DIME-FM : DIstilling Multimodal and Efficient Foundation Models

Ximeng Sun[1†] Pengchuan Zhang[2] Peizhao Zhang[2] Hardik Shah[2] Kate Saenko[1,2] Xide Xia[2]
[1] Boston University, [2] Meta AI

Abstract

*Large **V**ision-**L**anguage **F**oundation **M**odels (VLFM), such as CLIP, ALIGN and Florence, are trained on large-scale datasets of image-caption pairs and achieve superior transferability and robustness on downstream tasks, but they are difficult to use in many practical applications due to their large size, high latency and fixed architectures. Unfortunately, recent work shows training a small custom VLFM for resource-limited applications is currently very difficult using public and smaller-scale data. In this paper, we introduce a new distillation mechanism (**DIME-FM**) that allows us to transfer the knowledge contained in large VLFMs to smaller, customized foundation models using a relatively small amount of inexpensive, unpaired images and sentences.* ***We transfer the knowledge from the pre-trained CLIP-ViT-L/14 model to a ViT-B/32 model, with only 40M public images and 28.4M unpaired public sentences****. The resulting model "Distill-ViT-B/32" rivals the CLIP-ViT-B/32 model pre-trained on its private WiT dataset (400M image-text pairs): Distill-ViT-B/32 achieves similar results in terms of zero-shot and linear-probing performance on both ImageNet and the ELEVATER (20 image classification tasks) benchmarks. It also displays comparable robustness when evaluated on five datasets with natural distribution shifts from ImageNet. Please refer to our project page for code and more details.*

1. Introduction

In contrast to neural networks learnt to solve a single target vision task (*i.e.* task-specific models) [26, 48, 70, 13, 68, 45], CLIP [55] and other **V**ision-**L**anguage "**F**oundation **M**odels" (VLFMs) [41, 84] achieve superior accuracy on diverse novel downstream tasks and improved robustness to natural domain shifts during inference. At the same time, small and customizable VLFMs are in high demand for many applications that have limited computational resources (AV, AR/VR and other edge devices). Unfortunately, only a few labs in the world can afford the large-scale vision-language datasets (e.g. WiT [55] with 400M image-text pairs) and the immense computing resources required to train VLFMs. Efforts to re-create VLFMs on public data [82, 64, 15]) either fall short on accuracy or require even more expensive training on huge datasets of images paired with captions (*e.g.* over 5B pairs [61]).

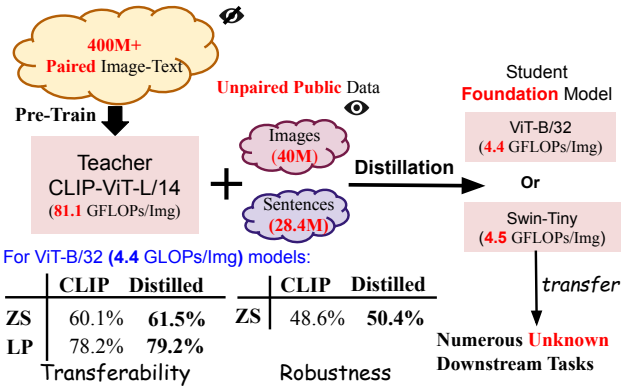


Figure 1: **Conceptual Figure of our Vision-Language Knowledge Distillation DIME-FM.** We distill the knowledge from a large VLFM "CLIP-ViT-L/14' pretrained on 400M private image-text paired dataset. We only use public unpaired image and text corpora as inputs. Our Distill-ViT-B/32 rivals CLIP-ViT-B/32 in both transferability and robustness. ZS: Zero-Shot, LP: Linear Probing.

Instead of pretraining, model distillation used to offer a convenient way to obtain a smaller custom model. Recent work distills CLIP specifically for one or a few target tasks (*i.e.* task-specific distillation). For example, some works [78, 79, 87, 54] distill the CLIP's image feature maps for better visual feature representations. BeamCLIP [36] distills CLIP logits for a single target image classification task, *e.g.* ImageNet-1K [16]. More recently, CLIP-TD [76] distills CLIP to solve three specific vision-language tasks. Even though these task-specific distillation works achieve good performance for the specialized downstream task, they are not scalable to solve new downstream tasks by zero-shot transferring. **There is no approach for distilling VLFMs to another foundation model** which preserves transferability to *novel* tasks and robustness to domain shifts. Due to the unaffordable large-scale pretraining and the lack of foundation-model distillation mechanism, **practitioners must rely on the few labs to release smaller VLFMs, and cannot easily customize their size or architecture.**

[†]Work done when interning at Meta AI.

In this work, we successfully distill smaller custom VLFMs using only smaller-scale public data, but achieving comparable transferability and robustness as if they were pre-trained on the large-scale data. Specifically , we transfer the knowledge from the released CLIP-ViT-L/14 [55] to our small VLFM "Distill-ViT-B/32". During the distillation, we adopt only **40M images from public datasets and 28.4M unpaired sentences**. Remarkably, with less than one-tenth of CLIP's pretraining dataset WiT, Distill-ViT-B/32 achieves comparable transferability and robustness to CLIP-ViT-B/32 [55] (see Fig. 1).

To accomplish this, we propose a novel distillation mechanism to **DI**still **M**ultimodal and **E**fficient **F**oundation **M**odels (**`DIME-FM`**) from CLIP. In standard distillation of image classification models with the fixed categories (*i.e*. fixed-vocabulary models), the class scores (logits) are matched between the teacher and student models [31, 37, 5, 52, 9]. However, since VLFMs do not have fixed-vocabulary logits, we instead match similarity of images to sentences (*i.e*. open-vocabulary logits) to retain the transferability (especially zero-shot ability) and robustness of VLFMs. We perform a careful ablation study of how "vocabulary", determined by training sentences, affects the student model's performance and find that it is crucial to perform distillation with a visually-related vocabulary rather than a random vocabulary. To construct a visually-related distillation text corpus, we propose an efficient algorithm that selects visually-grounded sentences (*i.e*. sentences which describe the visual world) from an NLP corpus rather than require the expensive human-annotated image captions or use noisy web-crawled image-related text. On top of text selection algorithm, we design two distillation losses to augment open-vocabulary logits in VLFM and empirically show that our novel distillation losses benefit vision-language (VL) knowledge distillation.

To summarize, we make three contributions in this paper:

1. We propose a vision-language knowledge distillation mechanism **`DIME-FM`** to transfer knowledge of pre-trained huge VLFMs to **small foundation models** with smaller-scale public images and unpaired sentences.

2. We distill the pre-trained CLIP-ViT-L/14 to Distill-ViT-B/32, with only unpaired 40M public images and 28.4M sentences. Notably, our Distill-ViT-B/32 rivals the CLIP-ViT-B/32 that was pre-trained on private 400M image-text paired data in both transferability and robustness.

3. Our proposed **`DIME-FM`** consists of an efficient algorithm to construct a visually-grounded text corpus from an NLP corpus and two specific distillation losses to augment open-vocabulary logits in VL distillation.

2. Related Works

Vision-Language Foundation Models. Many previous works focus on learning a generic alignment between language and vision features extracted by pretrained encoders [23, 38, 42, 49, 66, 77, 86] to improve many downstream tasks, *e.g*. Visual Question Answering (VQA) [4, 81, 34], Image Captioning [2, 44, 59, 27] *etc*. Recently, inspired by the great success on generic NLP model transferring to the downstream tasks [56, 57, 10], CLIP [55] and other large VLFMs [35, 41, 84, 43, 83] pretrain on hundreds of million image-text pairs to learn transferable visual representation from natural language supervision with contrastive learning. These works have shown astonishing transferring performance, such as zero-shot and linear probing evaluations, on various downstream tasks [40] as well as a great robustness to the distribution shift from ImageNet [55]. Without the use of private large-scale data, it is challenging to learn small custom foundation models that possess comparable transferability and robustness. ELEVATER evaluation [40] shows that training VLFMs [82, 64] using relative small public datasets ($\leq$ 40M image-text pairs) and even with help of external knowledge, *e.g*. WordNet [51] and Wiktonary [50], cannot close the performance gap in comparison to CLIP [55] or Florence [84]. Trained with CLIP-Filtered 400M image-text pairs [62], OpenCLIP [15] still performs worse than CLIP at each model size because of the possible poorer quality of paired data. In this paper, instead of pretraining the model using contrastive loss with paired data, we distill from CLIP-ViT-L/14 to different models with smaller-scale public images and unpaired sentences.

Uni-modal Knowledge Distillation. In general, knowledge distillation [31] transfers knowledge from one model (teacher) to another (student). It optimizes a student model to match some certain output of the teacher model. With a single modality, there are two main ways of distillations: (1) knowledge distillation of the fixed-vocabulary prediction logits [31, 37, 5, 52, 9]. (2). feature distillation on the final or intermediate activation of the network [60, 33, 3, 30, 85, 69]. In this paper, we do not require the same feature dimension in both teacher and student foundation models. To avoid complex tricks to circumvent the mismatch of feature dimensions using feature distillation methods, we adopt the simple logit distillation for the vision-language distillation. Instead of applying KL divergence loss to fixed-vocabulary logits in the uni-modal logit distillation, we apply KL divergence loss to feature similarity scores (*i.e*. open-vocabulary logits) in VLFMs. Moreover, we still use the uni-modal logit distillation as a regularizer in the distillation.

Model Distillation from CLIP. Some works [78, 79, 87, 54] perform feature distillation of CLIP image encoder with Masked Image Modeling [7, 17, 25, 18, 6, 19, 75, 80] to learn a new image encoder which claim superior finetuning performance on ImageNet-1K [16] and ADE20K [88]. They ignore the language encoder during the distillation and do not maintain the alignment of image and text in the feature space. BeamCLIP [36] distills the CLIP using logits computed by

images from the public image datasets and class names of ImageNet-1K, and achieves better ImageNet-1K Top-1 linear probe accuracy than vision-only self-supervised learning (SSL) methods [14, 11]. CLIP-TD [76] distills knowledge from CLIP into existing architectures to solve targeted vision-language (VL) tasks. Even though these works achieve better performance in their specific tasks, their student models lose the capability of VLFMs, as they are not scalable to solve new tasks by zero-shot transferring. Instead of distilling CLIP and tuning it for specific downstream task(s), we wish to distill another foundation model from CLIP, and our result model yields the comparable transferability and robustness performance to the foundation models with the similar model size but pretrained on hundreds of million image-text pairs.

3. Vision-Language Knowledge Distillation

In this paper, we propose our VL knowledge distillation **DIME-FM** which uses the public unpaired images and text to distill a small VLFM from a pretrained large VLFM (CLIP-ViT-L/14). First, we mathematically define VLFMs and our VL distillation setting. Then, we introduce our novel training losses and our text construction algorithm.

Preliminaries. A dual-encoder VLFM consists of an image and text encoder to extract image/text embeddings respectively, then project the image and text embeddings to the common feature space. To get more flexible design choices for the dimensions of the separate image/text feature spaces and the final shared feature space, we separate the image and text projection layers from the image and text feature encoders. Therefore, a standard dual-encoder VLFM can be defined as a quartet $[f_\theta, g_\phi, \mathbf{A}, \mathbf{B}]$. f_θ and g_ϕ are image and text encoders which encode the image $\boldsymbol{x}$ and text $\boldsymbol{t}$ into their own feature spaces (as $\boldsymbol{u}' \in \mathbb{R}^{d^v}$ and $\boldsymbol{v}' \in \mathbb{R}^{d^l}$)[1] respectively. $\mathbf{A} \in \mathbb{R}^{d \times d^v}$ and $\mathbf{B} \in \mathbb{R}^{d \times d^l}$ are two linear layers projecting image and text embeddings ($\boldsymbol{u}'$ and $\boldsymbol{v}'$) to $\boldsymbol{u}$ and $\boldsymbol{v}$ in a shared d-dim feature space:

$$\boldsymbol{u}' = f_{\boldsymbol{\theta}}(\boldsymbol{x}), \quad \boldsymbol{v}' = g_{\boldsymbol{\phi}}(\boldsymbol{t}), \quad \boldsymbol{u} = \mathbf{A}\boldsymbol{u}', \quad \boldsymbol{v} = \mathbf{B}\boldsymbol{v}' \quad (1)$$

The similarity score between the image and text embeddings

$$s(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T \boldsymbol{v} / (\|\boldsymbol{u}\| \|\boldsymbol{v}\|) \quad (2)$$

reveals the semantic relationship between image and text encoded in VLFMs. It plays an important role in transferring to downstream tasks and being robust to domain shift.

Problem Definition. Given a public unpaired image corpus $\mathcal{X}$ and text corpus $\mathcal{T}$, we distill a small VLFM $[f_{\widehat{\boldsymbol{\theta}}}, g_{\widehat{\boldsymbol{\phi}}}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}]$ from a pretrained large VLFM $[f_\theta, g_\phi, \mathbf{A}, \mathbf{B}]$, where

$$\widehat{\boldsymbol{u}}' = f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}) \in \mathbb{R}^{\widehat{d^v}}, \quad \widehat{\boldsymbol{v}}' = g_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{t}) \in \mathbb{R}^{\widehat{d^l}}, \quad (3)$$

$$\widehat{\boldsymbol{u}} = \widehat{\mathbf{A}}\widehat{\boldsymbol{u}}' \in \mathbb{R}^{\widehat{d}}, \quad \widehat{\boldsymbol{v}} = \widehat{\mathbf{B}}\widehat{\boldsymbol{v}}' \in \mathbb{R}^{\widehat{d}}, \quad (4)$$

[1]The upper scripts v and l are short for vision and language, respectively.

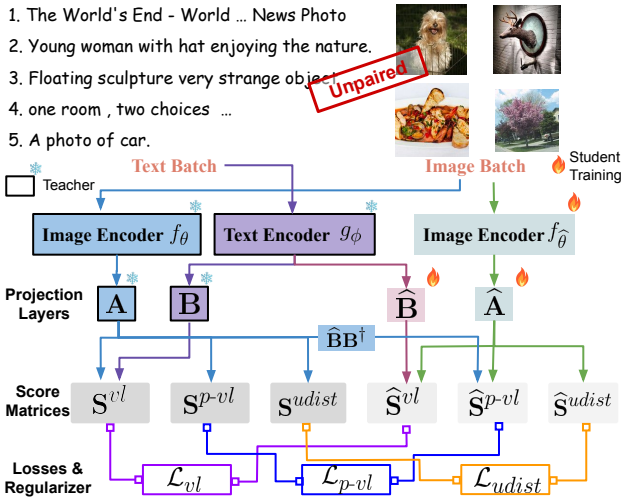


Figure 2: **Illustration of our proposed distillation losses.** In each iteration, we compute two losses ($\mathcal{L}_{vl}$, $\mathcal{L}_{p\text{-}vl}$) and one regularizer ($\mathcal{L}_{udist}$) with a min-batch of images and texts to distill knowledge from the teacher to the student. We freeze all parameters in the teacher model and learn the student model from scratch.

where $\widehat{(\cdot)}$ is the component in the student model corresponding to $(\cdot)$ in the teacher model. Notably, we can freely choose the image, text and projected embeddings' dimensions ($\widehat{d^v}$, $\widehat{d^l}$ and $\widehat{d}$) in the student VLFM, which can be different from those (d^v, d^l and d) in the teacher VLFM.

In contrast to the expensive pretraining VLFMs [55, 41, 15] with large-scale image-text pairs, we do not require any paired data for optimization. During the distillation, we match the similarity scores of feature embeddings between the teacher and student VLFMs, which ensures our distilled small image encoder $f_{\widehat{\boldsymbol{\theta}}}$ is still superior in transferability and robustness, as if it were trained on large-scale paired data. To this end, we propose our VL distillation mechanism **DIME-FM** including two novel distillation losses (Sec. 3.1) and an efficient text selection algorithm to construct the training text corpus (Sec. 3.2).

The capacity of CLIP text encoder has few effects on CLIP performance [55]. Also, the inference latency for close-vocabulary downstream visual tasks [40, 67, 32]. To make the presentation simple while keeping the essential idea, we fix the text encoder, i.e., $g_{\widehat{\boldsymbol{\phi}}} = g_{\boldsymbol{\phi}}$, and focus on VL knowledge to distill a small custom image encoder $f_{\widehat{\boldsymbol{\theta}}}$ as a transferable and robust vision backbone. If a small text encoder $g_{\widehat{\boldsymbol{\phi}}}$ is desired for open-vocabulary downstream tasks, we can apply the proposed method to distill $g_{\boldsymbol{\phi}}$ while fixing $f_{\widehat{\boldsymbol{\theta}}}$, which is left as an interesting area of future work.

3.1. Optimization for VL Knowledge Distillation

In standard uni-modal logit distillation, the objective is to match the fixed-vocabulary logits predicted by the student to the logits predicted by the teacher on the same input sample [31]. In VLFMs, the vocabulary is not fixed and the

outputs are similarity score between images and sentences. Thus we change our objective to match the distribution of these scores produced by the student to the distribution produced by the teacher. Specifically, we minimize the KL divergence of score distributions computed over image dataset $\mathcal{X}$ and text dataset $\mathcal{T}$ (see Eq. 2) using three separate losses.

We first define the general form of applying KL divergence to distill the similarity scores and then define the three losses.. Suppose we have two batches of embeddings $\{\boldsymbol{w}_i^1\}_{i=1}^{B_1}$ and $\{\boldsymbol{w}_j^2\}_{j=1}^{B_2}$ [2] in the teacher model's shared d-dim feature space. All similarity scores form a teacher score matrix $\mathbf{S} \in \mathbb{R}^{B_1 \times B_2}$, where $\mathbf{S}_{i,j} = s(\boldsymbol{w}_i^1, \boldsymbol{w}_j^2)$. Similarly, we have the student's score matrix as $\widehat{\mathbf{S}} \in \mathbb{R}^{B_1 \times B_2}$, where $\widehat{\mathbf{S}}_{i,j} = s(\widehat{\boldsymbol{w}}_i^1, \widehat{\boldsymbol{w}}_j^2)$. Each row and column of the score matrix can be seen as *open-vocabulary logits*. We measure the row-wise (indexing i) and column-wise (indexing j) discrepancy between $\mathbf{S}$ and $\widehat{\mathbf{S}}$ with KL divergence:

$$\begin{aligned}\mathcal{L}_{KL}(\widehat{\mathbf{S}}; \mathbf{S}, \mu) = &\sum\nolimits_i \mathrm{KL}(\sigma(\mu \mathbf{S}_i) || \sigma(\mu \widehat{\mathbf{S}}_i)) \\ &+ \sum\nolimits_j \mathrm{KL}(\sigma(\mu \mathbf{S}_j^T) || \sigma(\mu \widehat{\mathbf{S}}_j^T)),\end{aligned} \tag{5}$$

where σ is the softmax function and μ is a temperature.

In particular, we propose two losses ($\mathcal{L}_{vl}$ and $\mathcal{L}_{p\text{-}vl}$, described in detail below) in form of Eq. 5 with different $\mathbf{S}$ and $\widehat{\mathbf{S}}$'s. The third loss is a regularizer $\mathcal{L}_{udist}$ to maintain the Euclidean Distance between every pair of image embeddings (a.k.a geometry of image embeddings) during the distillation. Our final VL distillation objective is (see Fig 2):

$$\min_{f_{\widehat{\boldsymbol{\theta}}}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}} (1 - \lambda_1)\mathcal{L}_{vl} + \lambda_1 \mathcal{L}_{p\text{-}vl} + \lambda_2 \mathcal{L}_{udist}\ , \tag{6}$$

where $\lambda_1 \in [0, 1]$ and $\lambda_2 \in \mathbb{R}^+$ are two hyperparameters to control each loss weight. We study the efficacy of three losses with various λ_1 and λ_2's in Sec. 4.5.

VL Score Distillation Loss $\mathcal{L}_{vl}$. We distill the VL score matrices in form of Eq. 5. Given an image batch $\{\boldsymbol{x}_i\}_{i=1}^{B^v} \subset \mathcal{X}$ and a text batch $\{\boldsymbol{t}_j\}_{j=1}^{B^l} \subset \mathcal{T}$, they are projected to $\{\boldsymbol{u}_i\}_{i=1}^{B^v}$ and $\{\boldsymbol{v}_i\}_{i=1}^{B^l}$ in the teacher's shared feature space respectively and projected to $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^{B^v}$ and $\{\widehat{\boldsymbol{v}}_i\}_{i=1}^{B^l}$ in the student's feature space. Therefore, we define the teacher's and student's VL score matrices as

$$\mathbf{S}_{i,j}^{vl} = s(\boldsymbol{u}_i, \boldsymbol{v}_j), \quad \widehat{\mathbf{S}}_{i,j}^{vl} = s(\widehat{\boldsymbol{u}}_i, \widehat{\boldsymbol{v}}_j), \tag{7}$$

with which we define VL Score Distillation Loss as:

$$\mathcal{L}_{vl} = \mathcal{L}_{KL}(\widehat{\mathbf{S}}^{vl}, \mathbf{S}^{vl}, \mu^{vl}) \tag{8}$$

Pseudo-VL Score Distillation Loss $\mathcal{L}_{p\text{-}vl}$. Our study on the efficacy of text corpus (see Sec 4.4) shows that enlarging the text corpus $\mathcal{T}$ introduces more text embeddings and results in more open-vocabulary logits, which in turn benefits the VL knowledge distillation.

Motivated by this, besides adding more visually-grounded sentences to $\mathcal{T}$, we introduce image embeddings as additional pseudo text embeddings. Since image and text embedding are trained to live in a shared sphere (*i.e.* $\forall i,\ \|\boldsymbol{u}_i\|_2 = \|\boldsymbol{v}_i\|_2 = 1$), image embeddings are a reasonable substitute for embeddings of visually-grounded text. For a given image $\boldsymbol{x}_j$ and its image embedding $\boldsymbol{u}_j$, we assume that there is a sentence $\boldsymbol{t}_j$ whose text embedding $\boldsymbol{v}_j$ perfectly matches $\boldsymbol{u}_j$ in the shared sphere:

$$\boldsymbol{v}_j = \boldsymbol{u}_j, \quad \boldsymbol{v}_j = \mathbf{B}\boldsymbol{v}'_j \Rightarrow \boldsymbol{v}'_j \approx \mathbf{B}^\dagger \boldsymbol{v}_j = \mathbf{B}^\dagger \boldsymbol{u}_j, \tag{9}$$

where $\mathbf{B}^\dagger$ is the pseudo-inverse[3] of matrix $\mathbf{B}$. We treat the image embedding $\boldsymbol{u}_j$ as the pseudo paired text embedding of the input image $\boldsymbol{x}_j$ in the teacher model. For the student model, based on Eq. 4, 9 and $\boldsymbol{v}'_j = \widehat{\boldsymbol{v}'}_j$ (due to the fixed text encoder), we get the pseudo paired text embedding $\widehat{\boldsymbol{v}}_j$ of the image $\boldsymbol{x}_j$ as $\widehat{\boldsymbol{v}}_j = \widehat{\mathbf{B}}\widehat{\boldsymbol{v}}'_j = \widehat{\mathbf{B}}\boldsymbol{v}'_j \approx \widehat{\mathbf{B}}\mathbf{B}^\dagger \boldsymbol{u}_j$. We note that $\widehat{\mathbf{B}}\mathbf{B}^\dagger \boldsymbol{u}_j \equiv \boldsymbol{u}_j$ when we do not reduce the projected dimension ($\widehat{d} = d$) and keep $\widehat{\mathbf{B}} = \mathbf{B}$. By replacing the text embeddings $\boldsymbol{v}_j$ and $\widehat{\boldsymbol{v}}_j$ in Eq. 7 with pseudo text embeddings $\boldsymbol{u}_j$ and $\widehat{\mathbf{B}}\mathbf{B}^\dagger \boldsymbol{u}_j$ respectively, we get the pseudo VL score matrices as:

$$\mathbf{S}_{i,j}^{p\text{-}vl} = s(\boldsymbol{u}_i, \boldsymbol{u}_j), \quad \widehat{\mathbf{S}}_{i,j}^{p\text{-}vl} = s(\widehat{\boldsymbol{u}}_i, \widehat{\mathbf{B}}\mathbf{B}^\dagger \boldsymbol{u}_j) \tag{10}$$

with which we define pseudo-VL Score Distillation Loss as:

$$\mathcal{L}_{p\text{-}vl} = \mathcal{L}_{KL}(\widehat{\mathbf{S}}^{p\text{-}vl}; \mathbf{S}^{p\text{-}vl}, \mu^{p\text{-}vl}). \tag{11}$$

Some uni-modal self-supervised learning (SSL) works [65, 53] also compute the similarity score matrix (similar to $\mathbf{S}^{p\text{-}vl}$) from the same image batch, and then assign the positive/negative ground-truth label for each element in the score matrix. However, in the VL distillation, we treat $\mathbf{S}^{p\text{-}vl}$ as the supplement to $\mathbf{S}^{vl}$ which further augments text embeddings. Moreover, we use $\mathbf{S}^{p\text{-}vl}$ as the pseudo label from the teacher and minimize the discrepancy between $\widehat{\mathbf{S}}^{p\text{-}vl}$ and $\mathbf{S}^{p\text{-}vl}$ without any ground-truth labels.

Uni-Modal Distance Preserving Regularizer $\mathcal{L}_{udist}$. In addition to matching the similarity score of a student image embedding and a teacher image embedding in $\mathcal{L}_{p\text{-}vl}$, we introduce a regularizer $\mathcal{L}_{udist}$, which distills similarity score s of two normalized student image embeddings[4] from the teacher model, to keep the geometry of image embeddings in the student model close to that in the teacher model.

Suppose we have two images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as well as their projected embeddings ($\boldsymbol{u}_i$ and $\boldsymbol{u}_j$) in the teacher's feature

[2] they can be image or text's embeddings. This will be further explained

[3] also known as Moore–Penrose inverse

[4] The similarity score of two normalized embeddings already encodes their relative locations.

space and projected embeddings ($\widehat{\boldsymbol{u}}_i$ and $\widehat{\boldsymbol{u}}_j$) in the student's feature space . We define the score matrices to preserve the distances of image embeddings as:

$$\mathbf{S}_{i,j}^{udist} = s(\boldsymbol{u}_i, \boldsymbol{u}_j), \quad \widehat{\mathbf{S}}_{i,j}^{udist} = s(\widehat{\boldsymbol{u}}_i, \widehat{\boldsymbol{u}}_j). \tag{12}$$

We define the uni-distance preserving loss as a regularization term in the VL distillation as:

$$\mathcal{L}_{udist} = \mathcal{L}_{KL}(\widehat{\mathbf{S}}^{udist}; \mathbf{S}^{udist}, \mu^{udist}). \tag{13}$$

Although $\mathcal{L}_{udist}$ and $\mathcal{L}_{p\text{-}vl}$ only differ in $\widehat{\mathbf{B}}\mathbf{B}^{\dagger}u_j$ and $\widehat{u_j}$, their mechanisms are theoretically different: $\mathcal{L}_{p\text{-}vl}$ views image features as pseudo text features and utilizes them for distillation (thus $\mathcal{L}_{p\text{-}vl}$ is a text-encoder aware loss), while $\mathcal{L}_{udist}$ simply preserves the geometric structure in the visual encoder (thus $\mathcal{L}_{udist}$ is a text-encoder agnostic regularizer).

3.2. Constructing Visually-Grounded Text Corpus

To effectively distill the information from the pre-trained VLFMs, the choice of image corpus $\mathcal{X}$ and text corpus $\mathcal{T}$ is crucial. Constructing image corpus $\mathcal{X}$ is relatively easy due to large scale natural images available on web although care must be taken to filter them to avoid duplicates and harmful content and increase diversity. However, we cannot simply use text crawled from web as $\mathcal{T}$, because the concept distribution of natural language corpus is very different from that of a visual-grounded sentence corpus. As we show in Sec. 4.6, we use 3 million unfiltered natural sentences as $\mathcal{T}$ which gives much worse performance than using 3 million image captions of GCC-3M [63]. So it is important to select $\mathcal{T}$ relating to visual concepts.

With an image-text paired dataset $\{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^N$, a simple option is that we take $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ and $\mathcal{T} = \{\boldsymbol{t}_i\}_{i=1}^N$, where $\mathcal{T}$ and $\mathcal{X}$ have overlapped semantic meanings. However, we do not assume the availability of any image-text paired data and this simple option is not achievable .

Since the vision-language teacher model $[f_\theta, g_\phi, \mathbf{A}, \mathbf{B}]$ maps images and text into the same feature space, we can quantify the modality gap between $\mathcal{T}$ and $\mathcal{X}$ by measuring the distribution discrepancy between their projected embedding distributions. Given a large NLP corpus $\mathcal{T}_{large}$, we can select $\mathcal{T}$ from $\mathcal{T}_{large}$, by minimizing the discrepancy between $\mathcal{T}$'s and $\mathcal{X}$'s embedding distributions:

$$\min_{\mathcal{T} \subset \mathcal{T}_{large}} \quad \text{Discrepancy}(\mathcal{U}, \mathcal{V}) \tag{14}$$

$$\text{s.t.} \quad \mathcal{U} = \{\mathbf{A} f_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}, \tag{15}$$

$$\mathcal{V} = \{\mathbf{B} g_{\boldsymbol{\phi}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{T}\} \tag{16}$$

This is a combinatorial optimization problem. It is expected to be NP-hard to find the exact global minimum. We propose Algorithm 1 using greedy search to approximately solve the problem. Specifically, For each image in image datasets, we select the sentence with the highest similarity score (computed by the teacher) from NLP Corpus. We then form a visual-grounded text corpus with the selected text. We assume that the cardinality of $\mathcal{T}$ and $\mathcal{U}$ is similar. If we want $|\mathcal{T}| < |\mathcal{U}|$, we can simply do a $|\mathcal{T}|$-mean clustering of Algorithm 1's outputs, and construct $\mathcal{T}$ with the resulting cluster centers. We do not see a need for $|\mathcal{T}| > |\mathcal{U}|$, since the image corpus $\mathcal{X}$ can be as large as we want.

Generally, the downstream tasks are unknown before distillation. We use our constructed $\mathcal{T}$ as the text input and call this **Task-Agnostic VL Distillation**. However, in practice, sometimes we know some of the class names used in the downstream tasks before distillation. In this case, we can incorporate those class names into the training text corpus $\mathcal{T}$. We refer to this as **Task-Aware VL Distillation**. We compare these two VL distillations in Sec 4.3.

Algorithm 1: Constructing text corpus $\mathcal{T}$

Input: image embeddings $\mathcal{U}$ as defined in Eq.15. A large text corpus $\mathcal{T}_{large}$.
Output: Selected text corpus $\mathcal{T}$, and $|\mathcal{T}| \approx |\mathcal{U}|$

$\mathcal{U}_{left} \longleftarrow \mathcal{U}$, $\mathcal{T}_{avail} \longleftarrow \mathcal{T}_{large}$, $\mathcal{T} \longleftarrow \emptyset$, $U_p = \infty$
while $\mathcal{U}_{left} \neq \emptyset$ and $|\mathcal{U}_{left}|/U_p < 0.95$ **do**
  $U_p = |\mathcal{U}_{left}|$, $Matched = dict()$
  for $\boldsymbol{u} \in \mathcal{U}_{left}$:
    /* find the best text that matches the image */
    $\boldsymbol{t}(\boldsymbol{u}) = \arg\max_{\boldsymbol{t} \in \mathcal{T}_{avail}} s(\boldsymbol{u}, \mathbf{B} \cdot g_{\boldsymbol{\phi}}(\boldsymbol{t}))$ //
    $Matched[\boldsymbol{u}] = \boldsymbol{t}(\boldsymbol{u})$
  for $\boldsymbol{u}, \boldsymbol{t} \in Matched.items()$:
    /* For all images matching to the same text, pick the first match */
    if $\boldsymbol{t} \in \mathcal{T}_{avail}$:
      $\mathcal{U}_{left} \longleftarrow \mathcal{U} \backslash \{\boldsymbol{u}\}$
      $\mathcal{T}_{avail} \longleftarrow \mathcal{T}_{avail} \backslash \{\boldsymbol{t}\}$, $\mathcal{T}.add(\boldsymbol{t})$

4. Experiments

We first compare our distilled models to two state-of-the-art VLFMs with the same model capacity, CLIP [55] and UniCL [82]. We then compare task-agnostic and task-aware knowledge distillation, and investigate the influence of data scale on transferability and robustness. Finally, we carefully ablate our proposed distillation losses and our algorithm for text corpus construction.

4.1. Settings

Evaluation benchmarks. Foundation models are typically evaluated on transferability to downstream tasks (via zero-shot and linear probing) as well as robustness to data shifts. Following [40], we evaluate all baselines and our models in

Method	Train Data	Loss	Zero-Shot		Linear Probing	Robustness
			ELEVATER	IN-1K	ELEVATER	
CLIP-ViT-B/32	400M Image-Text Pairs	Contrastive Loss	**57.2%**	63.4%	78.2%	48.6%
Distill-ViT-B/32 (**DIME-FM**)	40M Images, 28.6M Captions	$\mathcal{L}_{vl}$	53.5%	64.2%	77.9%	48.1%
		$0.7\mathcal{L}_{vl} + 0.3\mathcal{L}_{p\text{-}vl}$	55.8%	64.8%	78.4%	49.4%
		$0.7\mathcal{L}_{vl} + 0.3\mathcal{L}_{p\text{-}vl} + 0.5\mathcal{L}_{udist}$	55.0%	65.1%	78.6%	49.4%
	40M Images, 28.4M NLP Text	$0.7\mathcal{L}_{vl} + 0.3\mathcal{L}_{p\text{-}vl}$	56.4%	**66.5%**	**79.2%**	**50.2%**

Table 1: **DIME-FM vs. CLIP.** We distill Distill-ViT-B/32 from CLIP-ViT-L/14 (81.1G FLOPs/img) and compare it with CLIP-ViT-B/32.

Dataset	Method	Zero-Shot		Linear Probing
		ELEVATER	IN-1K	ELEVATER
IN-21K	UniCL	27.2%	28.5%	74.8%
	UniCL*	40.9%	51.4%	75.3%
	Distill-UniCL*	**45.6%**	**59.5%**	**76.2%**
IN-21K + YFCC-14M	UniCL	37.1%	40.5%	**77.1%**
	UniCL*	44.6%	58.7%	75.4%
	Distill-UniCL*	**47.6%**	**60.0%**	76.6%

Table 2: **DIME-FM vs. UniCL.** We distill a Swin-Tiny Transformer from CLIP ViT-L/14 and compare it to Swin-Tiny UniCL model trained with the same dataset. UniCL* is defined in Sec. 4.2.

three settings: (1) Average **Zero-Shot on ELEVATER** [40], a dataset of 20 image-classification tasks; (2) **Zero-Shot on IN-1K**, the ImageNet-1K [16] validation set; (3) Average **Linear Probing on ELEVATER**. For robustness, we follow CLIP [55] to report average zero-shot performance on five datasets [58, 28, 8, 74, 29] with domain shifts from IN-1K.

Training Data. Following the academic track proposed in ELEVATER [40], we form our *image corpus* with images from ImageNet-21K (i.e. ImageNet-22K [39] excluding IN-1K classes), GCC-15M (including GCC-3M [63] and GCC-12M [12]) and YFCC-14M [71]. **We construct our *text corpus* in two different ways**: (1) Following UniCL [82], from GCC-15M and YFCC-14M captions and the prompt sentences with ImageNet-21K (IN-21K) class names and 80 templates; (2) Selecting $\mathcal{T}$ from $\mathcal{T}_{large}$ using images in GCC-15M and YFCC-14M with Algorithm 1. We use ROBERTa [46]'s pretraining datasets [24, 89, 1, 72, 21] (total of 1.58B sentences) as $\mathcal{T}_{large}$.

Note that we generally do not use paired image-text data in training. We never load image-text pairs and never use pair labels explicitly in our loss function unless specified. For each experiment, we specify the exact image and text corpora used for distillation.

Other Settings. We find that $\mathcal{L}_{vl}$ alone achieves good performance, so we use it as our loss function in most experiments except for Table 1 & 3 and Fig. 4. In all experiments we distill only the image encoder and use it together with the teacher's text encoder in evaluation. See Supplementary Material Sec.A for implementation and evaluation details.

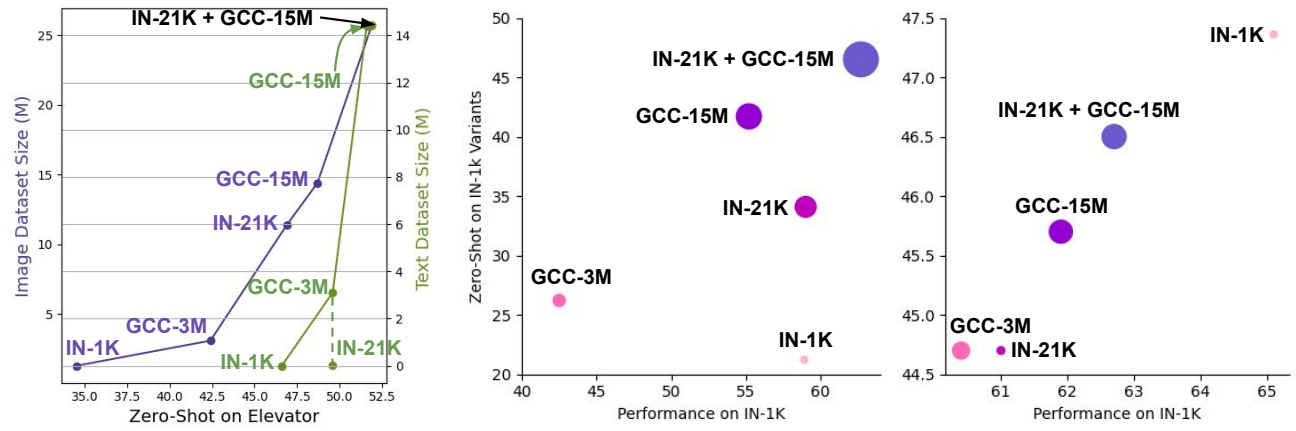
4.2. Comparison with CLIP and UniCL

Comparison with CLIP. We distill a small model from the released CLIP-ViT-L/14 checkpoint using the ViT-B/32 image encoder [20]. We compare Distill-ViT-B/32 with CLIP-ViT-B/32 in Table 1. Both models have the same inference cost (4.4 G FLOPs/img), but CLIP is trained on the private 400M WiT dataset [55], while ours uses 40M images and 28.6M sentences from public datasets (IN-21K, GCC-15M and YFCC-14M). Training with just $\mathcal{L}_{vl}$ slightly underperforms CLIP, but after adding $\mathcal{L}_{p\text{-}vl}$ to expand the vocabulary, the two models' performance becomes similar across the zero-shot and linear-probing testbeds. $\mathcal{L}_{udist}$ improves our zero-shot accuracy on IN-1K and linear-probing on ELEVATER but reduces zero-shot accuracy on ELEVATER. The robustness score of our distilled model is higher than CLIP-ViT-B/32 when training with $\mathcal{L}_{p\text{-}vl}$ and $\mathcal{L}_{udist}$. While this can be partially explained by our higher accuracy on IN-1K, it is still remarkable as we use less than one-tenth of CLIP training data and no image-text pairs.

Instead of captions, we also try using a text corpus $\mathcal{T}$ consisting of 28.4M sentences selected using Algorithm 1 from a language-only corpus $\mathcal{T}_{large}$, using query images from GCC-15M and YFCC-14M. Distilling on $\mathcal{T}$ and IN-21K prompt sentences, Distill-ViT-B/32 yields better average zero-shot performance on ELEVATER and IN-1K than CLIP-ViT-B/32 (61.4% vs. 60.3%), better linear probing performance (79.2% vs. 78.2%) as well as better robustness (50.2% vs. 48.6%). We analyze the quality of our constructed $\mathcal{T}$ and the human-annotated captions in Sec. 4.6.

We note that Distill-ViT-B/32 falls short on Zero-Shot on ELEVATOR. After the careful analysis, we find the large CLIP-ViT-L/14 performs much worse than the small CLIP-ViT-B/32 on PatchCamelyon [73] (51.2% vs. 60.7%) and KITTI Distance [22] (13.8% vs. 29.0%) in ELEVATER. After removing these two tasks, Distill-ViT-B/32 yields the same zero-shot score (61.0%) on ELEVATER as CLIP-ViT-B/32. See Supplementary Material Sec. E for more analysis for each individual downstream dataset.

Comparison with UniCL. In Table 2, we compare our distillation approach with UniCL [82], which trains contrastively on smaller-scale public image-text pairs, unifying captioning datasets and pseudo-captioned classification datasets. Following the settings in UniCL, we adopt "IN-21K" and "IN-21K + YFCC14M" as our training datasets and use Swin-Tiny Transformer [47] as our student image encoder. In UniCL, both image and text encoders are trained from scratch. We report UniCL's performance by evaluating its released checkpoints trained with two different



(a) Transferability of varying Image/Text data Size **(b.i) Robustness** when changing Image data **(b.ii) Robustness** when changing Text data

Figure 3: **Transferability and Robustness for different Image/Text Dataset Sizes.** (a) zero-shot transferability of our student model increases with larger training image/text corpus; (b.i) shows robustness strongly correlates to the training image dataset size (represented as the dot size); (b.ii) shows robust score strongly correlates to IN-1K performance when changing the training text.

λ_1	Input Text Corpus	Zero-Shot ELEVATER	Zero-Shot IN-1K	Linear Probing ELEVATER
0	*28.6M Text*	53.5%	64.2%	77.9%
	IN-1K Prompt Text[36]	47.4%	**67.2%**	76.9%
	DS Prompt Text	56.8%	57.2%	78.9%
	28.6M + DS Prompt Text	**57.5%**	65.6%	**79.2%**
0.3	*28.6M Text*	55.8%	64.8%	78.4%
	IN-1K Prompt Text[36]	50.8%	**66.6%**	77.1%
	DS Prompt Text	**57.8%**	60.3%	**79.4%**
	28.6M + DS Prompt Text	57.7%	66.1%	**79.4%**
0.8	*28.6M Text*	55.5%	64.2%	78.8%
	IN-1K Prompt Text[36]	53.1%	**65.4%**	78.0%
	DS Prompt Text	**59.4%**	61.7%	**79.8%**
	28.6M + DS Prompt Text	57.5%	64.9%	79.5%

Table 3: **Task-Agnostic vs. Task-Aware.** DS = Downstream.

data sources. For a fair comparison, we further introduce UniCL* in which we use the pretrained CLIP-ViT-L/14 text encoder as UniCL's. During training, we fix the text encoder's weights and only optimize the image encoder with contrastive loss. UniCL* achieves better zero-shot performance than UniCL due to CLIP's strong text encoder. Nevertheless, our Distilled-UniCL* significantly outperforms UniCL* on all evaluation benchmarks with only the $\mathcal{L}_{vl}$ loss. This indicates that distilling a small VLFM using strong pseudo-labels from large VLFMs is better than contrastive pretraining when we do not have large-scale datasets. Even though our experiment with "IN-21K + YFCC-14M" shows that enlarging data scale reduces the performance gap between distillation and pretraining (more analysis in Supplementary Material Sec. E), **DIME-FM** is more data-efficient, since it does not require any expensive image-text pairs.

4.3. Task-Agnostic vs. Task-Aware

We evaluate Task-agnostic VL Distillation and Task-aware VL Distillation in Table 3. We show performance on downstream tasks with known classes and generalization to other downstream tasks with unknown classes, under different loss weight λ_1. BeamCLIP [36] uses *only* the IN-1K prompt text to distill CLIP's image encoder. Table 3 shows that this generalizes poorly to other unknown downstream tasks (*e.g.* ELEVATER). With larger weight λ_1 on $\mathcal{L}_{p\text{-}vl}$ to expand text embeddings, BeamCLIP's student model generalizes better on ELEVATER but is still worse than our task-agnostic knowledge distillation. When we target multiple downstream tasks and only use prompt text with their class names (*i.e.* IN-1K and ELEVATER) as the input text corpus (denoted as "DS Prompt Text" in Table 3), it is hard to balance different tasks, *e.g.* zero-shot performance on ELEVATER improves while zero-shot performance on IN-1K worsens compared to [36]. Combining the large text corpus $\mathcal{T}$ and the prompt sentences of downstream class names is a good practice for task-aware distillation.

4.4. Influence of Dataset Scale

We investigate the influence of image and text datasets' scale on transferability and robustness of the student model by fixing the dataset scale of one modality and varying the other. For the fixed-size modality, we use images or text from "IN-21K + GCC-15M".

From Fig. 3 (a), we find that the transferability of student foundation models improves with larger image or text corpus, but it is more sensitive to the image corpus size. Also, prompt sentences with IN-21K class names describe diverse visual concepts, so training with these achieves comparable transferability to training with 3M captions from GCC-3M.

In Fig. 3 (b), we study the correlation between robustness

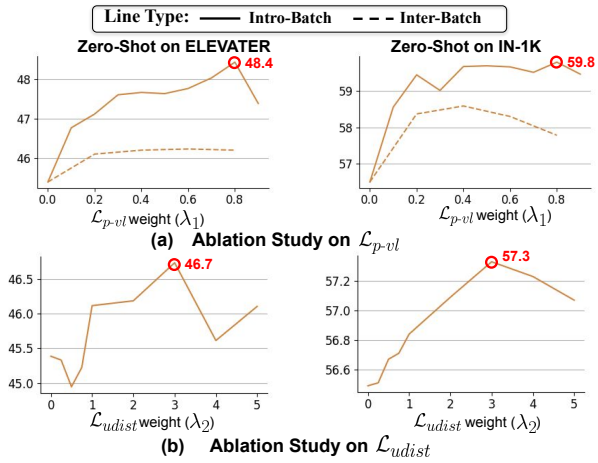


Figure 4: **Ablation Studies on $\mathcal{L}_{p\text{-}vl}$ and $\mathcal{L}_{udist}$.**

on IN-1K variant datasets and original performance on IN-1K, as well as the correlation between robustness and the size of image/text corpus. Fig. 3 (b.i) shows that when we fix the text corpus size, robustness correlates with the training image corpus size more strongly than with IN-1K performance. Even though distilling directly with IN-1K images produces better performance on IN-1K, it does not guarantee better robustness to domain shifts from IN-1K. In Fig. 3 (b.ii), we freeze the image corpus size and find a different trend, in which robustness directly relates to performance on IN-1K regardless of the text corpus size.

We conclude that VL distillation methods should focus on increasing the training image set to achieve better transferability and robustness. If downstream class names are unknown, it is critical to construct a text corpus that covers more visual concepts. If downstream class names are known, using them during distillation greatly benefits robustness.

4.5. Ablation Studies on Losses

We examine the effect of $\mathcal{L}_{p\text{-}vl}$ and $\mathcal{L}_{udist}$ by Zero-Shot on ELEVATER and IN-1K with IN-21K images and part of GCC-3M captions as the training data. See Supplementary Material Sec. G for more ablation studies on losses.

Ablation on $\mathcal{L}_{p\text{-}vl}$. We gradually put more weights on the $\mathcal{L}_{p\text{-}vl}$ by increasing λ_1 in Eq. 6 from 0 to 1 in Fig. 4 (a). We compare the inter-batch version (i.e. $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ in Eq. 10 from different batches) and intro-batch version (i.e. $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ from the same batch) of $\mathcal{L}_{p\text{-}vl}$ and find the intro-batch $\mathcal{L}_{p\text{-}vl}$ performs better than inter-batch $\mathcal{L}_{p\text{-}vl}$, so we keep intro-batch version in other experiments. Furthermore, adding $\mathcal{L}_{p\text{-}vl}$ with $\lambda_1 \leq 0.9$ brings better zero-shot performance than only using $\mathcal{L}_{vl}$ in all three settings. However, we observe the dramatic performance drop when we totally replace $\mathcal{L}_{vl}$ with $\mathcal{L}_{p\text{-}vl}$ (*i.e.* $\lambda_1 = 1$). We argue improvement with smaller λ_1's and drop at $\lambda_1 = 1$ both due to the gap between images and text embeddings in the shared feature space (More analysis in Sec. 4.6.).

Text Corpus	Zero-Shot		Linear Probing
	ELEVATER	IN-1K	ELEVATER
GCC-3M (Text)	38.6%	39.0%	**68.2%**
Unfiltered NLP (3M)	35.9%	33.2%	65.2%
Our Constructed $\mathcal{T}$ (3M)	**40.4%**	**39.2%**	67.7%

Table 4: **Distillation with Different Text Corpora of the Same Size.** Images from GCC-3M serve as the image dataset.

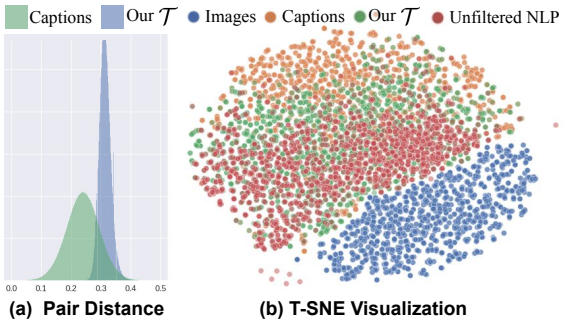


Figure 5: **Analysis of Text Corpus selected from the ROBERTa NLP Corpus.** Best viewed in color.

Ablation on $\mathcal{L}_{udist}$. We increase λ_2 from 0 to 5, to introduce $\mathcal{L}_{udist}$ as a regularization term. Generally, $\mathcal{L}_{udist}$ benefits the Zero-Shot on ELEVATER since it tries to preserve the geometry of image features. $\mathcal{L}_{udist}$ slightly improves IN-1K performance when λ_2 is small but it quickly harms IN-1K performance when λ_2 gets larger. We suspect the poor student embedding in early training along with the large regularization term detours the gradient decent trajectory. We find $\mathcal{L}_{udist}$ is less effective than the similar $\mathcal{L}_{p\text{-}vl}$, so we only use $\mathcal{L}_{udist}$ as a regularizer. Our main experiment (Table. 1) further shows $\mathcal{L}_{udist}$ is less effective when applying $\mathcal{L}_{p\text{-}vl}$ and $\mathcal{L}_{udist}$ together.

4.6. Analysis of Constructed Text Corpus

Text Corpus based on GCC-3M [63] Images. We first compare the distillation performance of our constructed $\mathcal{T}$ with the original GCC-3M captions and with randomly sampled NLP sentences in Table 4. $\mathcal{T}$ is constructed using Algorithm 1 on the large-scale NLP corpus and the GCC-3M image set. We find that our constructed $\mathcal{T}$ yields better zero-shot and compatible linear-probing performance compared to the original GCC-3M Captions, while the unfiltered NLP corpus at the same size performs poorly.

Pair-level Analysis. We analyze the quality of image-text pairs via the similarity score computed with CLIP-ViT-L/14 (teacher model). In Fig. 5 (a), we compute the histogram of similarity scores for image-caption pairs in GCC-15M and YFCC-14M. We also compute the same similarity score histogram for our selected sentence with its query image. Our sentences selected from 1.5B candidate sentences yield higher average similarity score than the human annotations. See our visualization for the selected text and the effect of each NLP dataset in Supplementary Material Sec. B-C.

Distribution-level Analysis. We also analyze the distribution of our constructed $\mathcal{T}$ in the shared feature space. We plot T-SNE (see Fig. 5 (b)) of normalized embeddings for samples from four different corpora: images, human-annotated captions, our selected sentences and random sentences from the NLP corpus. Even though original CLIP models [55] yield astonishing zero-shot performance on a large variety of downstream tasks, the images and their human annotated captions surprisingly do not overlap in the T-SNE visualization (We also provide MMD among these four corpora in Supplementary Material Sec. F). We conclude that the contrastive loss used in CLIP only pushes the text closer to its related image does not close the distribution gap between image and text corpora in feature space. This explains the effectiveness of $\mathcal{L}_{p\text{-}vl}$ where we use the image embeddings (Blue dots in Fig. 5 (b)) as the pseudo text embeddings. $\mathcal{L}_{p\text{-}vl}$ expands the text feature space and unsurprisingly leads to better performance. When we completely replace $\mathcal{L}_{vl}$ with $\mathcal{L}_{p\text{-}vl}$, the distillation performance drops a lot due to the large gap between the image and text modalities in the feature space. Moreover, the distributions of our selected $\mathcal{T}$ and the human-annotated captions are more similar. On the other hand, samples of the ROBERTa NLP Corpus have a different distribution from the visually-grounded sentences. These results provide further support for our Algorithm 1.

5. Conclusion

In this paper, we propose a vision-language knowledge distillation mechanism **DIME-FM** that distills knowledge in pre-trained VLFMs to small foundation models, without using any paired image-text data. We distill pre-trained CLIP-ViT-L/14 to our Distill-ViT-B/32 model, with only 40M public images and 28.4M unpaired text and our model rivals the CLIP-ViT-B/32 model that was pretrained on private large-scale WiT dataset in both transferability to novel tasks and robustness to natural domain shifts. Particularly, we propose an efficient text selection algorithm and two novel distillation losses for vision-language knowledge distillation. This paper shows how to achieve a small custom foundation model with limited unpaired data and released huge CLIP foundation models, which is the first trial in distilling a multi-modal foundation model while preserving its foundation properties. There are many interesting directions not covered in this paper and left for exploration in the future, such as VL distillation with large-scale paired image-text data (*e.g.* 400M+) and distillation of foundation models of other multi-modalities (*e.g.* video-language).

References

[1] Ellie Pavlick Stefanie Tellex Aaron Gokaslan*, Vanya Cohen*. Openwebtext corpus. 6

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 2

[3] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019. 2

[4] 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. 2

[5] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014. 2

[6] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 2

[7] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[8] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 6

[9] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10925–10934, 2022. 2

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020. 3

[12] 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. 6

[13] Hongwei Chen, Douglas Hendry, Phillip Weinberg, and Adrian Feiguin. Systematic improvement of neural network quantum states using lanczos. *Advances in Neural Information Processing Systems*, 35:7490–7503, 2022. 1

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3

[15] 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. *arXiv preprint arXiv:2212.07143*, 2022. 1, 2, 3

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1, 2, 6

[17] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 2

[18] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *European Conference on Computer Vision*, pages 247–264. Springer, 2022. 2

[19] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, 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. *arXiv preprint arXiv:2208.12262*, 2022. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[21] Wikimedia Foundation. Wikimedia downloads. 6

[22] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700. IEEE, 2013. 6

[23] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6589–6598, 2017. 2

[24] Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223, March 2017. 6

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[26] 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. 1

[27] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the*

European Conference on Computer Vision (ECCV), pages 771–787, 2018. 2

[28] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 6

[29] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[30] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 2

[31] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2, 3

[32] Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2308.01890*, 2023. 3

[33] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 2

[34] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[35] 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. 2

[36] Byoungjip Kim, Sungik Choi, Dasol Hwang, Moontae Lee, and Honglak Lee. Transferring pre-trained multimodal representations with cross-modal similarity matching. In *Advances in Neural Information Processing Systems*. 1, 2, 7

[37] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 2

[38] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[39] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020. 6

[40] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*, 2022. 2, 3, 5, 6

[41] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2, 3

[42] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[43] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022. 2

[44] 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. 2

[45] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1

[46] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[49] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[50] Christian M Meyer and Iryna Gurevych. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012. 2

[51] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998. 2

[52] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020. 2

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[54] 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. 1, 2

[55] 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, 6, 9

[56] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[58] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 6

[59] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2

[60] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[61] 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. *arXiv preprint arXiv:2210.08402*, 2022. 1

[62] 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 preprint arXiv:2111.02114*, 2021. 2

[63] 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. 5, 6, 8

[64] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, Kurt Keutzer, Trevor Darrell, Anna Rohrbach, and Jianfeng Gao. K-LITE: Learning transferable visual models with external knowledge. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 2

[65] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 4

[66] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[67] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3

[68] Ximeng Sun, Rameswar Panda, Chun-Fu (Richard) Chen, Aude Oliva, Rogerio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7375–7385, October 2021. 1

[69] Ximeng Sun, Rameswar Panda, Chun-Fu Richard Chen, Aude Oliva, Rogerio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7375–7385, 2021. 2

[70] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[71] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 6

[72] Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018. 6

[73] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. 6

[74] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[75] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14733–14743, 2022. 2

[76] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*, 2022. 1, 3

[77] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[78] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022. 1, 2

[79] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 1, 2

[80] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[81] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European conference on computer vision*, pages 451–466. Springer, 2016. 2

[82] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 1, 2, 5, 6

[83] 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. 2

[84] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1, 2

[85] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2

[86] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[87] Xinyu Zhang, Jiahui Chen, Junkun Yuan, Qiang Chen, Jian Wang, Xiaodi Wang, Shumin Han, Xiaokang Chen, Jimin Pi, Kun Yao, et al. Cae v2: Context autoencoder with clip target. *arXiv preprint arXiv:2211.09799*, 2022. 1, 2

[88] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 2

[89] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 6