This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Revisiting Multimodal Representation in Contrastive Learning: From Patch and Token Embeddings to Finite Discrete Tokens

Yuxiao Chen[1*], Jianbo Yuan[2], Yu Tian[2], Shijie Geng[1,2], Xinyu Li[2],
Ding Zhou[2], Dimitris N. Metaxas[1†], Hongxia Yang[3]
[1]Rutgers University [2]ByteDance Inc. [3]Zhejiang University
{yc984, sg1309, dnm}@rutgers.edu,
{jianbo.yuan, yutian.yt, lixinyu.arthur, ding.zhou}@bytedance.com
hongxia.yang1@gmail.com

Abstract

Contrastive learning-based vision-language pre-training approaches, such as CLIP, have demonstrated great success in many vision-language tasks. These methods achieve cross-modal alignment by encoding a matched image-text pair with similar feature embeddings, which are generated by aggregating information from visual patches and language tokens. However, direct aligning cross-modal information using such representations is challenging, as visual patches and text tokens differ in semantic levels and granularities. To alleviate this issue, we propose a Finite Discrete Tokens (FDT) based multimodal representation. FDT is a set of learnable tokens representing certain visual-semantic concepts. Both images and texts are embedded using shared FDT by first grounding multimodal inputs to FDT space and then aggregating the activated FDT representations. The matched visual and semantic concepts are enforced to be represented by the same set of discrete tokens by a sparse activation constraint. As a result, the granularity gap between the two modalities is reduced. Through both quantitative and qualitative analyses, we demonstrate that using FDT representations in CLIP-style models improves cross-modal alignment and performance in visual recognition and vision-language downstream tasks. Furthermore, we show that our method can learn more comprehensive representations, and the learned FDT capture meaningful cross-modal correspondence, ranging from objects to actions and attributes.[1]

1. Introduction

Recently, the Contrastive Language-Image Pre-training (CLIP) framework [16, 27] has demonstrated notable capabilities for learning powerful and transferable feature representations [10, 22, 40–43]. In this framework, models are trained to align text and image information in a two-stream approach where image and text representations are extracted through two separate encoders. The InfoNCE loss [27] is used to train the encoders which enforces the representations of matched image-text pairs to be closer, while those of unmatched pairs to be far apart (as shown in Figure 1 (Left)).

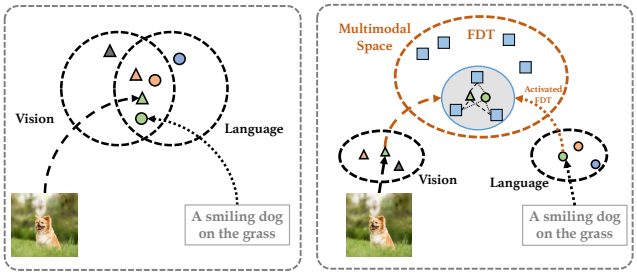

Figure 1. Comparison of different feature representation learning methods. **Left**: contrastive vision-language pre-training (CLIP). **Right**: CLIP with our proposed finite discrete tokens (FDT).

However, the fact that the information conveyed in images and text captions is naturally of different levels of granularities [29, 34] is not considered by such models. For example, an image of a dog also portrays various lower-level attributes, such as its breed, fur color, body size, and shape, while the textual description, such as "a smiling dog", is generally more abstract and compact. In CLIP, images and text captions are represented through the aggregation of visual patches and text tokens without explicitly aligning the visual and semantic concepts at the same level of granularity. It can cause challenges in multimodal representation learning, or even potentially result in performance degradation [35]. Additionally, the learned models may overlook certain semantic concepts [14]. Therefore, we argue that

*This work was done during a research internship at ByteDance.

†Dimitris N. Metaxas has been supported by NSF IUCRC CARTA-1747778, 2235405, 2212301, 1951890, 2003874.

[1]The source code can be found at https://github.com/yuxiaochen1103/FDT.

unifying the information granularities of images and texts can help generate better multimodal representations.

In this paper, we propose a new **F**inite **D**iscrete **T**okens (FDT) based representations. FDT is a set of *learnable tokens* that encode cross-modal shared semantic concepts. Both image and text are represented as the combinations of FDT shared between modalities so that the information granularities are unified (see Figure 1 (Right)). Figure 2 gives an overview of our method. For an image, its patch embeddings are first extracted by an image encoder. The correspondence between the FDT and the image is then measured by max pooling over the attention weights of FDT among all patches. Finally, the FDT-based representation of the image is calculated as the attention-weighted sum of FDT. The FDT-based embeddings for input texts can be constructed in the same way. The encoders and FDT are trained to pull close the FDT-based representations of matched image-text pairs while pushing away those of unmatched pairs by using the InfoNCE loss. To the point of leveraging a shared FDT across modalities is to enforce the matched visual and semantic concepts to be represented by the same discrete tokens. For example, the visual patches of a dog and the word "dog" should activate the same subsets of FDT. We empirically demonstrate that this can be achieved by simply enforcing relatively sparse attention-weights between FDT and the inputs.

We conduct extensive experiments covering a wide range of pre-training settings and downstream tasks to evaluate the proposed method. We conclude with the following key observations: (1) Our approach exhibits consistent performance enhancements across various pre-training dataset scales, CLIP-based pre-training frameworks [20], and encoder architectures. Notably, our method outperforms CLIP by 5.0% on zero-shot image classification when pre-training on 145M datasets, and by 33.4% in image-text retrieval with 30M datasets; (2) Our method tends to alleviate the model degradation problem and learns more comprehensive feature representations than CLIP; (3) The learned FDT exhibit better: we visualize FDT's correspondent patches and language tokens, and the results show that FDT successfully capture and align visual-semantic concepts including objects, attributes, and actions.

2. Related Work

Vision and Language Pre-training. Vision and language pre-training methods can be briefly classified into two-stream and single-stream models based on their architectures. A typical two-stream model leverages individual encoders to extract continuous feature embeddings from the inputs, and enforces the embeddings of a matched image-text pair to be similar by using contrastive learning [12, 16, 27] and additional self-supervised tasks [20, 36]. Inherited from the encoder design, these feature embeddings convey information aggregated from local vision patches and language tokens, which encompass different semantic levels and granularities and are constrained by how patches are generated. Therefore, we propose FDT-based representations to directly perform contrastive learning on FDT that denotes high-level vision-semantic concepts. The single-stream approaches feed all inputs together into a unified encoder (mostly transformers) to enhance the cross-modal interactions for a better cross-modal alignment [4, 5, 18, 19, 31, 36]. For simplicity, we also clarify models consisting of individual encoders followed by multimodal fusion operations (late-fusion) as one-stream, because it requires the inputs from all modalities for inference and hence does not support ANN, similar to a typical one-stream model (early-fusion). To combine the best of both worlds, FDT-based representations bridge the gap between different modalities with cross-modal interactions by vision-to-token and language-to-token information exchange, while maintaining a two-stream structure.

Vector-Quantization and Codebook. Vector-quantization is first proposed for image generation showing that image information can be encoded by discrete representations (namely *codebook*) [32]. Each image patch is represented by its nearest-neighbor code's embedding, and the decoder reconstructs the input image based on these code embeddings. Because finding *nearest-neighbor* is non-differentiable, the codebook is trained either by minimizing the distance between the code and image patch embeddings when the encoder is stop-gradient, or by exponential moving averages (EMA). Applying VQ to multimodal pre-training is more challenging, as the codebook now needs to accommodate multimodal contents and is often found to be sensitive to initialization (cold-start problem). To address these challenges, previous studies leverage encoder or code warm-up [23], knowledge distilled vision tokenizers from pre-trained vision-language models [26], one-stream models to enforce multimodal code learning [15, 19], and a combination of these techniques [33]. As a comparison, our approach is designed to be more intuitive where only differentiable operations are used and it can be trained end2end from scratch while still maintaining a two-stream structure for ANN in large-scale retrieval tasks. More technical details will be discussed in Section 3.2.

Dictionary Learning. Dictionary learning is another group of discrete representation learning in addition to VQ [2, 9, 11]. Given a dictionary matrix [11], the representation of a signal is the weights that can linearly combine the dictionary matrix to reconstruct the signal with minimal error. When learning multi-modal representations [2, 9], a shared dictionary matrix is used for facilitating cross-modal information alignment and fusion. The dictionary is served as the cross-modal information anchor, which shares the same idea as ours. However, the models are trained to solve a

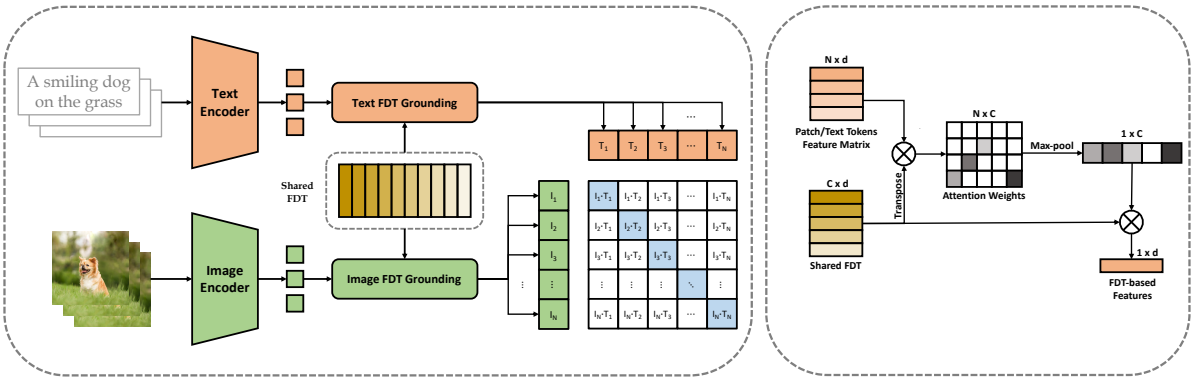

Figure 2. **Left:** Overview of the proposed method. Both the image and text information is encoded with shared FDT during cross-modal contrastive pre-training. **Right:** The process of grounding image or text features to FDT. The attention weights between visual patch/language token and FDT are first calculated, and then max-pooled over all visual patches/language tokens. The attention-weighted sum of FDT is calculated as the FDT-based features.

slow optimization problem, and the feature learned by solving the reconstruction or generative problem may have limited discriminative capability. By contrast, our model is trained end-to-end to learn discriminative information.

3. Method

3.1. Revisiting Feature Representations in CLIP

In CLIP, the image and text features are the aggregation of the embeddings of image patches or language tokens, respectively. Specifically, the image encoder takes an image as input and extracts the patch or local region embeddings based on the self-attention [8], or convolution operations [13]. The obtained patch features are then aggregated as the final representation of the image f_v by using the attention pooling or the [CLS] token [8, 27], which can be formulated as:

$$w_{p_i} = \frac{e^{<f_g, f_{p_i}>}}{\sum_j^{N_v} e^{<f_g, f_{p_j}>}}, \tag{1}$$

$$f_v = \sum_i^{N_v} (w_{p_i} \cdot f_{p_i}). \tag{2}$$

Here, w_{p_i} is the weight of i-th patch, which measures the importance of the patch to the final representation. $<,>$ is the inner-product function. N_v is the number of patches, and f_{p_i} denotes the embedding of i-th patch. f_g is the [CLS] token embedding or the average-pooled patch embedding, which embeds the global image information.

Similarly, for the text encoder, the extracted text representation of an input sentence can also be regarded as the weighted sum of language token embeddings:

$$f_t = \sum_i^{N_t} (w_{t_i} \cdot f_{t_i}), \tag{3}$$

where N_t is the number of language tokens. f_{t_i} is the embedding of the i-th language token. It is extracted with the self-attention operations [7, 28], which model the relationship among the language tokens. w_{t_i} is the weight of the i-th language token, which is calculated by the following Equation 1 using the text [CLS] token.

Equations 1 and 3 suggest that images or texts are represented by two different bases: visual patches and language tokens. However, the information conveyed by image patches and language tokens may have different semantic meanings and granularities. Additionally, the bases are dynamic, since the visual patches or language tokens of different images or texts are different. It may increase the difficulty of learning an optimal alignment between image and text features [14, 35]. Thus, the encoders may fail to capture important semantic concepts shared in both modalities and may encode irrelevant information.

3.2. FDT-based Representation

To address the aforementioned limitations of feature representation in CLIP, we propose the FDT-based representation. Figure 2 gives an overview of our proposed method. Instead of representing the image and text with different bases, FDT serve as the common bases for both the image and text representations. As a result, the granularities of cross-modal information are explicitly unified. Moreover, the FDT encode the semantic information shared by both modalities. It can be regarded as prior knowledge that guides image and text encoders to extract feature embeddings. In the following, we elaborate on the steps necessary to achieve FDT-based representations:

Grounding to FDT. Let $\{c_i | i = 1, ..., C\}$ be FDT, where C is the number of shared tokens, and c_i is the i-th discrete token. Given an input image, its patch embeddings are first extracted using the image encoder. The extracted patch em-

beddings are then projected to the FDT space by using a projecting function. The relevance between the image and a token is obtained by calculating the inner product between the projected patch embeddings and the token, and selecting the maximal value, which can be formulated as

$$r_i^v = \max_j < f_{p_j}, c_i >, \tag{4}$$

where r_i^v is the relevance between the image and the i-th tokens. Intuitively, the proposed patch-level relevance calculation mechanism may enjoy two advantages: (1) it can capture small objects that exist in a single patch; (2) it helps remove the influence of irrelevant noisy patches that have low relevance to all FDT.

The relevance between the image and FDT is normalized by a Softmax function, which generates the final weights of each token as follows:

$$w_i^v = \frac{e^{r_i^v}}{\sum_j^C e^{r_j^v}}, \tag{5}$$

where w_i^v is the weight of the i-th token with respect to the image. Similarly, the weight w_i^t of the i-th token assigned by an input text can be calculated using

$$r_i^t = \max_j < f_{t_j}, c_i >, \tag{6}$$

$$w_i^t = \frac{e^{r_i^t}}{\sum_j^C e^{r_j^t}}. \tag{7}$$

Intuitively, FDT can be treated as prior knowledge for the image or text information. With the help of FDT, the extracted features of both modalities are grounded to a shared manifold space, thus enabling the cross-modal interaction.

Normalizing Concept Weights with Sparse Constraints. We expect the normalized weights of FDT to be sparse, since it can largely reduce noise and make the results more interpretable [2, 11]. Additionally, we empirically show that sparsity is crucial for FDT to learn cross-modal correspondence, where a token corresponds to the same image and text semantic meaning. We use the Sparsemax function [24] for sparser weights, which is defined as:

$$\arg\min_{\boldsymbol{p} \in \Delta^{K-1}} \|\boldsymbol{p} - \boldsymbol{r}\|^2, \tag{8}$$

where $\boldsymbol{r}$ is the vector consisting of the relevance score between the image or text and FDT (Equation 4 and 6). This function first calculates a threshold, and then sets the weights below the threshold to zero for sparsity. In contrast, the commonly used Softmax function cannot explicitly assign FDT with exactly zero probabilities.

Generating FDT-based Embeddings. Given the normalized weights, the FDT-based features of the image f_v^{FDT} and text f_t^{FDT} are the weighted sum of FDT:

$$f_v^{\mathrm{FDT}} = \sum_i^C w_i^v \cdot c_i \tag{9}$$

$$f_t^{\mathrm{FDT}} = \sum_i^C w_i^t \cdot c_i \tag{10}$$

Equations 9 and 10 show that image and text features are represented by the same base FDT, which explicitly unifies the granularities of image and text information.

Given the FTD-based features, the encoders and FDT are trained to make the similarity between FDT-based features of matched image-text pairs larger than those of unmatched pairs:

$$\mathcal{L} = -\frac{1}{N}\sum_i^N \log \frac{\exp\left(\mathrm{sim}\left(f_{v_i}^{\mathrm{FDT}}, f_{t_i}^{\mathrm{FDT}}\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\mathrm{sim}\left(f_{v_i}^{\mathrm{FDT}}, f_{t_j}^{\mathrm{FDT}}\right)/\tau\right)} - \frac{1}{N}\sum_i^N \log \frac{\exp\left(\mathrm{sim}\left(f_{t_i}^{\mathrm{FDT}}, f_{v_i}^{\mathrm{FDT}}\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\mathrm{sim}\left(f_{t_i}^{\mathrm{FDT}}, f_{v_j}^{\mathrm{FDT}}\right)/\tau\right)}, \tag{11}$$

where N is the number of matched image-text pairs, sim is the cosine similarity function, and τ is the temperature hyper-parameter.

Intuitively, the equation shows that FDT are updated based on both the image and text modalities, and thus FDT is trained to learn the information shared by both modalities.

4. Experiments

4.1. Experimental Settings

Pre-training Datasets. We use four publicly available datasets, including YFCC-15M V2 [6], Conceptual Captions (CC3M) [30], Conceptual 12M (CC12M) [3] and LAION115M [17] datasets to pre-train our models. We construct three different pre-training settings, including **15M**, **30M**, and **145M** settings. Each of the settings uses different combinations of pre-training datasets, as shown in Table. The 15M setting is used for the ablation study and to compare our methods with state-of-the-art methods under a fair setup [6]. The 30M and 145M settings are used to evaluate the scalability of our model.

Setting	Dataset
15M	YFCC-15M V2
30M	YFCC-15M V2, CC3M, CC12M
145M	YFCC-15M V2, CC3M, CC12M, LAION115M

Table 1. The used pre-training datasets under different settings.

Evaluation Protocols. Following previous work [6, 20, 37], our method is evaluated on three commonly-used downstream tasks, including zero-shot image classification, linear probe image classification, and zero-shot image-text

	C10	C100	F101	PETS	FLOW	SUN	DTD	CAL	IN	AVG
SLIP [25]	50.7	25.5	33.3	23.5	49.0	34.7	14.4	59.9	34.3	36.1
MS-CLIP-S [38]	-	-	-	-	-	-	-	-	36.7	-
CLIP [27]	60.4	33.5	39.6	23.1	54.0	42.0	17.0	65.5	37.0	41.3
FILIP [37]	65.1	34.2	43.2	24.1	52.8	50.8	24	68.9	39.5	44.7
DeCLIP [20]	72.8	40.3	49.9	36.2	60.1	48.8	26.4	72.7	43.2	50.0
CLIP+FDT (Ours)	67.7	39.9	42.9	25.8	55.5	45.5	26.5	69.6	39.3	45.9
DeCLIP+FDT (Ours)	**75.7**	**45.2**	**52.9**	**40.7**	**64.6**	**52.0**	**30.7**	**76.2**	**45.8**	**53.8**

Table 2. Zero-shot image classification accuracy (%) under the 15M setting. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. IN is ImageNet-1K. "AVG" is the average accuracy over all datasets.

	C10	C100	F101	PETS	FLOW	SUN	CARS	DTD	CAL	AIR	AVG
SLIP [25]	87.4	69.5	71.3	70.5	91.9	66.9	27.5	65.6	86.2	27.7	66.5
MS-CLIP-S [38]	87.2	66.7	76.0	62.1	93.8	71.7	27.5	69.4	81.6	**32.9**	66.9
CLIP [27]	88.3	68.6	72.1	72.5	92.6	69.5	29.8	67.8	86.2	27.7	67.5
FILIP [37]	86.5	66.6	71.7	69.2	93	69.6	30.0	66.4	85.7	27.0	66.6
DeCLIP [20]	89.4	69.6	75.9	71.4	95.7	71.6	30.1	66.9	89.0	26.7	68.6
CLIP+FDT (Ours)	89.1	**71.2**	74.4	73.0	93.4	70.8	31.4	69.4	87.7	27.9	68.8
DeCLIP+FDT (Ours)	**89.8**	**71.2**	**77.7**	**73.9**	**95.7**	**72.9**	**33.7**	**69.6**	**89.4**	26.9	**70.1**

Table 3. Linear probing image classification accuracy (%) under the 15M setting. The dataset names are abbreviated. C10/100 is CIFAR10/100. F101 is Food101. FLOW is Flowers. CAL is Caltech. Air is Aircraft. "AVG" is the average accuracy over all datasets.

	Flickr30K				MSCOCO				VQAv2			
	Image Retrieval		Text Retrieval		Image Retrieval		Text Retrieval					
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	y/n	number	other	overall
SLIP [25]	23.3	47.2	35.7	65.8	13.2	31.3	21.0	44.6	69.8	34.3	38.1	50.7
MS-CLIP-S [38]	-	-	-	-	19.4	40.8	28.5	54.1	-	-	-	-
CLIP [27]	27.6	53.9	42.8	71.5	15.9	36.7	24.8	49.8	67.7	31.9	33.6	47.5
FILIP [37]	30.6	58.2	46.3	74.4	16.2	37.5	25.6	50.8	68.1	34.5	36.2	49.2
DeCLIP [20]	35.5	63.0	51.2	80.7	19.6	41.9	30.1	55.6	**70.3**	34.9	36.9	50.4
CLIP+FDT (Ours)	32.6	58.6	51.0	78.3	19.4	40.8	29.6	55.3	67.8	34.6	39.6	50.6
DECLIP+FDT (Ours)	**39.4**	**66.8**	**57.0**	**82.3**	**22.5**	**45.5**	**34.0**	**59.6**	67.8	**35.8**	**41.3**	**51.6**

Table 4. Results of the vision-language tasks under the 15M setting, including the zero-shot image-text retrieval on the Flickr30K and MSCOCO (5K) datasets, and the non-linear probing on VQA v2 dataset.

retrieval. Moreover, we propose a non-linear probe task to evaluate the effectiveness of the learned features for VQA [1]. The FDT-based features are used for all the downstream tasks.

Zero-shot image classification. In this task, image categories are represented by the text descriptions generated from their names. After extracting the embeddings of these text descriptions and input images by pre-trained encoders, the category of an image can be predicted by choosing the one whose text descriptions have the largest cosine similarity score. Following the setting of CLIP and DeCLIP, we construct 80 prompts to evaluate the performance of different approaches. We use 9 of the 11 commonly used datasets [20] for evaluation. The StanfordCars and Aircraft datasets are not used, because the pre-training datasets contain few captions about car models or aircraft types.

Linear Probe Image Classification. A linear classifier is trained to predict the categories of images based on the FDT-based features of the images. We use 10 of the 11 commonly used datasets for evaluation. We do not report the results on ImageNet-1K, since conducting hyperparameter sweeping on this dataset is computationally expensive.

Image-text retrieval. The image-text retrieval task is evaluated on the Flickr30K [39] and MSCOCO [21] dataset. The recalls at different K values (R@K, K = 1, 5, 10) are reported as the evaluation metrics. They are used to measure the percentage of relevant items that match the queries in top-K retrieved items. We also report rsum, which is obtained by summing all R@K values.

Non-linear probe task. The task is to evaluate the capability of learned features for vision-language reasoning tasks. The FDT-based embeddings of an image and its questions are concatenated and fed to two fully-connected layers with non-linear activation to predict the answer. More details can be found in the supplementary materials.

Implementation Details. We evaluate our method by incorporating it into two state-of-the-art contrastive vision-language pre-training approaches, namely CLIP [27] and DECLIP [20]. Our implementation is based on the open-

	Setting	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 overall
CLIP	15M	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP+FDT	15M	45.9(↑4.6)	68.8(↑1.3)	32.6(↑5.0)	51.0(↑8.2)	376.5(↑33.4)	19.4(↑3.5)	29.6(↑4.8)	263.1(↑26.3)	50.6(↑3.1)
CLIP	30M	56.8	73.8	43.6	58.8	431.3	23.3	34.8	300.8	50.6
CLIP+FDT	30M	61.2(↑ 4.4)	75.6 (↑ 1.8)	52.5(↑8.9)	70.8(↑12.0)	474.2(↑42.9)	28.3(↑5.0)	43(↑8.2)	337.1 (↑36.3)	53.4(↑2.8)
CLIP	145M	64	82.1	52.6	67.9	469.8	29.3	42.1	335.2	53.1
CLIP+FDT	145M	69.0(↑ 5.0)	82.3 (↑ 0.2)	56.3(↑3.7)	75.9(↑8.0)	489.4(↑19.6)	31.0(↑1.7)	46.4(↑4.3)	353.0(↑17.8)	55.2(↑2.1)

Table 5. Ablation study results when using different scales of training data. "ZS" means zero-shot. "AVG" is average. "ACC" is accuracy. "LP" stands for linear prob. "CLS" represents classification. "IR" and "TR" are image retrieval and text retrieval, respectively.

	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 Overall
CLIP-ViT-B/32	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP-ViT-B/32+FDT	45.9(↑4.6)	68.8(↑1.3)	32.6(↑5.0)	51.0(↑8.2)	376.5(↑33.4)	19.4(↑3.5)	29.6(↑4.8)	263.1(↑26.3)	50.6(↑3.1)
CLIP-ViT-B/16	45.2	68.8	35.3	50.5	387.8	19.3	29.7	263.6	49.2
CLIP-ViT-B/16+FDT	49.9(↑4.7)	71.3(↑2.5)	41.6(↑6.3)	60.8(↑10.3)	425.5(↑37.7)	23.4(↑4.1)	35.3(↑5.6)	295.4 (↑31.8)	54.3(↑5.1)
CLIP-Swin-B	39.6	68.5	30.5	48.5	368.1	17.7	26.0	247.6	46.5
CLIP-Swin-B+FDT	42.4(↑2.8)	70.7(↑2.2)	39.6(↑9.1)	57.9(↑9.4)	415.5(↑47.4)	22.3(↑4.6)	33.8(↑7.8)	288.3(↑40.7)	51.6(↑5.1)

Table 6. Ablation Study results when using different image encoder architectures. "ZS" means zero-shot. "AVG" is average. "ACC" is accuracy. "LP" stands for linear prob. "CLS" represents classification. "IR" and "TR" are image retrieval and text retrieval.

source PyTorch implementation[2] of the two methods. We use 16384 tokens, each with 512 dimensions. Please refer to the supplementary material for detailed information.

4.2. Comparison with State-of-the-Art Approaches

We compare our method with the state-of-the-art CLIP family approaches on the benchmark proposed in [6]. In this benchmark, methods are compared fairly by pre-training them using the same training recipe and data (our 15M setting). Note that the original paper only reports the results for zero-shot classification on the ImageNet dataset, and the results of other tasks are obtained by directly applying the released checkpoints for evaluation.

The results for zero-shot image classification, linear prob image classification, and vision-language reasoning tasks are reported in Table 2, 3, and 4, respectively. First, we observe that using the proposed FDT-based representation with CLIP (i.e., CLIP+FDT) can achieve significant performance improvement over CLIP on all the downstream tasks. Notably, CLIP+FDT can outperform FILIP [37], which aligns image and text information at the fine-grained patch and language token levels. The results suggest that aligning global cross-modal information in a unified space is more effective than directly aligning fine-grained patches and language tokens with different granularities. Interestingly, the linear probe results show that CLIP+FDT can learn a comparable image encoder with DeCLIP, which applies various self-supervised pretext tasks that have already been proven effective for visual recognition. One possible reason is that aligning the information in a unified space helps our model better leverage semantic supervision signals in the language domain. We can also see that our method can significantly improve DeCLIP for all the tasks and achieve state-of-the-art performance on the benchmark. It shows that our approach is compatible with self-supervised learning tasks to improve CLIP. Moreover, FDT can improve the VQAv2 task, which requires the capability of collaborative multi-modal reasoning and content understanding.

[2]https://github.com/Sense-GVT/DeCLIP

4.3. Ablation Study

In this section, we conduct ablation studies to investigate how different factors influence the performance of our approach. These factors include the pre-training data scale, image encoder architecture, and several design choices of our method. Throughout the ablation study, we use the CLIP model as the baseline to save computation costs.

Pretraining Data Scale. We evaluate the performance of our methods on different pre-training data scales by further pre-training the model on 30M and 145M data. According to the results presented in Table 5, our method still achieves improved performance for all the downstream tasks when pre-trained on larger datasets. We also note that the improvement for the linear probing setting is minor when pre-trained on 145M data. We assume this is because the performance of the model saturates. To further improve the performance of the image encoder, a more vision-specific training task is needed. Note that using FDT still achieves significant performance improvements on 145M data for other tasks. Interestingly, our model achieves significant improvements on the 30M data. One possible reason is that our FDT can benefit significantly from cleaning supervision information in the CC3M [30] and CC12M [3] datasets. We have similar observations for the VQAv2 task.

FDT size	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 overall
-	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
8192	42.8	67.9	32.7	50.6	374.6	18.5	29.1	258.1	50.1
16384	**45.9**	**68.8**	32.6	**51.0**	376.5	**19.4**	29.6	**263.1**	50.6
24576	45.2	68.6	**33.3**	50.4	**378.5**	18.6	**29.7**	**263.1**	**51.4**

Table 7. Results of the models with different FDT sizes. The row whose FDT value is "-" represents the original CLIP model. "ZS" means zero-shot. "AVG" is average. "ACC" is accuracy. "LP" stands for linear prob. "CLS" represents classification. "IR" and "TR" are image retrieval and text retrieval.

	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 overall
CLIP	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP+FDT$_{\text{Softmax}}$ *	5.2	-	5.4	1.7	45.5	2.4	0.8	26.2	-
CLIP+FDT$_{\text{Sparsemax}}$ *	32.4	-	10.5	32.5	242.4	6.0	18.3	157.5	-
CLIP+FDT$_{\text{Softmax}}$	43.9	68.7	33.3	47.9	377.6	19.2	28.3	258.8	47.9
CLIP+FDT$_{\text{Sparsemax}}$	45.9	68.8	32.6	51.0	376.5	19.4	29.6	263.1	50.6

Table 8. Results of models trained with (Sparsemax) and without (Softmax) sparse constraints. The rows marked with "*" are the results when using FDT weights as features (see Section 4.3). "ZS" means zero-shot. "AVG" is average. "ACC" is accuracy. "LP" stands for linear prob. "CLS" represents classification. "IR" and "TR" are image retrieval and text retrieval.

Image Encoder Architecture. We evaluate the influence of different image encoder architectures on our proposed method, and the results are reported in Table 6. We observe that our method still significantly outperforms CLIP when using different types of image encoders. Additionally, FDT slightly adds an average of 6% more parameters, 13% more training time, and 12% less throughput when using different encoder architectures. The detailed results can be found in the supplementary materials.

FDT Numbers. The performance of models trained with different learnable token numbers are shown in Table 7. We can see that using 8192 tokens can already achieve an improvement over CLIP. Increasing the FDT size to 16384 obtains a more significant improvement than 8192, since it can encode more types of information. Furthermore, growing the FDT size to 24576 achieves a slight improvement over 16384 for the zero-shot image-text retrieval task on the Flickr30K dataset and VQA task. We set the FDT size as 16384 in our implementation because it achieves the best performance-efficiency tradeoff.

Sparse Constraints. In this section, we aim to demonstrate that applying sparse constraints helps the model learn better cross-modal correspondence, where the same cross-modal information is represented using the same subset of FDT. To this end, we evaluate the performance when using the FDT weights (Equation 5, 7 and 8) of each image or sentence as the features for zero-shot image classification and image-text retrieval tasks. The results are reported in Table 8. From the table, we can see that using sparse constraints (Sparsemax) achieves significantly better performance for all tasks. The results demonstrate that adding sparse constraints to FDT weights can lead to better cross-modal correspondence. Additionally, we can also see that without sparse constraints (Softmax), FDT-based features can also achieve significant performance over CLIP. Adding a sparse constraint (Sparsemax) achieves a larger performance improvement. This is because the granularities are further unified by representing the same cross-modal information with the same token set.

Figure 3. Examples shows the top-5 retrieved images for the given text queries for the text-to-image retrieval task on MSCOCO.

4.4. Analysis of the Completeness of Alignment

Since the granularities of image and text information are inconsistent, the learned model may fail to capture key semantic concepts [14]. In this experiment, we empirically evaluate whether unifying the granularities through the proposed FDT can alleviate the problem. The model pretrained on the 145M dataset is used for this evaluation.

To this end, we design a probing experiment on the

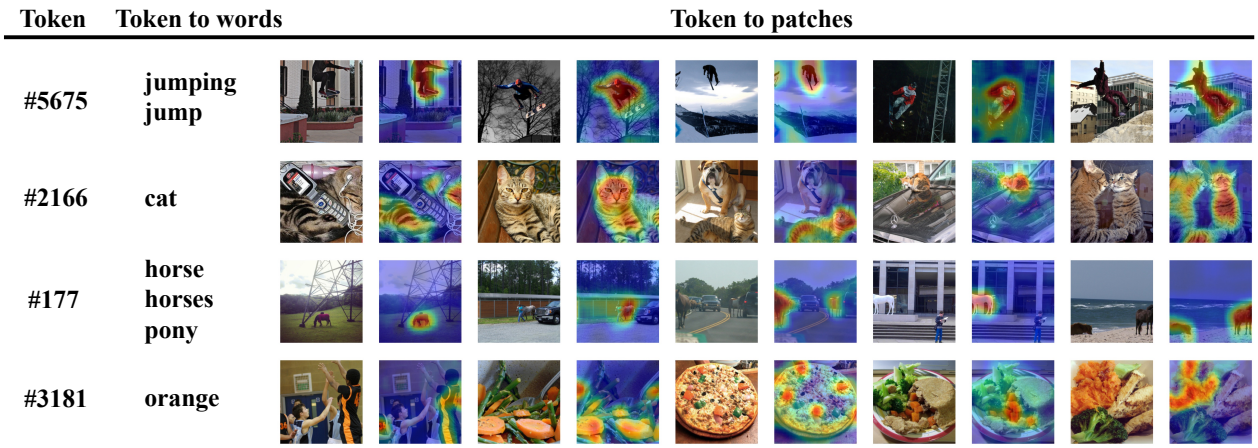

Figure 4. Example of the top-5 most relevant image patches and text tokens of four FDT tokens. Note that the redundant text tokens in the top-5 are removed. The color of the heatmap from blue to red denotes the relevance between patches and FDT from small to large.

MSCOCO dataset. Using the object detection annotations in the training split of MSCOCO, we construct 305,723 sentence pairs. For each sentence pair, one *matched sentence* describes all objects in an image, while the other *partially matched sentence* only captures part of the objects. Please refer to the supplementary material for more details about how we constructed these sentence pairs.

We then use pre-trained models to extract the embeddings of images and sentences and compute the similarity scores between the images and these constructed sentences. If the learned model comprehensively captures the semantic concepts, the similarity between an image and its matched sentence should be higher than that between the partially matched sentence. We found that the CLIP+FDT models can meet our expectation in 68.2 % of all sentence pairs, surpassing the CLIP model by 7.6%. The results demonstrate that FDT can help the CLIP model more comprehensively capture various semantic concepts. We assume that this is because the FDT serve as the prior knowledge that guides encoders to extract cross-modally shared high-level semantic concepts. This not only facilitates cross-modal interactions but also helps encoders capture semantic information from images and texts more comprehensively.

In addition, we show two cases for the text-to-image retrieval task in Figure 3. We can see that the images retrieved by CLIP ignore some important concepts described in the text queries. For example, in terms of the text query "baseball players entertaining a crowd of spectators", four out of the five images retrieved by the CLIP models contain baseball players only but with no spectators. Moreover, the image containing spectators is ranked lower than the two images without spectators. In contrast, FDT can retrieve images that contain both baseball players and spectators. More results are provided in the supplementary material.

4.5. Visualization of Learned FDT

To explicitly show the cross-modal correspondence learned by our FDT, we visualize the top-5 most relevant image patches and text tokens (using Equation 4 and 6) of four FDT tokens in Figure 4. The MSCOCO dataset and the model pretrained on the 145M dataset are used for visualization. The example cases show that each token captures different types of cross-modal correspondence, including actions (jump/jumping), objects, and attributes (orange color). Moreover, the learned FDT can potentially detect correspondent patches from the images. For example, the second token has high relevance values with patches of cats, while having low relevance with other patches. More results can be found in the supplementary material.

5. Conclusions

In this paper, we introduce a new multimodal presentation using finite discrete tokens (FDT). Specifically, a set of learnable tokens shared by all modalities are used to represent multimodal information conveyed in the image and text modalities. Our approach is a light-weighted way of fulfilling cross-modal interaction, where FDT serves as multimodal anchors to capture information from each input with better completeness. This help alleviate the model degradation problem commonly observed in vanilla CLIP models. Our FDT can be trained with the contrastive learning scheme from scratch without cold-start problems. Both quantitative and qualitative results demonstrate that FDT representations achieve better cross-modal alignment and performance on various downstream tasks, including image classification, cross-modal retrieval, and VQA. Additionally, the learned FDT capture meaningful cross-modal correspondence, ranging from objects to actions and attributes.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 5

[2] Soheil Bahrampour, Nasser M Nasrabadi, Asok Ray, and William Kenneth Jenkins. Multimodal task-driven dictionary learning for image classification. *IEEE transactions on Image Processing*, 25(1):24–38, 2015. 2, 4

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 4, 6

[4] Tianlang Chen, Yuxiao Chen, Han Guo, and Jiebo Luo. You type a few words and we do the rest: Image recommendation for social multimedia posts. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2124–2133. IEEE, 2018. 2

[5] Yuxiao Chen, Jianbo Yuan, Long Zhao, Tianlang Chen, Rui Luo, Larry Davis, and Dimitris N Metaxas. More than just attention: Improving cross-modal attentions with contrastive constraints for image-text matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4432–4440, 2023. 2

[6] 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 preprint arXiv:2203.05796*, 2022. 4, 6

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[9] Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE Transactions on Image Processing*, 31:1325–1339, 2022. 2

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1

[11] Cristina Garcia-Cardona and Brendt Wohlberg. Convolutional dictionary learning: A comparative review and new algorithms. *IEEE Transactions on Computational Imaging*, 4(3):366–381, 2018. 2, 4

[12] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. HiCLIP: Contrastive language-image pre-training with hierarchy-aware attention. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018. 1, 3, 7

[15] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 2

[16] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 4

[18] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2

[19] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo-2: End-to-end unified vision-language grounded learning. *arXiv preprint arXiv:2203.09067*, 2022. 2

[20] 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 preprint arXiv:2110.05208*, 2021. 2, 4, 5

[21] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[22] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 388–404. Springer, 2022. 1

[23] Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*, 2021. 2

[24] Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label clas-

sification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016. 4

[25] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 5

[26] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[27] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[29] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5566–5574, 2022. 1

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 4, 6

[31] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 2

[32] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[33] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[34] Lingxi Xie, Xiaopeng Zhang, Longhui Wei, Jianlong Chang, and Qi Tian. What is considered complete for visual recognition? *arXiv preprint arXiv:2105.13978*, 2021. 1

[35] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 1, 3

[36] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. 2022. 2

[37] 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 preprint arXiv:2111.07783*, 2021. 4, 5, 6

[38] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision*, pages 69–87. Springer, 2022. 5

[39] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5

[40] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[41] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[42] Renrui Zhang, Liuhui Wang, Yali Wang, Peng Gao, Hongsheng Li, and Jianbo Shi. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[43] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. *arXiv preprint arXiv:2212.04489*, 2022. 1