=1}^{i-1} w_j$ is a dynamic decay.

Channel Mixing. The spatial mixing module is followed by the channel-mixing module. Similarly, the R_x^c, K_x^c are obtained by Lerp:

$$\psi_x^c = \text{Lerp}_\psi(x) \cdot w_\psi^c, \quad \psi \in \{R, K\}. \tag{8}$$

After that, a linear projection and a gate mechanism are performed respectively and the final output O_x^c is formulated as:

$$O_x^c = (\sigma(R_x^c) \odot \rho(K_x^c)) \cdot w_o^c, \tag{9}$$

where ρ is the squaredReLU (Agarap, 2018). After passing through the stack RWKV-based image and text encoders E_I and E_T, we can get the image embeddings $\hat{I} = E_I(I)$ and text embeddings $\hat{T} = E_T(aug(T))$, the loss function L is defined as:

$$L = -\sum_{i=1}^{N} \left[\log \frac{e^{\hat{I}_i^\top \hat{T}_i/\tau}}{\sum_j e^{\hat{I}_i^\top \hat{T}_j/\tau}} + \log \frac{e^{\hat{I}_i^\top \hat{T}_i/\tau}}{\sum_j e^{\hat{I}_j^\top \hat{T}_i/\tau}} \right]. \tag{10}$$

4 Experiments

4.1 Experimental Settings

Pre-training Datasets. We train our model on the YFCC15M dataset, which is a subset of YFCC100M (Thomee et al., 2016) filtered by DeCLIP (Li et al., 2022b). To further verify the effectiveness and generalizability of RWKV-CLIP, following ALIP (Yang et al., 2023), we randomly select subsets of 10M and 30M from the LAION400M (Schuhmann et al., 2021). We then conduct a series of experiments with different model scales and pre-training datasets.

Implementation Details. Consistent with ALIP (Yang et al., 2023), we employ OFA_{base} to generate synthetic captions. The instruction dataset is constructed using ChatGPT-35-turbo, and we fine-tune LLaMA3-8B to enhance the generation of diverse descriptions. We employ AdamW (Loshchilov and Hutter, 2019) as the optimizer, initialized with a learning rate of $1\text{e}{-3}$ and a weight decay of 0.2. The parameters $\beta 1$ and β_2 are set to 0.9 and 0.98, respectively. The input image size is 224×224, and the input text sequence length is truncated or padded to 77. The temperature parameter τ is initialized to 0.07. We train RWKV-CLIP for 32 epochs with a batch size of 4096 on 8 NVIDIA A100 (80G) GPUs. We meticulously regulate the parameters and FLOPs of RWKV-CLIP to ensure the fairness of the experimental comparison. Please refer to the supplementary material for more detailed parameters, FLOPs, and settings of RWKV-CLIP.

4.2 Experimental Results

Linear Probe. Building upon previous works (Yang et al., 2023; Li et al., 2022b; Geng et al., 2023), we use RWKV-CLIP as

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	Average
CLIP-ViT-B/32(Radford et al., 2021)	YFCC15M	86.5	64.7	69.2	64.6	90.6	66.0	24.9	61.3	79.1	23.1	63.0
DeCLIP-ViT-B/32 (Li et al., 2022b)	YFCC15M	89.2	69.0	75.4	72.2	94.4	71.6	31.0	68.8	87.9	27.6	68.7
HiCLIP-ViT-B/32 (Geng et al., 2023)	YFCC15M	89.5	71.1	73.5	70.6	91.9	68.8	30.8	63.9	84.8	27.4	67.2
ALIP-ViT-B/32 (Yang et al., 2023)	YFCC15M	94.3	77.8	75.8	76.0	**95.1**	**73.3**	33.6	71.7	88.5	36.1	72.2
RWKV-CLIP-B/32	YFCC15M	**95.3**	**81.8**	**76.4**	**77.1**	92.4	73.1	**37.7**	**73.2**	**90.6**	**43.5**	**74.1**

Table 1: Linear probe performance on 10 downstream datasets. RWKV-CLIP achieves an average performance improvement of 1.9%∼11.1%.

	Text retrieval						Image retrieval					
	Flickr30k			MSCOCO			Flickr30k			MSCOCO		
Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-ViT-B/32(Radford et al., 2021)	34.9	63.9	75.9	20.8	43.9	55.7	23.4	47.2	58.9	13.0	31.7	42.7
SLIP-ViT-B/32 (Mu et al., 2022)	47.8	76.5	85.9	27.7	52.6	63.9	32.3	58.7	68.8	18.2	39.2	51.0
DeCLIP-ViT-B/32 (Li et al., 2022b)	51.4	80.2	88.9	28.3	53.2	64.5	34.3	60.3	70.7	18.4	39.6	51.4
UniCLIP-ViT-B/32 (Lee et al., 2022)	52.3	81.6	89.0	32.0	57.7	69.2	34.8	62.0	72.0	20.2	43.2	54.4
HiCLIP-ViT-B/32 (Geng et al., 2023)	-	-	-	34.2	60.3	70.9	-	-	-	20.6	43.8	55.3
ALIP-ViT-B/32 (Yang et al., 2023)	70.5	91.9	95.7	46.8	72.4	81.8	48.9	75.1	82.9	29.3	54.4	65.4
RWKV-CLIP-B/32	**76.0**	**94.7**	**97.6**	**50.3**	**76.2**	**85.2**	**57.6**	**82.3**	**88.7**	**34.0**	**60.9**	**71.7**

Table 2: Zero-shot image-text retrieval performance on the test splits of Flickr30k and MSCOCO. RWKV-CLIP achieves a significant improvement on all metrics.

a feature extractor and train only a logistic regression classifier. Tab. 1 details the linear probe performance across 10 downstream datasets, as referenced in ALIP (Yang et al., 2023). RWKV-CLIP achieves a significant performance improvement ranging from 1.9% $\sim$ 11.1% over the baseline models, outperforming ALIP in 8 of the 10 datasets. The observed performance improvements are primarily due to two main factors: (1) Our proposed description generation framework effectively synthesizes and refines information from web-based texts, synthetic captions, and detection tags, producing more accurate and semantically enriched descriptions. (2) RWKV-CLIP exhibits superior representation learning capabilities compared to Transformer-based models.

Zero-shot Image-text Retrieval. In Tab. 2, we compare our method with state-of-the-art approaches in zero-shot image-text retrieval on Flickr30k and MSCOCO. RWKV-CLIP achieves new state-of-the-art results on all evaluation metrics. Specifically, RWKV-CLIP achieves 76.0%/57.6% I2T/T2I retrieval Recall@1 on Flickr30K, surpassing ALIP by 5.5%/8.7%. Similarly, significant improvements of 3.5%/4.7% in I2T/T2I retrieval Recall@1 are observed for RWKV-CLIP on MSCOCO. This exceptional image-text retrieval capability indicates that the representations learned by RWKV-CLIP are robust and exhibit enhanced cross-modal alignment.

Zero-shot Classification. We present the zero-shot classification performance across 11 datasets. To ensure fair comparisons, we use the same prompt templates and class names as established in ALIP (Yang et al., 2023) and SLIP (Mu et al., 2022). As shown in Tab. 3, RWKV-CLIP achieves an average performance improvement of 2.6% $\sim$ 14.4% over baseline models. Notably, our model outperforms ALIP in 10 out of the 11 datasets, with significant enhancements on instance discrimination datasets such as Food101, and ImageNet. This improvement is mainly due to the diverse descriptions generated by our framework, providing more fine-grained semantic information.

Zero-Shot Robustness Evaluation. In Tab. 4, we present a robustness evaluation comparing ALIP and RWKV-CLIP. Our results show that RWKV-CLIP consistently outperforms ALIP in terms of robustness across all datasets with an average improvement of 2.0%. These experimental results establish the RWKV-driven model as a robust representation learner.

4.3 Ablation Study

Effectiveness of Model and Data Scaling. To evaluate the effectiveness of RWKV-CLIP on model and data scaling, we conduct experiments on randomly selected subsets of 10M and 30M from LAION400M. For a more comprehensive compari-

Method	Pre-train dataset	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	ImageNet	Average
CLIP-ViT-B/32(Radford et al., 2021)	YFCC15M	63.7	33.2	34.6	20.1	50.1	35.7	2.6	15.5	59.9	1.2	32.8	31.8
SLIP-ViT-B/32 (Mu et al., 2022)	YFCC15M	50.7	25.5	33.3	23.5	49.0	34.7	2.8	14.4	59.9	1.7	34.3	30.0
FILIP-ViT-B/32 (Yao et al., 2022)	YFCC15M	65.5	33.5	43.1	24.1	52.7	50.7	3.3	24.3	68.8	3.2	39.5	37.2
DeCLIP-ViT-B/32 (Li et al., 2022b)	YFCC15M	66.7	38.7	52.5	33.8	60.8	50.3	3.8	**27.7**	74.7	2.1	43.2	41.3
HiCLIP-ViT-B/32 (Geng et al., 2023)	YFCC15M	74.1	46.0	**51.2**	**37.8**	**60.9**	50.6	**4.5**	23.1	67.4	3.6	40.5	41.8
ALIP-ViT-B/32 (Yang et al., 2023)	YFCC15M	**83.8**	51.9	45.4	30.7	54.8	47.8	3.4	23.2	74.1	2.7	40.3	41.7
RWKV-CLIP-B/32	YFCC15M	79.8	**55.1**	50.6	37.6	57.1	**54.0**	4.1	24.6	**77.1**	**4.0**	**44.3**	**44.4**

Table 3: Zero-shot classification performance on 11 downstream datasets. RWKV-CLIP achieves an average performance improvement of 2.6%∼12.6%.

Method	IN-V2	IN-A	IN-R	IN-Sketch	Average
ALIP-ViT-B/32	34.1	16.1	35.2	12.1	24.4
RWKV-CLIP-B/32	**37.5**	**16.7**	**37.0**	**14.5**	**26.4**

Table 4: Zero-shot robustness comparison of ALIP and RWKV-CLIP pretrained on YFCC15M.

son, we report the linear probe performance on 26 downstream datasets. As shown in Fig. 5, RWKV-CLIP significantly improves performance across different model scales and pre-training datasets. These results demonstrate the robustness and extensibility of RWKV-CLIP. Detailed experimental results can be found in the supplementary material.

Comparision Analysis with CapsFusion. To further demonstrate the performance differences between our proposed diverse description generation framework and CapsFusion, we used CapsFusion-LLaMA to rewrite the YFCC15M dataset based on raw texts and synthetic captions. We then trained RWKV-CLIP using texts generated by our framework and CapsFusion. As shown in Tab. 5, our framework achieves a 0.9% and 2.1% improvement in the average linear probe and zero-shot classification performance, respectively. This improvement is primarily due to the detection tags introducing more semantic information from images, which further constrains LLMs and reduces hallucinations (as shown in Fig. 4).

Method	Text Generation Model	Linear probe Avg	Zero-shot Avg
RWKV-CLIP-B/32	CapsFusion	72.6	33.1
RWKV-CLIP-B/32	Ours	**73.5**	**35.2**

Table 5: Performance comparison using text generated by our proposed diverse description generation framework vs. CapsFusion.

Ablation on Different Types of Text. We conduct ablation experiments on different categories of text, the average linear probe results on 10 datasets and the average zero-shot classification accuracy on 11 datasets are shown in Tab. 6. Synthetic captions and generated diverse descriptions yielded superior linear probe performance compared to raw texts. This improvement is attributed to the high incidence of mismatched image-text pairs in raw texts, which can adversely affect representation learning. As shown in Fig. 6, our analysis of cosine similarity (computed by CLIP-L14) and token counts across different text types reveals that synthetic captions and generated diverse descriptions have

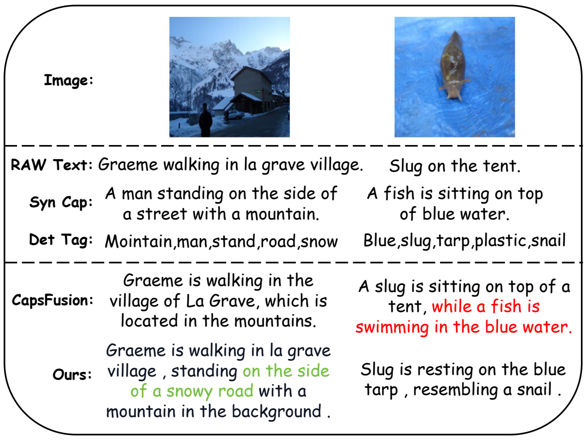

Figure 4: Comparison of our proposed diverse description generation framework vs. CapsFusion. Hallucinations are highlighted in red, and additional semantic information is highlighted in green.

T_r	T_s	T_g	Dataset	Linear probe Avg	Zero-shot Avg
✓	✗	✗	YFCC15M	71.3	38.7
✗	✓	✗	YFCC15M	72.4	23.1
✗	✗	✓	YFCC15M	73.5	35.2
✓	✓	✗	YFCC15M	73.0	43.0
✓	✗	✓	YFCC15M	73.8	43.4
✓	✓	✓	YFCC15M	**74.1**	**44.4**

Table 6: Ablation experiment results using different types of text. T_r: raw text. T_s: synthetic caption. T_g: generated diverse description using our framework.

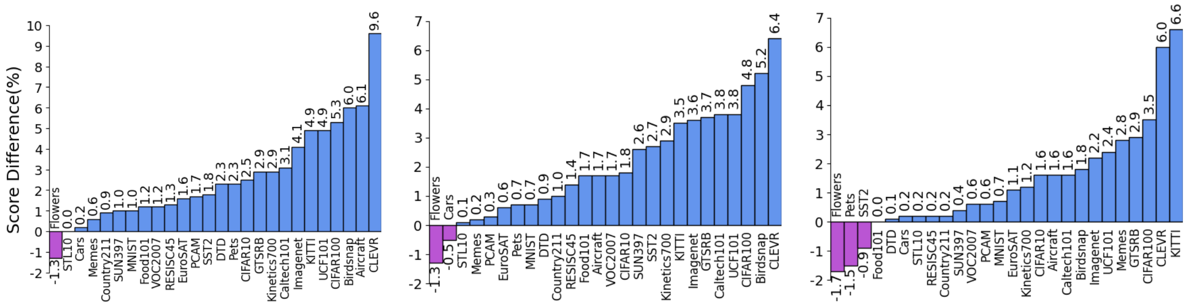

Figure 5: Linear probe performance comparison between RWKV-CLIP and ALIP on 26 downstream datasets. The comparisons include RWKV-CLIP-B/32 vs. ALIP-ViT-B/32 on LAION10M, RWKV-CLIP-B/16 vs. ALIP-ViT-B/16 on LAION10M, and RWKV-CLIP-B/32 vs. ALIP-ViT-B/32 on LAION30M, presented from left to right.

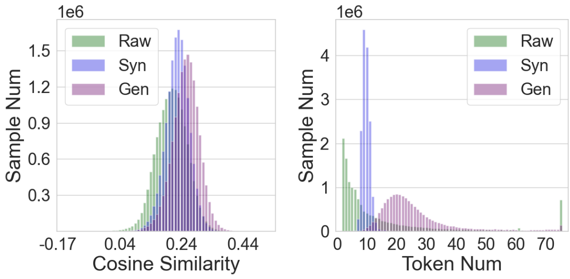

Figure 6: Statistical analysis of raw texts, synthetic captions, and generated diverse descriptions on the YFCC15M.

higher average similarity and token counts than raw texts. Furthermore, despite these advantages, raw texts achieve superior zero-shot classification results, mainly due to the constraints imposed by the prompt template.

Ablation on Model Architecture. In Tab. 7, based on text augmentation, we perform an ablation study combining RWKV and Transformer architectures. Compared with Transformer_I and Transformer_T, the integration of RWKV_I and Transformer_T achieves a 2.7% improvement on the linear probe but the zero-shot classification performance declines by 10.8%. This reduction is primarily due to the poor compatibility between the RWKV and Transformer architectures. Conversely, the combination of RWKV_I and RWKV_T yields improvements of 3.2% and 2.7% in linear probe and zero-shot classification, respectively, indicating that RWKV outperforms Transformer in vision-language representation learning.

Image		Text		Linear Probe	Zero-shot
RWKV_I	Transformer_I	RWKV_T	Transformer_T	Avg	Avg
✗	✓	✗	✓	70.9	41.7
✓	✗	✗	✓	73.6	30.9
✗	✓	✓	✗	71.0	41.1
✓	✗	✓	✗	**74.1**	**44.4**

Table 7: Ablation on model architecture.

Analysis of Feature Embedding. To understand what makes RWKV-CLIP effective, we randomly select 250 image-text pairs from YFCC15M and visualize the modality gaps of ALIP and RWKV-CLIP. Specifically, each image and its corresponding text are encoded into embedding space and reduced to two dimensions using UMAP (McInnes et al., 2018). As shown in Fig. 7, we found that the representations learned by RWKV-CLIP exhibit clearer discriminability within the same modality. Additionally, compared to ALIP, RWKV-CLIP demonstrates closer distances in the image-text modality space, indicating superior cross-modal alignment performance.

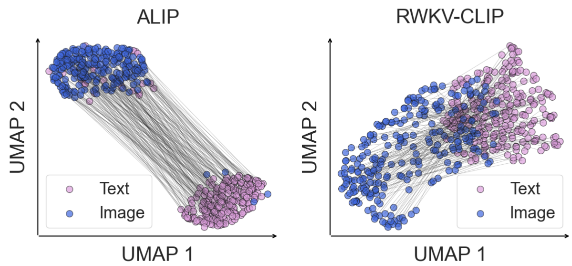

Figure 7: Visualization of modality gaps.

5 Conclusion

In this paper, we further explore CLIP from the perspectives of data and model architecture. We introduce a diverse description generation framework that can leverage Large Language Models (LLMs) to combine and refine information from web-based image-text pairs, synthetic captions, and detection tags. Besides, we propose RWKV-CLIP, the first RWKV-driven vision-language representation learning model that combines the effective parallel training of transformers with the efficient inference of RNNs. Our method demonstrates superior performance across various model scales and pretraining datasets on different downstream tasks. We hope that our work provides insights into vision-language representation learning models.

Limitations

Our proposed framework for diverse description generation leverages the existing caption generation model and detection tags model, both of which can directly influence the quality of the final generated descriptions. Furthermore, due to limitations in computational resources, this study only executes experiments at tens of millions of scales of image-text pairs. Conducting experiments at a billion-scale necessitates substantial computational resources.

References

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv:1803.08375*.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *ECCV*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv:1907.06987*.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *CVPR*.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. 2024. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv:2403.02308*.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11.

Mark Everingham. 2007. The pascal visual object classes challenge,(voc2007) results. *http://pascallin. ecs. soton. ac. uk/challenges/VOC/voc2007/index. html*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. In *NeurIPS*.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. 2024. Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. *arXiv:2404.04478*.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *NeurIPS*.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.

Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. 2023. HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention. In *ICLR*.

Qingdong He, Jiangning Zhang, Jinlong Peng, Haoyang He, Yabiao Wang, and Chengjie Wang. 2024. Pointrwkv: Efficient rwkv-like model for hierarchical point cloud learning. *arXiv:2405.15214*.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2023. Open-set image tagging with multi-grained text supervision. *arXiv:2310.15200*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *CVPR*, pages 4015–4026.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *ICCVW*.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *ACM SIGOPS*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. 2022. Uniclip: Unified framework for contrastive language-image pre-training. *NeurIPS*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv:1306.5151*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *ECCV*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *ICCV*.

Bo Peng, Eric Alcaide, Quentin Gregory Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Nguyen Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan S. Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the transformer era. In *EMNLP*.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. 2024. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv:2404.05892*.

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 *ICML*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

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. In *NeurIPS*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian

Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *MICCAI*.

Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. 2023a. Learning bottleneck concepts in image classification. In *CVPR*, pages 10962–10971.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. 2023b. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *ICCV*.

Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*.

Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. 2023. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. *CVPR*.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained interactive language-image pre-training. In *ICLR*.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. 2024. Capsfusion: Rethinking image-text data at scale. In *CVPR*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv:2403.13372*.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *CVPR*.

A Detail Experimental Settings

A.1 Model Architectures

We meticulously regulate the parameters and FLOPs of RWKV-CLIP to ensure the fairness of the experimental comparison. The detailed parameters and FLOPs of RWKV-CLIP-B/32 and RWKV-CLIP-B/16 are shown in Tab. 8. The detailed settings of RWKV-CLIP-B/32 and RWKV-CLIP-B/16 are shown in Tab. 10.

Method	Image		Text		Total	
	Params(M)	FLOPs(G)	Params(M)	FLOPs(G)	Params(M)	FLOPs(G)
CLIP-ViT-B/32	87.85	8.73	63.44	5.82	151.29	14.55
RWKV-CLIP-B/32	84.21	7.91	65.35	4.93	149.56	12.84
CLIP-ViT-B/16	86.19	33.72	63.44	5.82	149.63	39.54
RWKV-CLIP-B/16	82.83	31.05	65.35	4.93	148.18	35.98

Table 8: Parameters and FLOPs comparison between CLIP and RWKV-CLIP.

A.2 Detail Instruction Prompt

The prompt used to input ChatGPT is present in the following:

"Please merge the information from the given raw text and the synthetic caption with the help of the highly relevant detection tags. The raw caption offers detailed real-world information, yet it suffers from flaws in sentence structure and grammar. The synthetic caption exhibits impeccable sentence structure but often lacks in-depth real-world details and may contain false information. The highly relevant detection tags are provided to enrich the semantic information of the raw caption, while some are redundant and noisy. You are a great information integration and summary expert, you are also good at enriching semantic information. Ensure a well-structured sentence while retaining the detailed real-world information provided in the raw caption. Avoid simply concatenating the sentences and avoid adding external information to describe. Correctness and simplify sentences finally. Raw caption:<raw caption>, synthetic caption:<synthetic caption>, and highly relevant detection tags:<detection tags>".

A.3 Experimental Settings

We present the settings used in the training RWKV-CLIP in Tab. 9.

A.4 Prompts for Zero-shot Classification

In this work, we evaluate the zero-shot performance of RWKV-CLIP on 11 downstream datasets. All the prompts for the 11 downstream datasets are presented in Tab. 13.

Hyperparameter	Value
Initial temperature	0.07
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	10^{-6}
Weight decay	0.2
Batch size	4096
Learning rate	0.001
Learning rate scheduler	OneCycleLR
Pct start	0.1
Training epochs	32
GPU	8×A100

Table 9: Hyperparameters used for RWKV-CLIP pre-training.

B Detail Linear Probe on LAION

B.1 Downstream Datasets

To comprehensively demonstrate the performance of RWKV-CLIP, we compared the linear probe results of RWKV-CLIP and ALIP across 26 datasets. These datasets include Food101 (Bossard et al., 2014), CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), Birdsnap (Berg et al., 2014), SUN397 (Xiao et al., 2010), Stanford Cars (Krause et al., 2013), FGVC Aircraft (Maji et al., 2013), VOC2007 (Everingham, 2007), DTD (Cimpoi et al., 2014), Pets (Parkhi et al., 2012), Caltech101 (Fei-Fei et al., 2004), Flowers102 (Nilsback and Zisserman, 2008), MNIST (LeCun et al., 1998), SLT10 (Coates et al., 2011), EuroSAT (Helber et al., 2019), RESISC45 (Cheng et al., 2017), GTSRB (Stallkamp et al., 2012), KITTI (Geiger et al., 2012), Country211 (Radford et al., 2021), PCAM (Veeling et al., 2018), UCF101 (Soomro et al., 2012), Kinetics700 (Carreira et al., 2019), CLEVR (Johnson et al., 2017), Hateful Memes (Kiela et al., 2020), SST2 (Radford et al., 2021), ImageNet (Deng et al., 2009). Details on each dataset and the corresponding evaluation metrics are provided in Tab. 12.

B.2 Detail Linear Probe Results

Following ALIP, we conduct experiments on randomly selected subsets of 10M and 30M from the LAION400M dataset. For a comprehensive comparison, we report the linear probe performance on 26 downstream datasets. The complete experimental results are shown in Tab.11. RWKV-CLIP-B/32 outperforms ALIP-ViT-B/32 2.6% and 1.4% when training on LAION10M and LAION30M, respectively. Additionally, RWKV-CLIP-B/16 also surpasses ALIP-ViT-B/16 by 2.1% on average across the 26 datasets. These experimental results indicate

Model	Embedding dimension	Input resolution	Image Encoder layers	hidden rate	heads	Init	Text Encoder layers	hidden rate	heads	Init
RWKV-CLIP-B/32	640	224	12	5	8	✓	6	3.5	10	✓
RWKV-CLIP-B/16	640	224	12	5	8	✓	6	3.5	10	✓

Table 10: The detail architecture parameters for our proposed RWKV-CLIP.

Method	Pre-train data	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	Memes	SST2	ImageNet	Average
ALIP-ViT-B/32	LAION10M	71.5	92.2	76.1	36.3	67.3	70.1	41.8	85.3	71.3	74.3	86.9	**90.7**	98.0	94.6	95.4	84.3	84.1	70.0	12.9	83.4	75.9	46.4	51.0	54.8	56.5	59.6	70.4
RWKV-CLIP-B/32	LAION10M	**72.7**	**94.7**	**81.4**	**42.3**	**68.3**	**70.3**	**47.9**	**86.5**	**73.6**	**76.6**	**90.0**	89.4	**99.0**	**94.6**	**97.0**	**85.6**	**87.0**	**74.9**	**13.8**	**85.1**	**80.8**	**49.3**	**60.6**	**55.4**	**58.3**	**63.7**	**73.0**
ALIP-ViT-B/16	LAION10M	77.2	93.3	77.0	45.1	69.4	**77.3**	48.6	87.7	74.5	79.0	88.1	**93.0**	98.3	96.3	96.3	86.4	83.7	72.2	14.2	85.2	80.1	50.1	55.4	55.7	57.3	64.8	73.3
RWKV-CLIP-B/16	LAION10M	**78.9**	**95.1**	**81.8**	**50.3**	**72.0**	76.8	**50.3**	**89.4**	**75.4**	**79.7**	**91.9**	91.7	**99.0**	**96.4**	**96.9**	**87.8**	**87.4**	**75.7**	**15.2**	**85.5**	**83.9**	**53.0**	**61.8**	**55.9**	**60.0**	**68.4**	**75.4**
ALIP-ViT-B/32	LAION30M	76.6	94.0	79.3	44.2	70.6	77.7	48.4	87.6	74.4	**80.4**	90.0	**93.8**	98.3	96.3	96.0	86.7	84.7	72.3	15.0	85.0	81.0	50.6	55.6	56.1	**59.8**	65.0	73.8
RWKV-CLIP-B/32	LAION30M	**76.6**	**95.6**	**82.8**	**46.0**	**71.0**	**77.9**	**50.0**	**88.2**	**74.5**	78.9	**91.6**	92.1	**99.0**	**96.5**	**97.1**	**86.9**	**87.6**	**78.9**	**15.2**	**85.6**	**83.4**	**51.8**	**61.6**	**58.9**	58.9	**67.2**	**75.2**

Table 11: Top-1 accuracy(%) of linear probe on 26 image classification datasets.

Dataset	Classes	Train size	Test size	Evaluation metric
Food101	102	75,750	25,250	accuracy
CIFAR10	10	50,000	10,000	accuracy
CIFAR100	100	50,000	10,000	accuracy
Birdsnap	500	42,138	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Cars	196	8,144	8,041	accuracy
Aircraft	100	6,667	3,333	mean per class
VOC2007	20	5011	4952	11-point mAP
DTD	47	3,760	1,880	accuracy
Pets	37	3,680	3,669	mean per class
Caltech101	101	3,000	5,677	mean-per-class
Flowers	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
STL10	10	5,000	8,000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6770	711	accuracy
Country211	211	42,200	21,100	accuracy
PCAM	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	530,779	33,944	mean(top1,top5)
CLEVR	8	2,000	500	accuracy
Memes	2	8,500	500	ROC AUC
SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

Table 12: List of linear probe datasets with the data distribution and evaluation metrics.

that RWKV-CLIP demonstrates both robustness and extensibility.

C More Visualize and Analysis

C.1 Class Activation Map

As shown in Fig. 8, we visualize the class activation maps of ALIP and RWKV-CLIP on different classes from ImageNet. RWKV-CLIP performs superior in aligning the image patches and textual tokens. For example, RWKV-CLIP captures corresponding text semantic entities in images more accurately.

C.2 Cross Modal Alignment Analysis

To evaluate the performance of the cross-modal alignment of RWKV-CLIP, we random select 50

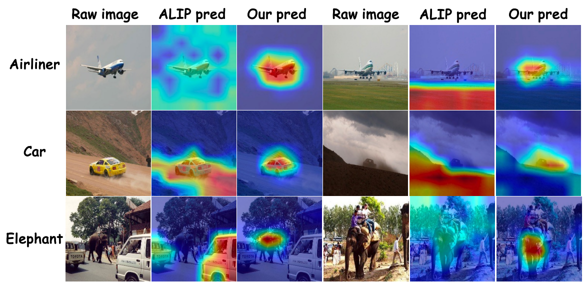

Figure 8: Class activation maps for ALIP and RWKV-CLIP on different classes from ImageNet.

samples from YFCC15M and visualize the cross-modal cosine similarity matrix in Fig. 9. We observe that the diagonal of the RWKV-CLIP matrix is significantly clearer compared to ALIP, indicating that the representations learned by RWKV-CLIP exhibit greater distinctiveness and improved cross-modal alignment capability.

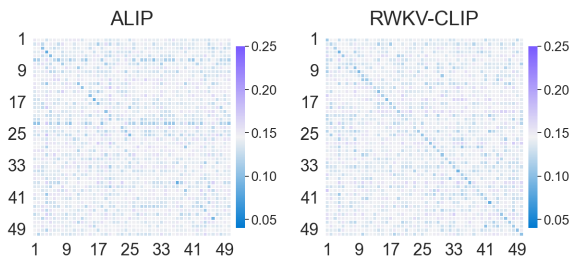

Figure 9: Visualization of modality gaps.

C.3 Case Study

In Fig. 10, we visualize additional generated text using CapsFusion and our proposed framework. The introduction of detection tags enhances semantic information from images, thereby constraining LLMs and significantly reducing hallucinations.

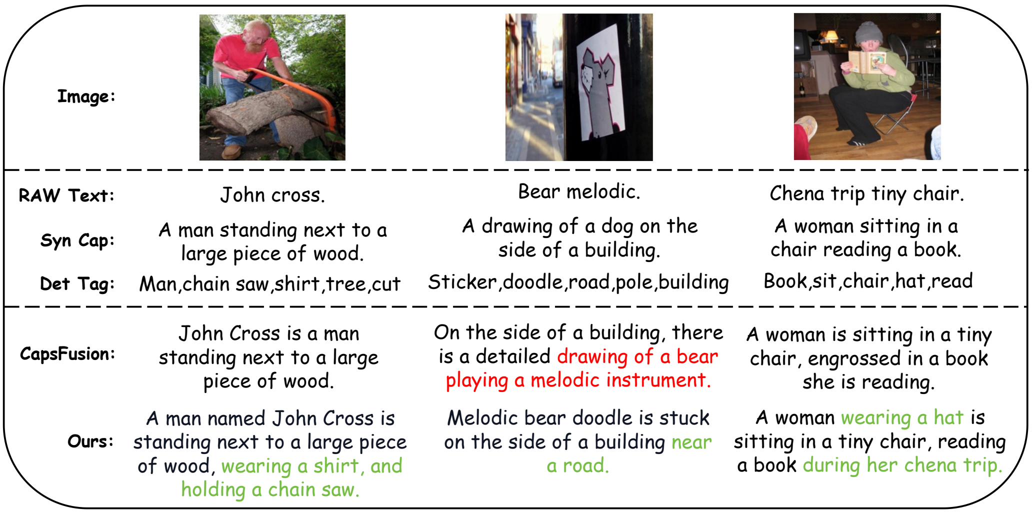

Figure 10: Comparison of generated text using our proposed diverse description generation framework vs. CapsFusion. Hallucinations are highlighted in red, and additional semantic information is highlighted in green.

CIFAR 10 & CIFAR 100			
a photo of a {label}.	a blurry photo of a {label}.	a black and white photo of a {label}.	a low contrast photo of a {label}.
a high contrast photo of a {label}.	a bad photo of a {label}.	a good photo of a {label}.	a photo of a small {label}.
a photo of a big {label}.	a photo of the {label}.	a blurry photo of the {label}.	a black and white photo of the {label}.
a low contrast photo of the {label}.	a high contrast photo of the {label}.	a bad photo of the {label}.	a good photo of the {label}.
a photo of the small {label}.	a photo of the big {label}.		
Food101			
a photo of {label}, a type of food.			
Caltech101			
a photo of a {label}.	a painting of a {label}.	a plastic {label}.	a sculpture of a {label}.
a sketch of a {label}.	a tattoo of a {label}.	a toy {label}.	a rendition of a {label}.
a embroidered {label}.	a cartoon {label}.	a {label} in a video game.	a plushie {label}.
a origami {label}.	art of a {label}.	graffiti of a {label}.	a drawing of a {label}.
a doodle of a {label}.	a photo of the {label}.	a painting of the {label}.	the plastic {label}.
a sculpture of the {label}.	a sketch of the {label}.	a tattoo of the {label}.	the toy {label}.
a rendition of the {label}.	the embroidered {label}.	the cartoon {label}.	the {label} in a video game.
the plushie {label}.	the origami {label}.	art of the {label}.	graffiti of the {label}.
a drawing of the {label}.	a doodle of the {label}.		
Stanford Cars			
a photo of a {label}.	a photo of the {label}.	a photo of my {label}.	i love my {label}!
a photo of my dirty {label}.	a photo of my clean {label}.	a photo of my new {label}.	a photo of my old {label}.
DTD			
a photo of a {label} texture.	a photo of a {label} pattern.	a photo of a {label} thing.	a photo of a {label} object.
a photo of the {label} texture.	a photo of the {label} pattern.	a photo of the {label} thing.	a photo of the {label} object.
FGVC Aircraft			
a photo of a {label}, a type of aircraft.	a photo of the {label}, a type of aircraft.		
Flowers102			
a photo of a {label}, a type of flower.			
Pets			
a photo of a {label}, a type of pet.			
SUN39			
a photo of a {label}.	a photo of the {label}.		
ImageNet			
a bad photo of a {label}.	a photo of many {label}.	a sculpture of a {label}.	a photo of the hard to see {label}.
a low resolution photo of the {label}.	a rendering of a {label}.	graffiti of a {label}.	a bad photo of the {label}.
a cropped photo of the {label}.	a tattoo of a {label}.	the embroidered {label}.	a photo of a hard to see {label}.
a bright photo of a {label}.	a photo of a clean {label}.	a photo of a dirty {label}.	a dark photo of the {label}.
a drawing of a {label}.	a photo of my {label}.	the plastic {label}.	a photo of the cool {label}.
a close-up photo of a {label}.	a black and white photo of the {label}.	a painting of the {label}.	a painting of a {label}.
a pixelated photo of the {label}.	a sculpture of the {label}.	a bright photo of the {label}.	a cropped photo of a {label}.
a plastic {label}.	a photo of the dirty {label}.	a jpeg corrupted photo of a {label}.	a blurry photo of the {label}.
a photo of the {label}.	a good photo of the {label}.	a rendering of the {label}.	a {label} in a video game.
a photo of one {label}.	a doodle of a {label}.	a close-up photo of the {label}.	a photo of a {label}.
the origami {label}.	the {label} in a video game.	a sketch of a {label}.	a doodle of the {label}.
a origami {label}.	a low resolution photo of a {label}.	the toy {label}.	a rendition of the {label}.
a photo of the clean {label}.	a photo of a large {label}.	a rendition of a {label}.	a photo of a nice {label}.
a photo of a weird {label}.	a blurry photo of a {label}.	a cartoon {label}.	art of a {label}.
a sketch of the {label}.	a embroidered {label}.	a pixelated photo of a {label}.	itap of the {label}.
a jpeg corrupted photo of the {label}.	a good photo of a {label}.	a plushie {label}.	a photo of the nice {label}.
a photo of the small {label}.	a photo of the weird {label}.	the cartoon {label}.	art of the {label}.
a drawing of the {label}.	a photo of the large {label}.	a black and white photo of a {label}.	the plushie {label}.
a dark photo of a {label}.	itap of a {label}.	graffiti of the {label}.	a toy {label}.
itap of my {label}.	a photo of a cool {label}.	a photo of a small {label}.	a tattoo of the {label}.

Table 13: Full list of prompts to evaluate the performance of zero-shot classification on 11 visual recognition datasets.