

FLAIR: VLM with Fine-grained Language-informed Image Representations

Rui Xiao¹, Sanghwan Kim^{1,2,3,4}, Mariana-Iuliana Georgescu^{1,2,3,4}, Zeynep Akata^{1,2,3,4}, Stephan Alaniz^{1,2,3,4}

¹Technical University of Munich ²Helmholtz Munich ³MCML ⁴MDSI

Abstract

CLIP has shown impressive results in aligning images and texts at scale. However, its ability to capture detailed visual features remains limited because CLIP matches images and texts at a global level. To address this issue, we propose FLAIR, **Fine-grained Language-informed Image Representations**, an approach that utilizes long and detailed image descriptions to learn localized image embeddings. By sampling diverse sub-captions that describe fine-grained details about an image, we train our vision-language model to produce not only global embeddings but also text-specific image representations. Our model introduces text-conditioned attention pooling on top of local image tokens to produce fine-grained image representations that excel at retrieving detailed image content. We achieve state-of-the-art performance on both, existing multimodal retrieval benchmarks, as well as, our newly introduced fine-grained retrieval task which evaluates vision-language models' ability to retrieve partial image content. Furthermore, our experiments demonstrate the effectiveness of FLAIR trained on 30M image-text pairs in capturing fine-grained visual information, including zero-shot semantic segmentation, outperforming models trained on billions of pairs. Code is available at <https://github.com/ExplainableML/flair>.

1. Introduction

By encoding images and texts into global embeddings, CLIP achieves coarse-grained semantic understanding. However it loses track of the local image details, e.g. CLIP is not able to perceive the difference between “background” and “frappuccino”, resulting in the inability to highlight the relevant regions specified in the text prompt, as illustrated in Fig. 1. Recently, it has been shown that CLIP models and other vision language models (VLMs) often lack visual details [46, 47]. Thus, our goal is to improve the fine-grained visual understanding of CLIP models which is essential for a wide range of downstream applications, such as image-text retrieval or semantic segmentation.

Previous works [14, 59] propose to generate detailed de-

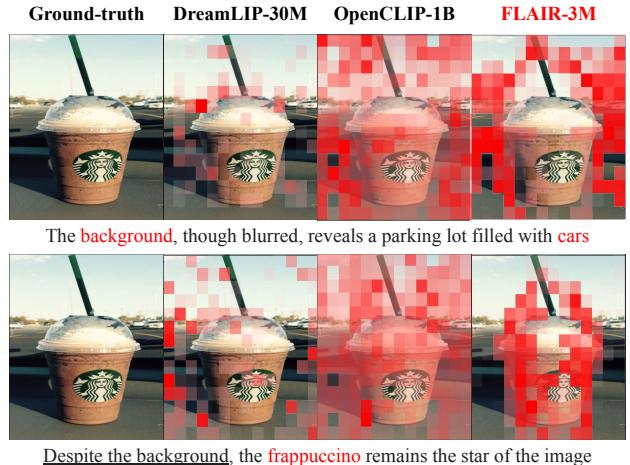


Figure 1. Visualization of the similarity scores between local image tokens and different text queries. While previous works [40, 59] lack fine-grained alignment, FLAIR matches text and image semantics at the token level.

scriptions for images to achieve more localized image-text alignment in CLIP models. However, these methods are restricted by the conventional learning mechanism of CLIP, since the detailed text descriptions enhance visual representations indirectly by matching them through the contrastive loss. Although DreamLIP [59] proposed to supervise local image tokens with textual information, we find that, without a careful selection of the negative pairs in the contrastive loss, the VLM does not learn to align the image tokens with semantically matching text, as illustrated in Fig. 1.

To address these issues, we propose FLAIR, to learn **Fine-grained Language-informed Image Representations**, where image embeddings are generated by conditioning on a relevant text embedding for a more targeted alignment, instead of an indirect alignment through a global loss function. To obtain image descriptions with maximum semantic richness, our method leverages long-caption datasets generated by Multimodal Large Language Models (MLLMs). These captions provide a rich source of information about specific objects or regions in the image. Given a long caption, we sample diverse sub-captions, some of which focus on local regions, while others describe the image globally.

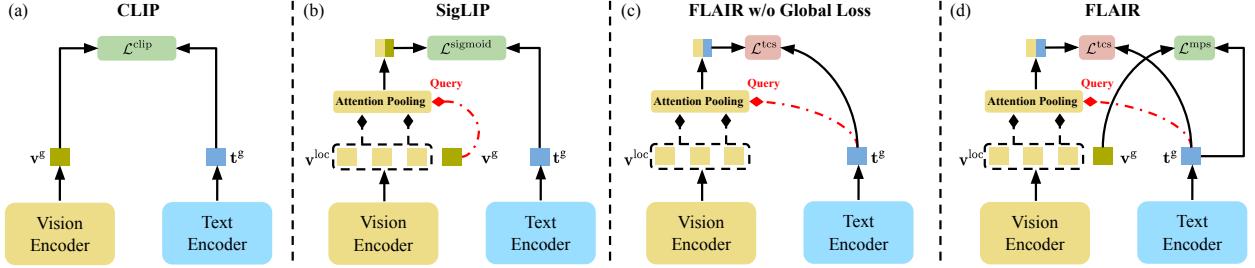


Figure 2. Comparison of text-conditioned attention pooling with previous methods. (a) Vanilla CLIP ($\mathcal{L}^{\text{clip}}$) aligns global image v^g and text t^g tokens. (b) SigLIP ($\mathcal{L}^{\text{sigmoid}}$) employs a global learnable image token v^g as query for a cross-attention to pool the local tokens v^{loc} . (c) FLAIR (\mathcal{L}^{tes}) employs text-conditioned attention pooling that leverages t^g as query, aggregating v^{loc} to capture language-informed visual features. (d) FLAIR ($\mathcal{L}^{\text{tes}} + \mathcal{L}^{\text{mps}}$) adds an extra multi-positive global sigmoid loss to refine global-level image-text alignment.

Considering that these captions describe the image to varying extents, we design our image encoder to produce text-conditioned image representations. To be specific, we introduce an attention pooling operation that uses the caption as a query to pool relevant image-token embeddings together.

As a result, FLAIR learns fine-grained image embeddings that demonstrate strong performance at retrieving fine-grained visual information. As shown in Figure 1, FLAIR can localize image regions relevant to the fine-grained textual description simply by computing the embedding similarity with respect to the individual image tokens. This is in contrast to previous methods that fail to capture local similarity. To analyze the text-image retrieval capabilities of our model, we consider three settings: standard (global) captions, long captions, and a newly proposed fine-grained retrieval setting, where the goal is to match short captions that describe a local region of the image. Our experimental evaluation on multimodal retrieval and zero-shot semantic segmentation demonstrates that FLAIR, trained on 30M image-text pairs with long synthetic captions, significantly outperforms previous vision-language models trained on billions of image-text pairs. While excelling at fine-grained tasks, FLAIR demonstrates comparable performance on global-level tasks, such as image classification, when trained on the same amount of data.

Our key contributions can be summarized as follows: 1) We propose FLAIR, a model architecture that employs text-conditioned attention pooling to produce fine-grained and localized image embeddings. 2) Building upon long synthetic captions, we introduce a diverse caption sampling strategy to obtain a rich set of positive and negative image-text pairs facilitating the learning of global and local multimodal relations. 3) Our experimental evaluation on fine-grained downstream tasks shows that FLAIR, trained on 30M samples, outperforms previous models by up to 10.8% R@1 on coarse-to-fine multimodal retrieval and by up to 11.2% R@1 on long retrieval tasks. Comparing with CLIP models trained on billions of data, FLAIR achieves an average of 14.4% increase in mIOU on segmentation tasks.

2. Related Works

Vision-Language Pre-training. CLIP [40] and ALIGN [21] have scaled up vision-language pre-training datasets to 400M and 1B samples, using a contrastive loss to match global image tokens with global text tokens (Fig. 2 (a)). However, there is a growing demand for more fine-grained alignment between modalities [47]. Several approaches have been proposed to achieve this goal, including token-level alignment [54], hierarchical alignment from global to local [17], soft assignments allowing many-to-many mappings [18], and the use of intra-modal contrastive losses [27]. CoCa [55] utilizes cross-attention to pool the local image tokens and achieves more refined image-to-text alignment by additionally training with a captioning objective. Along with attention pooling to form the global image embeddings (Fig. 2 (b)), SigLIP [57] replaced the Softmax loss of vision-language pre-training with a Sigmoid-based loss. Concurrent to our work, Llip [26] proposed an architecture incorporating language information into learnable image tokens to form contextualized visual representations. However, Llip [26] lacks the pooling of local image tokens (Fig. 2 (c)) and, thus, it does not ensure a fine-grained alignment between modalities. In contrast, FLAIR leverages diverse and detailed captions with both local and global alignment (Fig. 2 (d)), outperforming previous approaches even when training on a significantly smaller dataset.

Text Augmentation. Several works [14, 50, 58, 59] proposed to improve the visual-language alignment through text augmentation. Notably, LaCLIP [14] rewrites captions in large datasets with Large Language Models (LLMs), showing significant performance gains when training on synthetic captions. Similarly, large Vision-Language Models (VLMs) have been exploited to create synthetic images and captions, augmenting existing datasets [20, 30, 53]. DreamLIP [59] re-captions 30 million images from CC3M [44], CC12M [4] and YFCC15M [9] with detailed descriptions generated by pre-trained MLLMs. Employing

these synthetic captions, several models have been trained to handle long texts going beyond the 77-token limit of CLIP. Long-CLIP [58], LoTLIP [50], and TULIP [34] all leverage synthetic captions to achieve this goal. Although trained on the same 30M re-captioned images as DreamLIP, FLAIR changes the image-text interaction by directly using text-conditioned attention pooling to aggregate the local image tokens and choosing informative negative pairs in the loss function. Notably, without modifying the text encoder, the diverse sampling strategy empowers FLAIR to surpass models specialized for long caption retrieval task.

3. FLAIR: Fine-grained Language-informed Image Representations

In this section, we present the FLAIR architecture and methodology for language-image pre-training. We provide an overview of the main components of FLAIR in Fig. 3, including the sampling of diverse captions from long synthetic descriptions (Sec. 3.1), the text-conditioned attention pooling of image tokens (Sec. 3.2), and the local and global loss functions (Secs. 3.3 and 3.4).

3.1. Sampling Diverse Captions

Pre-training data for vision-language models is typically collected by scraping and filtering large amounts of web data, as performed by CC3M [44] or LAION [42, 43]. While a large amount of image-text pairs helps in discovering a comprehensive set of visual concepts, the text descriptions in these datasets often do not describe the image content in detail. As a result, it is not possible to extract fine-grained concepts in an image, such as scene composition, and small object features. To alleviate this issue, we employ image datasets that are synthetically re-captioned and contain a long and detailed description of each image [45, 59]. A single sentence of these long captions typically describes a particular image detail, e.g., one object, a feature of an object, the background, the image style, or context.

Using these captions, our goal is to align vision and language representations at the fine-grained level of individual caption sentences, while retaining global image understanding. We devise a sampling strategy to cover both local and global captions, and learn their similarity with adaptively pooled image features through a contrastive loss. Specifically, when constructing a batch of B images from the augmented dataset, we sample K sub-captions from a caption T_i belonging to an image I_i . Each sub-caption consists of $s \in \{1, \dots, S\}$ sentences that are either randomly sampled (i.e., independently sampled and concatenated), or extracted as a consecutive sequence of sentences. As a result, a batch contains B images and $B \times K$ texts, where each image is associated with K matching captions. At each iteration, we randomly choose the number of sentences s for every sub-caption, where a lower s result in a more localized caption,

while more sentences (a higher s) describe multiple parts of the image, resulting in global descriptions. We provide examples on the original long caption and our sampled captions in Sec. B in the supplementary.

3.2. Text-conditioned Attention Pooling

Having access to a diverse set of captions, some describing local regions of an image and others explaining the global content, motivates creating a model architecture that is capable of adapting to both scenarios. Naively applying a contrastive loss between a global image embedding and the individual text embeddings would collapse the carefully separated information content of our K captions into an averaged image representation.

Instead, we propose to contextualize the image representations with the individual captions, producing a unique image representation for every image-text pair. We start with the VLM architecture as proposed by Radford et al. [40], which uses two independent transformer encoders f_{img} and f_{txt} to project the tokenized image and text samples into per-token embeddings and global embeddings (i.e., $f_{\text{img}}(I) = [\{\mathbf{v}^{(p)}\}_{p=1}^n, \mathbf{v}^g]$ and $f_{\text{txt}}(T) = [\{\mathbf{t}^{(p)}\}_{p=1}^m, \mathbf{t}^g]$), where n is the number of image tokens and m the number of text tokens. For simplicity, we refer to the local image tokens $\{\mathbf{v}^{(p)}\}_{p=1}^n$ as $\mathbf{v}^{\text{loc}} \in \mathbb{R}^{n \times d}$ where d denotes the embedding dimension.

To effectively contextualize the image representation with semantics from the sampled captions, we introduce an attention pooling layer f_{AttnPool} , that produces a text-conditioned image representation \mathbf{v}^{tc} from the local image patch embeddings and the global text embedding. We define $\mathbf{v}^{\text{tc}} = f_{\text{AttnPool}}(\mathbf{t}^g, \mathbf{v}^{\text{loc}})$ as follows:

$$f_{\text{AttnPool}}(\mathbf{t}^g, \mathbf{v}^{\text{loc}}) = \text{softmax} \left(\frac{\mathbf{t}^g W_q (\mathbf{v}^{\text{loc}} W_k)^T}{\sqrt{d}} \right) \mathbf{v}^{\text{loc}} W_v \quad (1)$$

where W_q, W_k, W_v are the query, key, and value weight matrices. In other words, we use the global text embeddings of a caption as a query to pool the local image embeddings creating a text-conditioned image representation \mathbf{v}^{tc} . In practice, we use a multi-head attention layer. We append an empty token (zero vector) to \mathbf{v}^{loc} to allow \mathbf{t}^g to attend to the empty token when \mathbf{t}^g and \mathbf{v}^{loc} are not semantically related.

Choosing Negative Pairs. With text-conditioned attention pooling, FLAIR produces a different image representation for every image-text pair. However, to learn semantically rich and nuanced image representations, we need to carefully define the positive and negative pairs for vision-language pre-training. To simplify notation, we assume a single caption per image (i.e., $K = 1$). Let \mathbf{t}_i^g be the caption of the i -th image in the batch, and $\mathbf{v}_{i,j}^{\text{tc}}$ be the image embedding from the i -th image conditioned on the caption of image j . For the explanation in this section only, we enforce $i \neq j$. In the context of contrastive learning, image-text

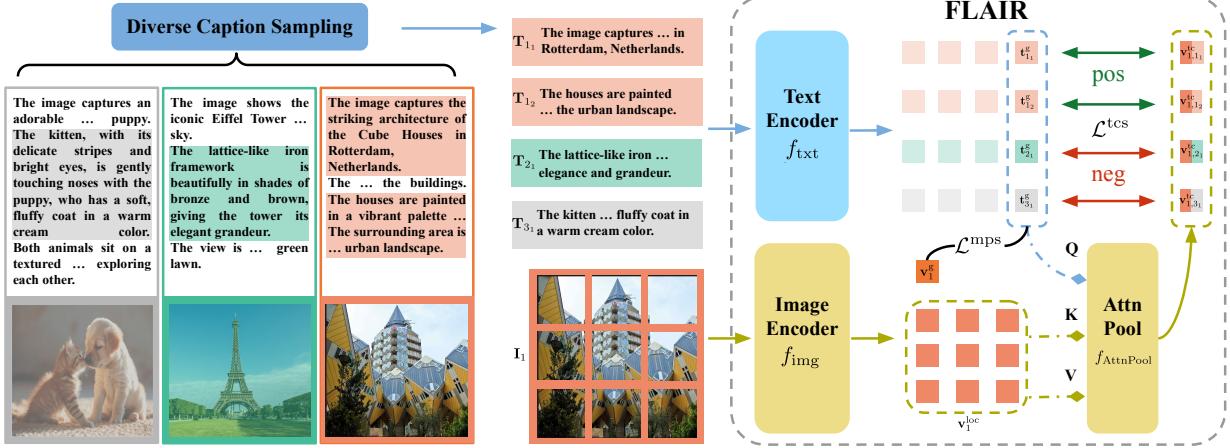


Figure 3. Overview of FLAIR; We sample diverse positive and negative captions $\{T_{11} \dots T_{31}\}$ for an image I_1 . f_{txt} and f_{img} then produce the global text tokens $\{t_{11}^g \dots t_{31}^g\}$, the global image token v_1^g , and local image tokens v_1^{loc} . Conditioned on $\{t_{11}^g \dots t_{31}^g\}$, $f_{\text{AttnPool}}(\cdot)$ generates fine-grained text-conditioned image representations $\{v_{11}^{\text{tc}} \dots v_{31}^{\text{tc}}\}$. The text-conditioned sigmoid loss \mathcal{L}^{tcs} aligns $\{t_{11}^g \dots t_{31}^g\}$ with $\{v_{11}^{\text{tc}} \dots v_{31}^{\text{tc}}\}$ contrastively, while the multi-positive sigmoid loss \mathcal{L}^{mps} refines the global alignment between v_1^g and $\{t_{11}^g \dots t_{31}^g\}$.

pairs where image, caption, and condition come from the same sample are considered positive pairs. In other words, $\langle v_{i,i}^{\text{tc}}, t_i^g \rangle$ is maximized during training, where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. For negative pairs, the text condition introduces multiple options. DreamLIP [59] proposed a loss that uses negatives defined as $\langle v_{i,j}^{\text{tc}}, t_i^g \rangle$ in our formulation. However, this allows to solve the contrastive objective by comparing the text condition with the text embedding, ignoring image information and creating an undesired shortcut. To overcome this problem, we instead propose to adopt $\langle v_{i,j}^{\text{tc}}, t_j^g \rangle$ as negative pairs. This ensures that image and text representations are contrasted meaningfully, i.e., neither image nor text information can be ignored. This definition generalizes to multiple captions per image (i.e., $K > 1$), where each sub-caption of the same image is considered a positive match, and negative otherwise. In this case, we can write the full notation as $\langle v_{i,i_k}^{\text{tc}}, t_{i_k}^g \rangle$ for positive pairs and $\langle v_{i,j_k}^{\text{tc}}, t_{j_k}^g \rangle$ for negative pairs, where k is the sub-caption index of the j -th image. Consequently, these positive and negative pairs allow FLAIR to learn text-aware image representations. Extended analysis is provided in Sec. C in the supplementary.

3.3. Text-conditioned Sigmoid Loss

After constructing the positives and negatives pairs and applying $f_{\text{AttnPool}}(\cdot)$, we adopt a contrastive loss based on the sigmoid function as proposed by SigLIP [57]. It is preferred over the InfoNCE loss [37], as it enables multiple positive pairs in the same batch, and is more efficient for multi-GPU training. Accordingly, we define our text-conditioned sigmoid loss as

$$\mathcal{L}_{i,j,k}^{\text{tcs}} = \frac{1}{1 + e^{y_{i,j}(-t(v_{i,j_k}^{\text{tc}}, t_{j_k}^g) + b)}} \quad (2)$$

where t is a learnable temperature, b is a learnable bias, and $\langle \cdot, \cdot \rangle$ is the cosine similarity. $y_{i,j}$ is $+1$ for positive pairs when $i = j$ for all $k \in [1, \dots, K]$, and -1 for negative pairs otherwise. Since every batch contains B images and BK captions, we reduce the compute and memory usage of $\mathcal{L}_{i,j,k}^{\text{tcs}}$ by considering all K positive pairs, but only $B - 1$ negative pairs per image, i.e., 1 out of K captions for every negative. Therefore, we compute the similarity of $B \times (K + B - 1)$ pairs, instead of $B \times BK$ pairs for every batch.

\mathcal{L}^{tcs} aligns the text-conditioned image embedding with the corresponding text embedding. Intuitively, this allows the image encoder to store semantic information locally in each token and pool the relevant tokens based on the text query producing context-aware representations. Our main experiments demonstrate that the text-conditioned image embeddings contribute significantly to fine-grained image-text alignment, providing the majority of the performance improvement in zero-shot semantic segmentation.

3.4. Multi-positive Sigmoid Loss

We find that FLAIR can be trained exclusively with the \mathcal{L}^{tcs} loss. At the same time, it proves beneficial to additionally match the global image embedding v^g with every sub-caption, to also learn a coarse alignment. Following previous works [14, 28, 59], we introduce a multi-positive loss to align the global image embedding v^g with the text embedding t^g of every sub-caption. Different from previous works, we employ the contrastive sigmoid loss to handle multiple positive captions per image in a more natural way. Our multi-positive sigmoid loss is defined as

$$\mathcal{L}_{i,j,k}^{\text{mps}} = \frac{1}{1 + e^{y_{i,j}(-t(v_i^g, t_{j_k}^g) + b)}} \quad (3)$$

where $y_{i,j}$ is $+1$ for all $k \in [1, \dots, K]$ positive pairs with $i = j$, and -1 for negative pairs otherwise. Equivalently to \mathcal{L}^{tcs} , we use all K positive pairs per image and 1 caption from every K negative sub-captions per match.

Since \mathcal{L}^{mps} mainly optimizes the global image and text embeddings, it is beneficial for coarse-grained tasks. We empirically find that combining \mathcal{L}^{mps} with \mathcal{L}^{tcs} consistently improves performance across all tasks, particularly in zero-shot image classification, where global-level alignment is crucial. Our final loss \mathcal{L} is an average of both losses and it is defined by

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}^{\text{tcs}} + \mathcal{L}^{\text{mps}}). \quad (4)$$

4. Experiments

We present our experimental evaluation of FLAIR on the three image-text retrieval settings: standard (Sec. 4.2), fine-grained (Sec. 4.3), and long (Sec. 4.4). In addition, we conduct experiments on zero-shot semantic segmentation (Sec. 4.5) and image classification (Sec. 4.6), qualitatively evaluate the attention maps of FLAIR (Sec. 4.7), and ablate important model components (Sec. 4.8).

4.1. Experimental Setup

Pre-training Datasets. To learn fine-grained image embeddings from descriptive local captions, we pre-train FLAIR on DreamLIP’s [59] re-captioned datasets, which we refer to as CC3M-recap, CC12M-recap, and YFCC15M-recap. Following DreamLIP, we also merged these three datasets into a combined set of 30M samples.

Implementation Details. Our model is based on the Open-CLIP [6] code implementation, adopting their default settings. We use ViT-B/16 as the vision encoder, with the default pre-processing: images are resized to 224×224 pixels, and text sequences are tokenized to a maximum length of 77 tokens. For direct comparison with DreamLIP, we follow their training configuration and caption pre-processing, splitting the MLLM-generated and original captions into individual sentences. To obtain diverse training captions, we sample $K = 8$ captions per image, with each caption randomly merging 1 to 3 sentences (i.e., $S = 3$). To maximize sampling variability while retaining context, we randomly construct our sub-caption by either sampling consecutive sentences or merging sentences from random positions in the original text. For fair comparison, we reproduce CLIP [40] and SigLIP [57] on all re-captioned datasets under identical training configurations as FLAIR. Further details are available in Sec. E in the supplementary.

Inference with FLAIR. To utilize the fine-grained embeddings from FLAIR for image-to-text (I2T) retrieval, each image i is first conditioned on all j texts to generate the conditioned embeddings $\mathbf{v}_{i,j}^{\text{tc}}$. Then we compute the similarity scores between the conditioned embeddings and each text

embedding ($\langle \mathbf{v}_{i,j}^{\text{tc}}, \mathbf{t}_j^{\text{g}} \rangle$) to obtain Recall@K from the top-K retrieval items. The text-to-image (T2I) retrieval score matrix is the transpose of the image-to-text retrieval matrix.

4.2. Standard Zero-shot Image-text Retrieval

As a standard assessment of image-text alignment, we follow prior works [26, 53, 59] to evaluate image-text retrieval on the validation splits of MSCOCO [29] and Flickr30K [39], where each image is typically paired with five global captions.

Results. We report the results on the standard retrieval task in the left side of Tab. 1. FLAIR outperforms the three baselines, CLIP, SigLIP, and DreamLIP on all pre-training datasets by a large margin. Comparing models trained on CC3M-recap, FLAIR surpasses DreamLIP in the retrieval task, obtaining higher R@1 scores on both COCO (T2I: +7.9%, I2T: +10.8%) and Flickr30k (T2I: +12.1%, I2T: +9.5%) datasets. When including SOTA models, FLAIR trained on CC12M-recap obtains a similar performance to SigLIP trained on 10B samples, and surpasses it significantly once we move to larger datasets with YFCC15M-recap and the merged 30M samples. FLAIR-30M (vs. SigLIP-10B) achieves 81.1% (vs. 75.6%) T2I, 94.7% (vs. 89.1%) I2T on Flickr30k and is similarly better on COCO. We also notice that CLIP and SigLIP trained on YFCC15M-recap can match or surpass their counterparts trained on billions of data samples. This suggests two key insights: 1) text augmentations from long synthetic captions empowers VLMs with better retrieval capability, and 2) FLAIR with text-conditioned attention pooling generates more targeted image embeddings for retrieval, maximizing the benefits from text augmentations, and resulting in a significant improvement with much less image data.

4.3. Fine-grained Zero-shot Image-text Retrieval

Standard retrieval tasks do not fully capture a model’s ability to align detailed descriptions with images. To address this, we introduce a fine-grained retrieval task aimed at evaluating how well a model can associate an image with fine-grained captions. Our benchmark is constructed as follows: 1) We use the recently released densely-captioned datasets DOCCI [36] and IIW [19]. Due to the careful human annotation process, their long captions are free from hallucinations; 2) For each test image in DOCCI and IIW, we split the long captions into individual sentences, yielding an average of 7.1 captions per image in DOCCI and 10.1 in IIW. As shown in Fig. 10, each caption focuses on a specific local part of the image, making both T2I and I2T tasks significantly more challenging than standard retrieval. We refer to our split datasets as DOCCI-FG and IIW-FG.

Results. The results obtained on DOCCI-FG and IIW-FG are reported in the right side of Tab. 1. The difficulty of this task is apparent by the significantly lower text-to-image

Setting	Method	Standard Retrieval								Fine-grained Retrieval							
		MSCOCO				Flickr30k				DOCCI-FG				IIW-FG			
		T2I	I2T	T2I	I2T	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	T2I	I2T	T2I	I2T
CC3M-recap	CLIP [40]	27.0	52.6	38.9	66.1	49.9	77.0	67.8	88.5	10.3	23.4	25.0	50.9	24.4	45.0	61.1	85.6
	SigLIP [57]	28.3	54.4	40.1	67.5	53.2	78.5	69.9	90.4	10.4	23.8	24.9	50.6	24.7	45.1	62.6	86.9
	DreamLIP [59]	29.8	55.4	40.8	68.4	53.6	78.4	69.2	91.5	10.3	22.8	23.3	47.5	22.7	41.6	59.2	83.3
	FLAIR	37.7	64.5	51.6	77.2	65.7	86.8	78.7	95.2	15.1	30.9	35.7	63.5	30.5	52.3	70.6	90.9
CC12M-recap	CLIP [40]	39.8	66.4	56.2	80.5	67.0	87.7	81.7	96.5	16.0	31.6	39.5	66.5	31.8	52.9	76.4	93.6
	SigLIP [57]	40.4	67.0	55.3	79.7	66.7	88.1	82.5	96.1	16.2	31.9	40.0	66.8	31.9	53.2	78.4	94.3
	DreamLIP [59]	40.6	66.5	54.0	78.3	68.3	89.3	84.1	97.8	17.2	33.0	41.6	68.5	31.9	52.1	77.8	94.9
	FLAIR	47.8	73.5	64.1	85.0	75.4	92.15	90.8	98.4	21.4	38.8	50.4	76.7	38.7	59.9	83.8	96.9
YFCC15M-recap	CLIP [40]	44.7	71.2	61.0	85.0	72.3	90.8	89.1	97.6	18.1	35.3	43.1	71.9	34.4	56.5	81.4	96.7
	SigLIP [57]	46.6	72.8	62.6	85.3	73.6	92.1	90.0	97.6	18.9	35.8	46.3	74.8	35.5	56.6	84.3	96.2
	DreamLIP [59]	42.4	68.5	57.0	81.0	70.0	89.2	87.3	98.1	17.3	33.6	41.4	69.8	32.0	53.0	76.1	95.4
	FLAIR	51.2	76.0	67.3	88.1	79.2	94.2	93.3	99.1	23.0	41.2	53.7	79.7	39.5	62.1	85.5	96.4
SOTA Comparison	OpenCLIP (2B) [6]	41.7	67.1	59.3	82.4	71.9	90.4	87.5	97.7	17.4	31.9	49.7	75.9	30.6	48.4	84.1	95.4
	SigLIP (10B) [57]	47.2	72.1	65.5	86.2	75.6	92.8	89.1	98.6	20.6	35.9	57.5	82.1	33.8	53.0	83.7	97.7
	MetaCLIP (2.5B) [52]	41.4	67.2	59.4	80.6	76.2	90.7	85.9	97.3	-	-	-	-	-	-	-	-
	Llip (2.5B) [26]	45.6	70.8	63.4	84.3	75.1	92.8	90.1	98.5	-	-	-	-	-	-	-	-
	DreamLIP (30M) [59]	44.8	69.8	62.3	84.5	73.3	91.8	89.9	99.0	21.6	39.3	51.2	78.3	37.5	58.6	85.3	97.4
	FLAIR (30M)	53.3	77.5	68.0	87.8	81.1	94.9	94.7	99.3	25.0	43.8	59.0	84.1	41.7	63.4	91.5	98.9

Table 1. Zero-shot image-text retrieval on validation splits for standard benchmarks (Flickr30k [39] and MSCOCO [29]) and our introduced fine-grained retrieval setting (sentence-level on DOCCI [36] and IIW [19]). Except for “SOTA Comparison”, all models are pre-trained on CC3M-recap, CC12M-recap, YFCC15M-recap, under the same training configurations. All models use ViT-B/16 as the vision encoder.

Method	Data	DCI	SV-1k	SV-10k	Urban-1k												
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I
OpenCLIP [40]	2B	56.0	55.4	90.3	87.7	69.6	66.8	69.5	65.8								
LiT [56]	100M	41.7	40.9	86.0	80.0	61.4	50.6	-	-								
ALIGN [21]	700M	56.5	57.4	86.3	85.3	65.1	62.7	-	-								
SigLIP [57]	10B	57.7	56.2	85.8	83.4	83.4	63.0	62.7	62.1								
Long-CLIP [58]	400M	47.4	44.1	90.6	87.4	73.1	62.0	78.9	79.5								
LoTLIP [50]	100M	62.1	61.0	95.5	86.8	81.4	83.7	-	-								
FLAIR	3M	47.3	50.5	91.0	89.7	72.0	70.6	63.5	69.5								
FLAIR	12M	55.5	60.8	96.1	95.1	85.0	83.4	74.6	80.6								
FLAIR	15M	54.9	<u>62.4</u>	<u>97.4</u>	<u>96.7</u>	<u>88.8</u>	<u>86.8</u>	<u>82.4</u>	<u>86.6</u>								
FLAIR	30M	<u>61.3</u>	66.2	98.5	98.0	90.3	89.4	83.6	87.7								

Table 2. Zero-shot long text-image retrieval tasks. I2T and T2I indicate the R@1 score on image-to-text and text-to-image retrieval, respectively. The best results are **bold**, second-best are underlined. All models use ViT-B/16 as vision encoder.

(T2I) retrieval scores compared to standard retrieval. Despite that, FLAIR consistently outperforms baselines across all training configurations and even surpasses the CLIP and SigLIP models trained on billions of samples. Interestingly, FLAIR trained on CC12M-recap achieves higher R@1 scores (38.7%), in terms of T2I retrieval on IIW-FG, compared to SigLIP-10B (33.8%).

On the 30M dataset, the performance of FLAIR further improves to 41.7%, outperforming DreamLIP by 4.2% in T2I retrieval (R@1). Overall, FLAIR achieves an increased between 3.4% and 7.8% in R@1 scores compared to DreamLIP. These results demonstrate that FLAIR learns to align images with detailed, fine-grained captions more effectively than the baselines.

4.4. Long Zero-shot Image-Text Retrieval

Image-text retrieval with long captions imposes a unique challenge for CLIP models. Following LoTLIP [50] and Long-CLIP [58], we evaluate FLAIR on datasets with long captions, including DCI [48], 1k (SV-1k) and 10k (SV-10k) subsets of ShareGPT-4V [5], and Urban-1k [58], with the results presented in Tab. 2.

Unlike previous methods specifically designed for long-caption retrieval with extended token limits and larger text encoders, FLAIR employs the standard CLIP text encoder with a 77-token limit. The former SOTA, LoTLIP [50], was trained on a 100M-scale re-captioned dataset, while Long-CLIP [58] fine-tunes a 400M-scale CLIP model with an additional 1M images with long captions. Although FLAIR is trained on a smaller training set of 30M samples, it still outperforms these methods on SV-1k, SV-10k, and Urban-1k. Most notably on T2I, FLAIR obtained improvements

Method	Data Size	VOC20	Cityscapes	Context59	ADE20K	COCO-Stuff	Average
CLIP [40]	400M	41.8	5.5	9.2	3.2	4.4	12.8
OpenCLIP [6]	2B	47.2	5.1	9.0	2.9	5.0	13.9
MetaCLIP [52]	2.5B	35.4	5.0	8.1	2.2	4.3	11.0
CLIP [57]		11.3	5.0	4.5	1.3	2.8	5.0
SigLIP [57]	30M	14.5	5.5	5.8	2.2	3.8	6.4
DreamLIP [59]		1.8	0.9	0.4	0.1	0.1	0.7
FLAIR	3M	60.9	20.6	23.8	13.2	13.1	26.3
FLAIR	12M	69.7	20.1	22.9	13.3	15.4	28.3
FLAIR	15M	66.7	16.5	17.4	9.1	13.6	24.7
FLAIR	30M	73.0	13.6	18.6	10.4	13.3	25.8

Table 3. Mean intersection over union (mIoU) for zero-shot semantic segmentation on the VOC20 [13], Cityscapes [8], Context59 [33], ADE20K [60], and COCO-Stuff [2] datasets. All models employ ViT-B/16 as vision encoder.

of 10.4%, 5.7%, and 8.2% in terms of R@1 over the previous SOTA. Remarkably, FLAIR trained on 15M samples already surpasses all previous methods on 3 out of 4 datasets.

This significant performance gain can be explained as follows: 1) text-conditioned attention pooling can adapt to the rich semantics in long texts to extract all relevant visual information. 2) By sampling diverse captions the model becomes aligned with the distribution of long captions.

4.5. Zero-shot Semantic Segmentation

For VLMs, zero-shot semantic segmentation involves measuring the similarity $\{\langle \mathbf{v}_i^{\text{loc}}, \mathbf{t}_j^g \rangle \mid j \in \{1, 2, \dots, M\}\}$ for M different class names. Recent works [12, 25, 49] provide a framework to map these similarities to semantic segmentation outputs. To examine the raw alignment of local image tokens \mathbf{v}^{loc} with the corresponding input texts, we perform semantic segmentation following [49] without post-processing or segmentation-specific modifications.

As shown in Tab. 3, FLAIR trained on all subsets of data, consistently outperforms CLIP-based methods trained on significantly larger datasets, achieving an improvement of 10.1% - 25.8% mIOU increase across all datasets (14.4% on average). As illustrated in Fig. 1, DreamLIP’s \mathbf{v}^{loc} image token embeddings show weak correspondence to the input text, which we conclude is the result of their choice of negatives as discussed in Sec. 3.2. In contrast, FLAIR, optimized with \mathcal{L}^{tcs} , effectively aligns \mathbf{v}^{loc} with varying text prompts, demonstrating strong localization capabilities.

Method	Data Size	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	ImageNet	Average
LaCLIP [14]		14.2	57.1	27.5	35.1	1.6	1.6	16.6	15.6	52.7	14.7	21.5	23.5
MLLM-A [30]		18.7	58.4	32.4	43.8	3.9	1.5	20.2	32.1	63.5	17.5	25.0	28.8
CLIP [40]	3M	17.9	75.0	40.8	43.1	2.6	1.0	15.3	22.1	68.9	12.6	23.8	29.4
SigLIP [57]		18.4	<u>76.4</u>	41.9	46.9	3.0	1.4	17.6	20.6	70.4	10.8	25.4	30.3
DreamLIP [59]		<u>23.1</u>	75.9	<u>44.2</u>	46.6	<u>3.4</u>	<u>1.6</u>	<u>19.0</u>	<u>27.4</u>	<u>66.1</u>	<u>16.0</u>	<u>30.1</u>	<u>32.1</u>
FLAIR		24.2	82.0	51.5	53.8	3.7	1.7	23.9	34.2	70.1	19.1	33.8	36.2
CLIP [40]		61.3	92.2	66.9	62.2	19.3	5.7	30.9	49.3	83.7	43.4	50.0	51.4
SigLIP [57]	30M	64.2	91.0	67.6	<u>64.0</u>	22.0	5.7	33.5	53.3	84.3	43.6	51.0	52.7
DreamLIP [59]		75.4	<u>92.3</u>	70.7	63.7	<u>22.7</u>	7.9	<u>33.9</u>	64.1	88.0	51.1	58.1	57.0
FLAIR		<u>72.5</u>	93.1	<u>69.6</u>	66.9	31.1	<u>7.2</u>	37.3	<u>55.6</u>	<u>86.5</u>	<u>48.4</u>	<u>56.6</u>	<u>56.8</u>
OpenCLIP [6]	2B	86.2	94.8	76.5	<u>70.0</u>	87.4	25.8	54.9	89.5	93.2	69.8	70.2	74.4
MetaCLIP [52]	2.5B	88.3	95.7	<u>79.0</u>	68.5	<u>82.9</u>	<u>30.3</u>	<u>62.1</u>	91.7	<u>93.3</u>	<u>73.9</u>	<u>72.1</u>	76.2
Llip [26]	2.5B	89.0	95.7	81.4	70.9	88.2	41.5	63.7	93.5	94.7	74.9	75.3	79.0

Table 4. Top-1 accuracy for zero-shot classification on: Food-101 [1], CIFAR-10 & CIFAR-100 [24], SUN397 [51], Cars [23], Aircraft [32], DTD [7], Pets [38], Caltech-101 [15], Flowers [35], ImageNet [10]. All models use ViT-B/16 as vision encoder. The best and second-best results are **bold** and underlined.

4.6. Zero-shot Image Classification

Following [14, 59], we evaluate the zero-shot classification performance of FLAIR and baseline methods on ImageNet [10] and 10 additional datasets, as shown in Tab. 4. In retrieval tasks, text-augmented methods outperform VLMs trained on billions of images. However, in image classification they lag behind by around 20%. This demonstrates that scaling up the number of images remains a key factor in improving VLM’s classification performance.

When trained on CC3M-recap, FLAIR achieves a 4.1% higher average performance than DreamLIP and other baselines. This shows that FLAIR, although optimized to generate fine-grained visual representations, could still efficiently gain global-level visual understanding performance when images are relatively scarce. However, when scaled up to 30M samples, FLAIR, while still outperforming CLIP and SigLIP by 4%, is on par with DreamLIP (-0.2% on avg.). This shows that these methods trained on synthetic captions converge similarly, further suggesting the importance of scaling up images for the classification task. Therefore, we hypothesize that scaling FLAIR to larger datasets would extend the concept vocabulary and image coverage, closing the gap to large-scale models on zero-shot classification.

4.7. Attention Maps Visualization

For a given image with two different local captions, we visualize the attention maps of $f_{\text{AttnPool}}(\cdot)$, i.e., which image tokens are pooled together, in Fig. 4. As illustrated by the “truck” and “worker” example, FLAIR can locate both large and small objects in an image. The horses example shows



Figure 4. Visualization of attention maps in the attention pooling layer $f_{\text{AttnPool}}(\cdot)$. Regions of high attention are highlighted in red.

that FLAIR is able to differentiate objects by their individual properties such as color and location. Notably, FLAIR is also precise in identifying individual parts of an object, exemplified by the eyes, mouth, and paw of the dog, where the focus lies on the one that is raised. These results show FLAIR’s strong sensitivity to semantic details.

4.8. Ablation Study

Model Components. In Tab. 5, we analyze the components of FLAIR: text conditioning (TC), global loss (GL), multiple captions per image (MC), and diverse caption sampling instead of single sentences (DS). \mathcal{L}^{tes} and \mathcal{L}^{mps} correspond to TC+MC and GL+MC respectively.

SigLIP is equivalent to only using GL (1). Replacing GL with TC (2) leads to performance improvements across all metrics, achieving a 3.7%/5.5% increase in R@1 for COCO retrieval and a 33.8% boost in VOC20 segmentation demonstrating its contribution to the fine-grained alignment. Adding MC improves performance in both scenarios (3 and 4). Our diverse sampling (DS) is another significant improvement, especially in segmentation and long-retrieval performance on the Urban-1K which gains over 20% (5 and 6). FLAIR, combining all components, achieves the best performance in all but long-retrieval (7). In summary, \mathcal{L}^{tes} is foundational to our method’s performance, sampling diverse captions provides a substantial boost in long retrieval tasks, while combining \mathcal{L}^{mps} delivers additional gains, particularly for global-level tasks.

Additional Ablations. In supplementary Sec. D, we provide additional ablations to support important choices. We pre-trained FLAIR on the original CC3M dataset and on the PixelProse [45] dataset with synthetic captions generated by Gemini-Pro [41], showing that our method is not restricted to long captions and adaptable to a variety of data distri-

Method	COCO				DOCCI		Urban-1K		VOC20		ImageNet	
	GL	TC	MC	DS	T2I	I2T	T2I	I2T	mIOU	Top-1		
1 ✓					28.3	40.1	10.4	24.9	42.8	40.5	3.1	25.4
2 ✓					32.0	45.6	12.7	30.9	44.4	42.6	36.9	28.1
3 ✓			✓		32.9	44.6	13.3	31.0	47.9	46.5	1.7	27.9
4 ✓		✓	✓		34.8	47.1	14.1	30.2	46.6	40.9	34.1	29.4
5 ✓		✓	✓	✓	35.0	49.1	13.0	33.1	70.7	64.6	7.2	32.0
6 ✓		✓	✓	✓	36.2	50.0	13.8	34.6	68.3	63.1	46.5	31.5
7 ✓	✓	✓	✓	✓	37.7	51.6	15.1	35.7	69.5	63.5	59.7	33.8

Table 5. Ablation study on different components of FLAIR on the CC3M-recap dataset. **GL:** Global Loss, **TC:** Text Conditioning, **MC:** Multiple Captions, **DS:** Diverse Sampling.

butions. By varying the sampling strategy of the diverse captions, we find that it is crucial across tasks to sample both short and long captions instead of only a fixed length. By testing a different number of multiple captions K ranging from 2 to 10, we observe that performance converges at around 8 captions. Finally, we ablate the maximal number of sampled sentences S and observe that merging 3 sentences achieved the most balanced results.

5. Conclusion and Limitations

We introduce FLAIR, a VLM that learns **Fine-grained Language-informed Image Representations** by conditioning on the semantics in dense local captions. Trained on 30M recaptioned images, FLAIR outperforms baselines trained on billions of images across standard, fine-grained, and long-form image-text retrieval tasks. The significant improvements in zero-shot segmentation compared to the baselines as well as the qualitative results corroborate that FLAIR learns a fine-grained alignment between text and image at the token-level.

While FLAIR matches baselines trained on the same number of images in zero-shot classification, it still falls behind CLIP models trained on significantly larger datasets. This suggests that, although leveraging detailed synthetic captions enhances fine-grained image understanding, it does not replace the image coverage and conceptual richness of larger datasets for global-level tasks. To tackle this limitation, a natural future direction involves scaling the synthetic recaptioning to large-scale datasets and training variants of FLAIR with higher parameter count.

Acknowledgments This work was partially funded by the ERC (853489 - DEXIM) and the Alfried Krupp von Bohlen und Halbach Foundation, which we thank for their generous support. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer JUWELS [22] at Jülich Supercomputing Centre (JSC).

References

- [1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining Discriminative Components with Random Forests. In *ECCV*, 2014. 7
- [2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 7, 17
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 12
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 2, 15
- [5] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *ECCV*, 2024. 6, 14
- [6] 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. In *CVPR*, 2023. 5, 6, 7, 12, 13, 14, 16, 17
- [7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 7
- [8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 7, 17
- [9] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv:2203.05796*, 2022. 2
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7
- [11] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 15
- [12] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, 2023. 7
- [13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 7, 17
- [14] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *NeurIPS*, 2024. 1, 2, 4, 7, 17
- [15] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*, 2004. 7
- [16] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacom: In search of the next generation of multimodal datasets. *NeurIPS*, 2024. 12, 15
- [17] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *NeurIPS*, 2022. 2
- [18] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *AAAI*, 2024. 2
- [19] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Burner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv:2405.02793*, 2024. 5, 6, 13, 14
- [20] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv:2402.01832*, 2024. 2
- [21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2, 6
- [22] Jülich Supercomputing Centre. JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, 2021. 8
- [23] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a Large-Scale Dataset of Fine-Grained Cars, 2013. 7
- [24] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images, 2010. 7
- [25] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *arXiv:2407.12442*, 2024. 7
- [26] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mido Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. In *ICML*, 2024. 2, 5, 6, 7, 17
- [27] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uni-clip: Unified framework for contrastive language-image pre-training. *NeurIPS*, 2022. 2
- [28] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv:2110.05208*, 2021. 4
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 6, 14, 15, 16

- [30] Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. Mllms-augmented visual-language representation learning. *arXiv:2311.18765*, 2023. 2, 7
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017. 17
- [32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft, 2013. 7
- [33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 7, 17
- [34] Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M Asano, Nanne van Noord, Marcel Worring, and Cees GM Snoek. Tulip: Token-length upgraded clip. *arXiv:2410.10034*, 2024. 3
- [35] Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 7
- [36] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. *arXiv:2404.19753*, 2024. 5, 6, 12, 13, 14, 15
- [37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [38] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 7
- [39] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 5, 6, 14, 16
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 5, 6, 7, 15, 16, 17
- [41] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024. 8, 15
- [42] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021. 3
- [43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 3
- [44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2, 3, 15, 16
- [45] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to pose: A large dataset of dense image captions. *arXiv:2406.10328*, 2024. 3, 8, 15, 16
- [46] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv:2406.16860*, 2024. 1
- [47] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 1, 2
- [48] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *CVPR*, 2024. 6, 14
- [49] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv:2312.01597*, 2023. 7, 17
- [50] Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. Lotlip: Improving language-image pre-training for long text understanding. *arXiv:2410.05249*, 2024. 2, 3, 6, 14
- [51] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 7
- [52] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv:2309.16671*, 2023. 6, 7, 17
- [53] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*, 2023. 2, 5, 17
- [54] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv:2111.07783*, 2021. 2
- [55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv*, 2022. 2
- [56] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *CVPR*, 2022. 6
- [57] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 2, 4, 5, 6, 7, 16
- [58] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *ECCV*, 2024. 2, 3, 6

- [59] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, 2024. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [12](#), [13](#), [14](#), [15](#)
- [60] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. [7](#), [17](#)

FLAIR: VLM with Fine-grained Language-informed Image Representations

Supplementary Material

In this supplementary file, we illustrate more qualitative results in Sec. A, describe the datasets in Sec. B, and present an extensive analysis of the impact of the negative pairs on the FLAIR performance in Sec. C. We further present additional ablation experiments in Sec. D, and the implementation details in Sec. E.

A. Qualitative Results

Attention Maps Visualization. We provide a comprehensive visualization of attention maps of $f_{\text{AttnPool}}(\cdot)$ in Fig. 5 and Fig. 6. We follow DINO [3] to aggregate attention maps from multiple heads. We empirically found that heads 1,4,6,8 mainly focus on foreground objects and aggregate these attention maps to form the visualization. In Fig. 5, we show that the attention maps focus on different parts of an image w.r.t. the local captions. Interestingly, in the “fireplace” example (second row), the attention correctly localizes the “white candle” (second row, second column), which is exactly what the caption describes, although “fireplace” also appears in the sentence. This demonstrates that FLAIR is able to locate an object based on the main semantics of a prompt, instead of simply matching “a bag of words”.

In Fig. 6, we visualize the attention maps w.r.t. long captions. When multiple objects appear in a long caption, FLAIR is able to locate them at the same time. Notably, in the “room” example (second row), FLAIR ignores descriptions like “adding a touch of nature to the room” and solely focus on the main semantics: “black shelf”, “books” and “lamp”. This might reveal one possible future application of FLAIR, understanding the main semantics in complex prompts and grounding the main objects in the image.

Token-to-Text Similarity. We also visualize the similarity between local image tokens and text prompts in Fig. 1 of the main paper. This similarity between the local image tokens and the text prompts could reflect the model’s localization capability, which is closely related to the segmentation task. We provide extra visualizations in Fig. 7. We use FLAIR pre-trained on the CC3M-recap dataset to compare with DreamLIP [59] trained on Merged30M dataset and OpenCLIP trained on DataComp-XL [16]. As illustrated, compared to OpenCLIP [6] that tends to make over-predictions, FLAIR is able to accurately localize the tokens w.r.t. the text prompts, especially on fine-grained details such as “flower on the cake” and “bird on the branch”. This further validates that the fine-grained representations learned by FLAIR are indeed sensitive to the text semantics.

Retrieval Visualization. For the fine-grained image-text retrieval task on the DOCCI [36] benchmark, we visualize

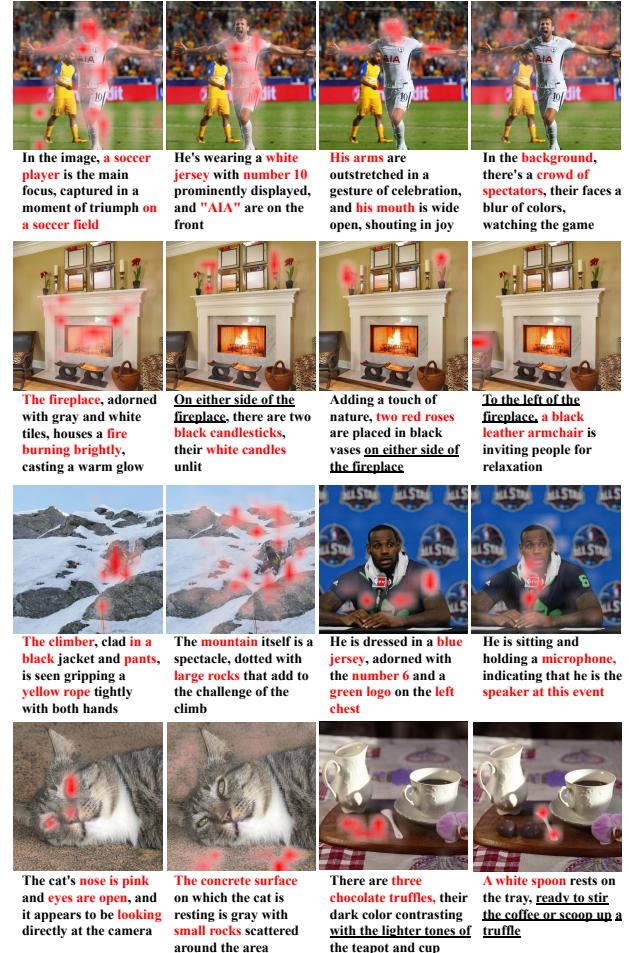


Figure 5. Visualization of the attention maps w.r.t. fine-grained captions. In the images, regions with high attention scores are marked in red; in the captions, objects representing the main semantics of the sentences are marked in red, while objects with less semantic significance are underlined.

the top-5 retrieved captions for a given image, highlighting incorrect captions in red. We compare FLAIR with OpenCLIP [6] trained on 2B samples in Fig. 8. From top to bottom, the similarity scores decrease. Interestingly, compared to OpenCLIP [6], FLAIR tends to retrieve “local” captions first. For example, the top-1 retrieved caption for FLAIR is only describing the “spotlight”, while OpenCLIP retrieves “a nighttime view of an artificial waterfall”, which can be considered a global description for this image. The incorrectly retrieved captions of OpenCLIP contain relevant keywords like “waterfall”, while FLAIR retrieves the captions

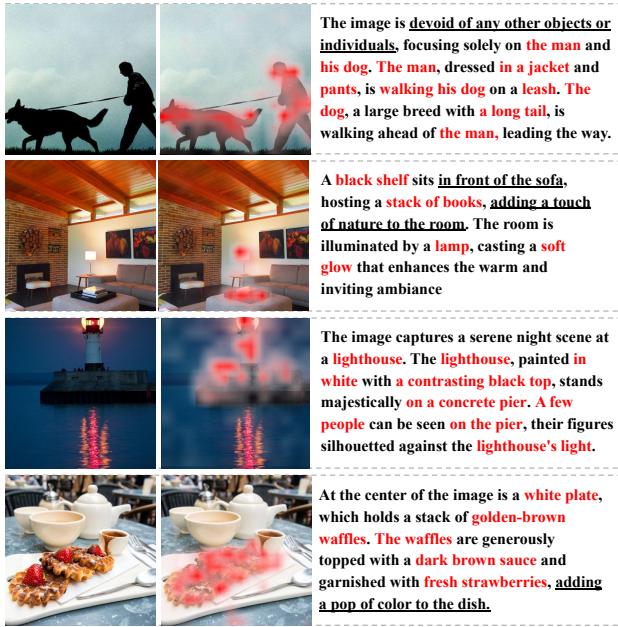


Figure 6. Visualization of the attention maps w.r.t. fine-grained long captions. In the images, regions with high attention scores are marked in red; in the captions, objects representing the main semantics of sentences are marked in red, while objects with less semantic significance are underlined.

correctly based on a more detailed understanding of the image semantics.

B. Dataset Details

Pre-training Data. FLAIR is pre-trained on CC3M-recap, CC12M-recap, YFCC15M-recap and Merged-15M [59], where each image is equipped with long synthetic captions generated by various MLLMs. Fig. 9 shows an example of the original long captions produced by DreamLIP [59] together with our diverse sampled captions. We take the whole paragraph of the long synthetic caption and split it into sentences. Our K diverse captions are sampled from these sentences, and each caption can contain $s \in \{1, \dots, S\}$ merged sentences. In our experiments, we set $S = 3$ and $K = 8$. We detail this choice in Sec. D.4 and Sec. D.3.

Fine-grained Retrieval Data. In order to create the new fine-grained retrieval task, we split the original long captions from DOCCI [36] and IIW [19] into separate sentences. Each sentence can either describe the image globally or describe the fine-grained details of an image. These captions, together with the original images, form our DOCCI-FG and IIW-FG retrieval benchmarks. We provide a visualization of DOCCI-FG containing two images with all the corresponding paired captions in Fig. 10. As illustrated in Fig. 10, the split captions are likely to describe one local part of an image, such as “The wings and chest of

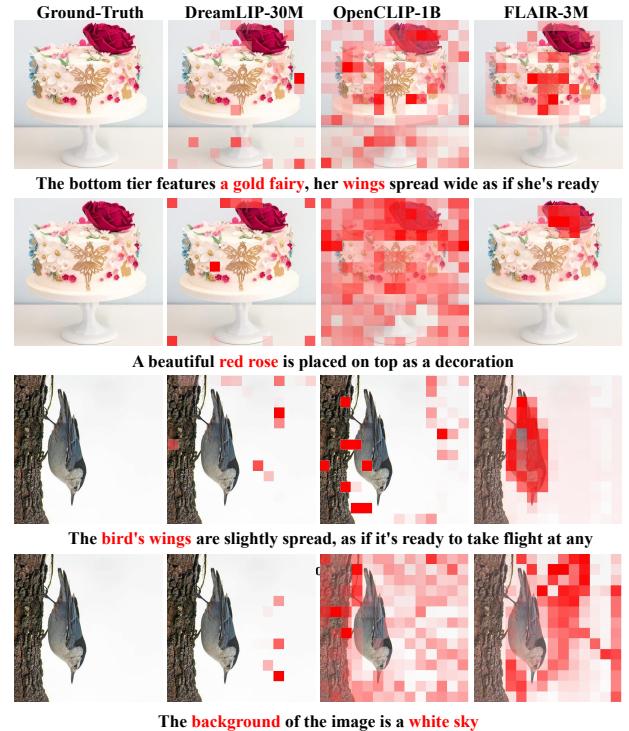


Figure 7. Visualization of the similarity scores between local image tokens and different text queries. While previous works [6, 59] lack fine-grained alignment, FLAIR matches text and image semantics at the token level.

the hawk are dark brown, and the left side of it is lit up by white light”. We provide detailed statistics on the number of images, captions, and the average number of tokens per caption for standard, fine-grained, and long retrieval benchmarks in Tab. 6. DOCCI-FG and IIW-FG contain an average of 7.1 and 10.1 captions per image, respectively, with each caption comprising approximately 18.76 and 22.56 tokens.

C. Extended Analysis of Negatives

As discussed in the methodology section of the main paper, FLAIR produces a unique image representation for each image-text pair using the text-conditioned attention pooling. Specifically, the text-conditioned embedding \mathbf{v}^{tc} is jointly conditioned by the local image tokens \mathbf{v}^{loc} and global text tokens \mathbf{t}^g :

$$\mathbf{v}^{tc} = f_{AttnPool}(\mathbf{v}^{loc}, \mathbf{t}^g)$$

When considering the global text token \mathbf{t}^g , which forms both positive and negative pairs in \mathcal{L}^{tes} , one positive pair ($\langle \mathbf{v}_{i,i_k}^{tc}, \mathbf{t}_{i_k}^g \rangle$) and five types of negative pairs can be identified. As visually depicted in Fig. 11, these negatives are:

$$\langle \mathbf{v}_{i,j_k}^{tc}, \mathbf{t}_{j_k}^g \rangle, \langle \mathbf{v}_{i,j_k}^{tc}, \mathbf{t}_{i_k}^g \rangle, \langle \mathbf{v}_{i,i_k}^{tc}, \mathbf{t}_{j_k}^g \rangle, \langle \mathbf{v}_{i,j_k}^{tc}, \mathbf{t}_{l_k}^g \rangle, \langle \mathbf{v}_{i,i_k}^{tc}, \mathbf{t}_{i_m}^g \rangle$$

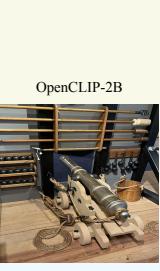
 OpenCLIP-2B	<ul style="list-style-type: none"> ■ A nighttime view of an artificial waterfall ■ The face of the waterfall is completely illuminated ■ A spotlight is on a tropical plant on the top right of the waterfall ■ Each water spout is backlit with warm white light ■ A brown dirt pile with light pink flowers is seen behind the flower pot, and three pink flowers are partially seen above the waterfall feature ■ A spotlight is on a tropical plant on the top right of the waterfall ■ The face of the waterfall is completely illuminated ■ Besides the waterfall, the remaining portion of the frame is filled with tropical plants that are mostly in the shadows ■ Water cascades evenly down to a little lit pool of water ■ A nighttime view of an artificial waterfall
 OpenCLIP-2B	<ul style="list-style-type: none"> ■ A view looking down at a metal cannon on a wooden stand with wooden wheels ■ A rope is tied to the back of the cannon and tied around a black hook that is on a black metal pole to the left of the cannon ■ Small cannonballs are on three rows of wooden shelves to the left of the cannon, and ten more small cannonballs are hanging from chains attached to a small wooden plank on the gray wall ■ The catapult is resting on a light cream-colored carpet, and the walls are a light gray color ■ A bird's eye view of the front of two wooden catapults sitting on beige carpet ■ Small cannonballs are on three rows of wooden shelves to the left of the cannon, and ten more small cannonballs are hanging from chains attached to a small wooden plank on the gray wall ■ A wooden bucket with a rope handle is to the right of the cannon ■ Another rope is tied to a black hook on the black pole to the right of the cannon ■ A view looking down at a metal cannon on a wooden stand with wooden wheels ■ A rope is tied to the back of the cannon and tied around a black hook that is on a black metal pole to the left of the cannon

Figure 8. Visualization of image-to-text retrieval samples on the DOCCI-FG [36] benchmark, comparing FLAIR with OpenCLIP [6]. For each image, the top-5 retrieved captions are displayed. The incorrect retrieved captions are marked in red. The top-1 retrieved captions are **bold**.

Dataset	#Images	#Captions	#Captions per Image	#Tokens per Caption
Standard Text-image Retrieval Dataset				
MSCOCO [29]	5,000	25,000	5.0	11.77
Flickr30K [39]	1,000	5,000	5.0	14.03
Fine-grained Text-image Retrieval Dataset				
DOCCI-FG [36]	5,000	35,533	7.1	18.76
IIW-FG [19]	612	6204	10.1	22.56
Long Text-image Retrieval Dataset				
DCI [48]	7,805	7,805	1.0	172.73
IIW [19]	612	612	1.0	239.73
SV-1k [5]	1,000	1,000	1.0	173.24
SV-10k [5]	10,000	10,000	1.0	173.66

Table 6. Dataset details of the standard, fine-grained and long image-text retrieval task. SV-1K and SV-10K denote the 1K and 10K subset from the ShareGPT4V [5] dataset. Values of long text-image retrieval are directly obtained from [50], since we follow their evaluation setting.

The notation $\{i, j, l\}$ indicates that this pair is constructed from the $\{\text{Image}, \text{Text Condition}, \text{Text}\}$, which stems from the $\{i\text{-th}, j\text{-th}, l\text{-th}\}$ image separately, while k represents the k -th caption for image i . The pair $\langle \mathbf{v}_{i,i_k}^{\text{tc}}, \mathbf{t}_{i_m}^{\text{g}} \rangle$ is unique, as it arises from the k -th and m -th captions of the same image.

Empirical Comparison. By introducing text-conditioned attention pooling for multi-caption settings, FLAIR consid-

Neg.	$\mathcal{L}_{\text{train}}$	T2I@1	T2I@5	I2T@1	I2T@5
$\langle \mathbf{v}_{i,j_k}^{\text{tc}}, \mathbf{t}_{l_k}^{\text{g}} \rangle$	5.8	0.0	0.1	0.0	0.1
$\langle \mathbf{v}_{i,i_k}^{\text{tc}}, \mathbf{t}_{i_m}^{\text{g}} \rangle$	0.0	0.0	0.1	0.0	0.0
$\langle \mathbf{v}_{i,j_k}^{\text{tc}}, \mathbf{t}_{i_k}^{\text{g}} \rangle$	0.0	0.0	0.1	0.0	0.0
$\langle \mathbf{v}_{i,i_k}^{\text{tc}}, \mathbf{t}_{j_k}^{\text{g}} \rangle$	1.53	2.4	7.8	0.3	1.2
$\langle \mathbf{v}_{i,j_k}^{\text{tc}}, \mathbf{t}_{j_k}^{\text{g}} \rangle$	0.68	24.5	49.1	36.4	62.7

Table 7. Retrieval performance if FLAIR on the MSCOCO [29] validation set when trained with different negative types on the CC3M-recap [59] dataset for 10 epochs. All models use ViT-B/16 as vision encoder. The best retrieval results are **bold**.

ers one positive and up to five distinct negative pairings. Modeling all five negatives simultaneously causes significant computational overhead. Thus, we investigate the importance of each negative type. To study their effects, we conducted a comprehensive ablation experiment (Tab. 7). For each setup, we trained FLAIR with one positive and only one negative pairing at a time, using a batch size of 1,024. All models were trained on the CC3M-recap [59] dataset for 10 epochs.

To evaluate training dynamics, we analyzed the training loss ($\mathcal{L}_{\text{train}}$) and validation performance using the MSCOCO retrieval task. Key findings include: 1. The negative $\langle \mathbf{v}_{i,j_k}^{\text{tc}}, \mathbf{t}_{l_k}^{\text{g}} \rangle$ suffers from high $\mathcal{L}_{\text{train}}$ and poor validation performance. As this negative spans across three different source images, it likely introduces noise rather than aiding learning. 2. The negatives $\langle \mathbf{v}_{i,j_k}^{\text{tc}}, \mathbf{t}_{i_k}^{\text{g}} \rangle$ and $\langle \mathbf{v}_{i,i_k}^{\text{tc}}, \mathbf{t}_{i_m}^{\text{g}} \rangle$ converge quickly during training, but their $\mathcal{L}_{\text{train}}$ swiftly drops to nearly zero. Their evaluation on MSCOCO reveals poor performance, suggesting the existence of shortcuts. For $\langle \mathbf{v}_{i,j_k}^{\text{tc}}, \mathbf{t}_{i_k}^{\text{g}} \rangle$, the model likely ignores image information and relies solely on text conditioning, thus failing in evaluation, when image information is vital. 3. The negative $\langle \mathbf{v}_{i,i_k}^{\text{tc}}, \mathbf{t}_{j_k}^{\text{g}} \rangle$ converges to a reasonable $\mathcal{L}_{\text{train}}$, but its performance (2.4% R@1 on T2I) indicates limited learning benefit. 4. The negative $\langle \mathbf{v}_{i,j_k}^{\text{tc}}, \mathbf{t}_{j_k}^{\text{g}} \rangle$, currently used in FLAIR, reaches the best retrieval results, demonstrating its effectiveness.

D. Additional Ablation Experiments

Aside from the main ablation study on the components of FLAIR described in the main paper, we conduct additional experiments to validate specific design choices. These include pre-training FLAIR on different data sources (Sec. D.1), comparing the diverse sampling strategy with a fixed merging strategy (Sec. D.2), ablating the maximum number of sampled sentences S (Sec. D.3), and examining how the number of sampled captions K affects the performance (Sec. D.4).

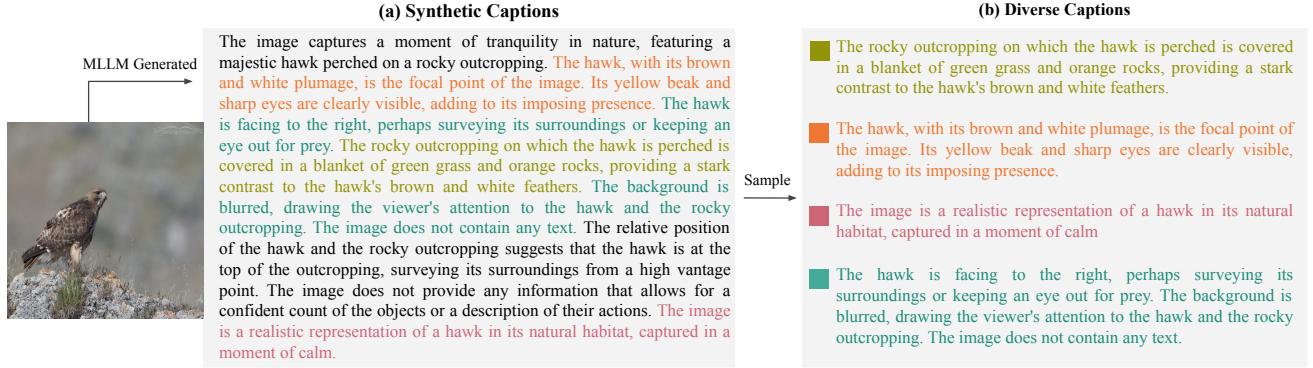


Figure 9. Examples of our diverse captions. Image and captions are taken from CC3M-recap [59]. Given the synthetic long captions generated by an MLLM, here we sample $K = 4$ sub-captions where each sub-caption consists of $s \in \{1, 2, 3\}$ sentences. In our main experiments, we use $K = 8$.

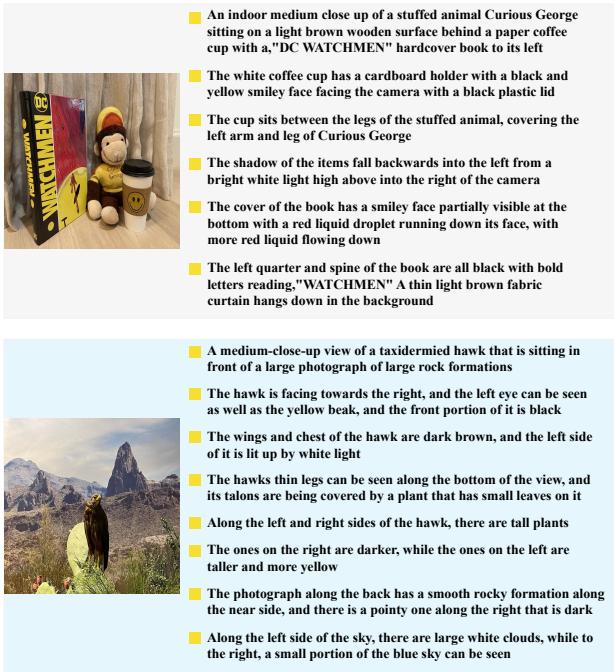


Figure 10. Dataset samples from DOCCI-FG [36] for the fine-grained retrieval task. For each image, we split the long caption into individual sentences each serving as a positive image-text pair for the benchmark.

D.1. Pre-training on Different Data Sources

To demonstrate that our model is not limited to data curated by DreamLIP [59], we also pre-train FLAIR on the original CC3M [44] (CC3M-orig) and PixelProse [45]. For CC3M-orig and PixelProse, we use the same pre-training configurations as CC3M-recap and CC12M-recap, respectively. Detailed configurations are available in Sec. E. CC3M-orig contains one conceptual caption per image, while

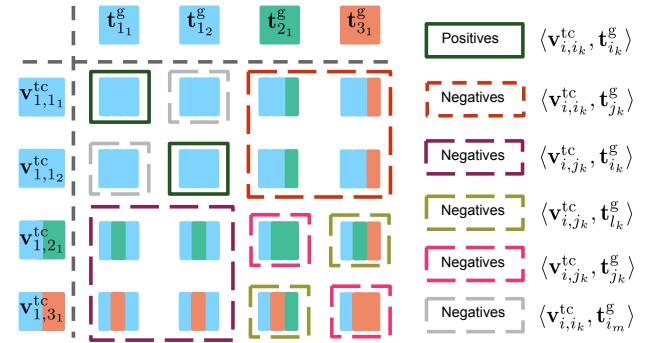


Figure 11. Illustration on all possible positive and negative pairs for FLAIR.

PixelProse re-captioned 15M images from CC12M [4], RedCaps [11], and CommonPool [16] using Gemini-Pro [41]. Unlike DreamLIP, which uses three MLLMs for re-captioning, PixelProse employs a single MLLM, resulting in shorter captions.

We evaluate FLAIR on the standard retrieval task and compare its performance to CLIP [40] trained on the same datasets. The results are summarized in Tab. 8.

As shown in Tab. 8, even when pre-trained on CC3M-orig, where FLAIR cannot leverage additional augmented captions, it still achieves a 2% improvement over CLIP in terms of R@1 on the MSCOCO dataset [29]. This demonstrates that FLAIR is capable of effectively enhancing the retrieval performance even on datasets with only global captions. Furthermore, when pre-trained on PixelProse, FLAIR achieves an 8% improvement in both text-to-image (T2I) and image-to-text (I2T) retrieval tasks on MSCOCO. These results indicate that FLAIR is versatile and can be applied to datasets where images are captioned by a different MLLM, while maintaining significant performance gains.

Data	Method	MSCOCO				Flickr30k			
		T2I		I2T		T2I		I2T	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CC3M-orig [44]	CLIP [40]	4.75	14.53	5.90	17.56	9.19	23.61	12.13	29.68
	FLAIR	6.45	18.14	8.00	22.48	12.70	30.20	17.55	38.66
PixelProse [45]	CLIP [40]	28.86	54.05	48.50	74.24	54.06	77.81	79.09	94.87
	FLAIR	36.08	61.18	56.56	79.06	64.87	85.03	86.69	97.14

Table 8. Standard zero-shot image-text retrieval on the validation splits of Flickr30k [39] and MSCOCO [29]. CLIP and FLAIR are pre-trained on CC3M-orig and PixelProse under the same training configurations, using ViT-B/16 as the vision encoder.

Merging	MSCOCO		DOCCI		Urban-1K		VOC20	ImageNet
	T2I@1	I2T@1	T2I@1	I2T@1	T2I@1	I2T@1	mIOU	Top-1
	No	35.8	47.1	14.8	33.2	46.4	42.4	52.2
Always	34.2	46.8	12.4	30.1	70.9	64.7	54.9	27.7
Random	37.7	51.6	15.1	35.7	69.5	63.5	59.7	33.8

Table 9. Ablation study on merging strategies for sampling captions. No: only sample 1 sentence as the sampled caption. Always: always merge 3 sampled sentences into one caption. Random: each caption is merged randomly from 1-3 sentences. We train FLAIR on CC3M-recap with 8 captions per image.

S	MSCOCO		DOCCI		Urban-1K		VOC20	ImageNet
	T2I@1	I2T@1	T2I@1	I2T@1	T2I@1	I2T@1	mIOU	Top-1
2	37.1	50.7	14.2	35.2	68.5	62.8	59.0	32.0
3	37.7	51.6	15.1	35.7	69.5	63.5	59.7	33.8
4	37.5	52.0	14.6	35.2	69.5	63.2	57.4	33.0

Table 10. Ablation on the maximum number of sentences (S) to be merged for create a new sub-caption. We trained FLAIR with $S \in [2, 4]$ on CC3M-recap dataset under the same training configuration. The best results are **bold**.

D.2. Sampling Strategy

When sampling diverse captions, we randomly merge $s \in \{1, \dots, S\}$ sentences in the original MLLM-generated long captions to form a single caption. To evaluate this strategy, we compare FLAIR with three settings: randomly merging 1–3 sentences, always merging 3 sentences, and no merging. The results, presented in Tab. 9, show that always merging 3 sentences improves Urban-1K T2I R@1 and I2T R@1 by 1.4% and 1.2%, respectively. However, it decreases T2I R@1 and I2T R@1 on MSCOCO by 3.5% and 5.2%, indicating a bias towards long retrieval tasks at the expense of short retrieval performance.

Conversely, random merging outperforms the no-merging setting across all metrics, effectively balancing short and long retrieval tasks. Additionally, it enhances model performance by introducing diverse data augmentations through caption variations.

K	MSCOCO		DOCCI		Urban-1K		VOC20	ImageNet
	T2I@1	I2T@1	T2I@1	I2T@1	T2I@1	I2T@1	mIOU	Top-1
CLIP [6]	27.0	38.9	10.3	25.0	41.3	37.7	3.16	23.8
SigLIP [57]	28.3	40.1	10.4	24.9	42.8	40.5	3.1	25.4
2	36.4	49.1	13.9	35.5	68.7	62.9	56.3	31.2
4	36.7	49.8	14.2	35.4	69.1	62.7	57.4	32.8
6	37.4	51.2	14.9	35.4	69.8	61.4	59.5	33.6
8	37.7	51.6	15.1	35.7	69.5	63.5	59.7	33.8
10	37.8	51.7	15.0	35.1	71.6	64.2	60.9	33.6

Table 11. Ablation results on the number of sub-captions K for FLAIR. OpenCLIP [6], SigLIP [57] and FLAIR are pre-trained on CC3M-recap datasets under the same configuration. All models use ViT-B/16 as vision encoder. The best results are **bold**.

D.3. Number of Merged Sentences

In the diverse caption sampling strategy, each new caption is created by merging up to S sentences. In Tab. 10, we train FLAIR with $S = 2$, $S = 3$, and $S = 4$. Compared to $S = 2$, $S = 3$ yields consistent improvements across all downstream tasks. However, increasing to $S = 4$ does not lead to further gains, likely because merging four sentences often exceeds the 77-token limit of the text encoder. Based on these findings, we set $S = 3$ for our main experiments.

D.4. Number of Sampled Sub-captions

In Tab. 11, we pre-train FLAIR with a different number of sampled captions K ranging from 2 to 10 on the CC3M-recap dataset. We also compared CLIP [40] and SigLIP [57] pre-trained on the same dataset. First, even when $K = 2$, FLAIR surpasses SigLIP by 8.1% (T2I R@1) and 9.0% (I2T R@1) on MSCOCO retrieval. Increasing to $K = 8$ further brings 1.3% and 2.5% increase in T2I R@1 and I2T R@1 on MSCOCO. Generally, we notice that the performance converges when $K \in (6, 10)$. However, increasing K introduces extra computation overhead, since the text encoder process K captions in every iteration. Therefore, we choose $K = 8$ as our main setting, as it achieves a good balance between optimal performance and computation.

E. Implementation Details

In this section, we describe the detailed implementation of pre-training and downstream tasks evaluation.

Pre-training. Our implementation is based on the OpenCLIP [6] code base with the ViT-B/16 architecture for the image encoder. Both image and text encoder consist of 12 transformer layers, and the embedding size is fixed at 512. Specifically for FLAIR, we replace the final pooling layer of image encoder with our text-conditioned attention pooling, while the rest of the layers remain unchanged. Our loss function initializes t at 0.07 and b at -10, consistent with the settings used in SigLIP. We follow DreamLIP’s pre-training configuration as displayed in Tab. 12. However,

Config	CC3M-recap	CC12M-recap	YFCC15M-recap	Merged-30M
Batch size	1,024	6,134	6,134	6,134
Optimizer		AdamW [31]		
Learning rate		5×10^{-4}		
Weight decay	0.5	0.5	0.5	0.2
Adam β		$\beta_1, \beta_2 = (0.9, 0.98)$		
Adam ϵ	1×10^{-8}	1×10^{-8}	1×10^{-8}	1×10^{-6}
Total epochs		32		
Warm up		2,000(steps)		
LR scheduler		cosine decay		

Table 12. Pre-training hyper-parameters for FLAIR and all re-trained baseline methods. LR scheduler: Learning Rate scheduler.

Method	Data Size	VOC20	Cityscapes	Context59	ADE20K	COCO-Stuff	Average
CLIP [40]	400M	41.8	5.5	9.2	3.2	4.4	12.8
OpenCLIP [6]	2B	47.2	5.1	9.0	2.9	5.0	13.9
MetaCLIP [52]	2.5B	35.4	5.0	8.1	2.2	4.3	11.0
FLAIR-CLIP	3M	60.9	8.9	15.6	8.0	9.7	20.6
FLAIR-TC		53.9	20.6	23.8	13.1	13.1	24.9
FLAIR-CLIP	12M	69.7	14.5	17.4	10.0	12.2	24.8
FLAIR-TC		55.1	20.1	22.9	13.3	15.4	25.4
FLAIR-CLIP	15M	71.5	13.3	18.4	9.0	12.5	24.9
FLAIR-TC		49.2	16.5	17.4	9.1	13.6	21.2
FLAIR-CLIP	30M	73.0	13.6	18.6	10.4	13.3	25.8
FLAIR-TC		48.3	13.6	17.4	10.8	14.4	20.9

Table 13. Mean Intersection over Union (mIoU) for zero-shot semantic segmentation on VOC20 [13], Cityscapes [8], Context59 [33], ADE20K [60], and COCO-Stuff [2]. All models employ ViT-B/16 as the vision encoder. The best results are **bold**.

we use 6K batch size for CC12M-recap, YFCC15M-recap and Merged30M due to GPU RAM limit. Experiments on CC3M-recap are trained on 8 NVIDIA A100 40GB GPUs and 32 GPUs on the other datasets. All baseline models, CLIP and SigLIP, follow the same pre-training configurations.

Large-scale Pre-trained CLIP Models. In the main paper, we report the values for OpenCLIP (2B) and SigLIP (10B). Both models employ ViT-B/16 as the vision encoder. Those values were obtained by evaluating the pre-trained weights of OpenCLIP. “OpenCLIP (2B)” refers to the ViT-B/16 model trained on the LAION-2B dataset with the pre-trained name of “laion2b_s34b_b88k”. “SigLIP (10B)” refers to the ViT-B/16-SigLIP model trained on the WebLI dataset with the pre-trained name of “webli”. The Llip [26] and MetaCLIP [52] results for zero-shot image classification are directly obtained from their papers.

Zero-shot Semantic Segmentation. As discussed in Sec. 4 of the main paper, zero-shot semantic segmentation

is based on the similarity between local image tokens and global text queries $\{\langle \mathbf{v}_i^{\text{loc}}, \mathbf{t}_j^g \rangle \mid j \in \{1, 2, \dots, M\}\}$, where M represents the number of class names in the dataset. Compared to CLIP, a key advantage of FLAIR is its flexibility during inference: it can either directly compute $\langle \mathbf{v}_i^{\text{loc}}, \mathbf{t}_j^g \rangle$ without applying $f_{\text{AttnPool}}(\cdot)$ (FLAIR-CLIP), or first use $f_{\text{AttnPool}}(\mathbf{v}_i^{\text{loc}}, \mathbf{t}_j^g)$ to generate fine-grained embeddings $\mathbf{v}_{i,j}^{\text{tc}}$, and then compute $\langle \mathbf{v}_{i,j}^{\text{tc}}, \mathbf{t}_j^g \rangle$ (FLAIR-TC). Segmentation results for both approaches are reported in Tab. 13. For implementation details, including window size, stride, and other parameters, we follow the design choices described in [49].

Interestingly, using the CLIP method increases mIoU on VOC20 by approximately 10%, while the TC method improves performance on other datasets. Both methods outperform OpenCLIP and SigLIP models trained on billions of images. This indicates that the segmentation capability of FLAIR is not solely reliant on the attention pooling layer, because the local image tokens \mathbf{v}_{loc} encode strong localization information independently.

Zero-shot Image Classification. We follow the prompt ensemble strategy described in LaCLIP [14] and ALIP [53], employing the same prompt templates. For each class name, we compute the average text embedding across all templates, which is then used to calculate the similarity between test images and class embeddings. For zero-shot ImageNet classification, we use the seven prompt templates recommended by [40], consistent with LaCLIP [14].