Enhancing Fine-Grained Vision-Language Pretraining with Negative Augmented Samples

Yeyuan Wang [1*], Dehong Gao [2*], Lei Yi [3], Linbo Jin [3], Jinxia Zhang [4], Libin Yang [2†], Xiaoyan Cai [1†]

[1]School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[2]School of Cybersecurity, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[3]Alibaba Group, Hangzhou, Zhejiang, China
[4]School of Automation, Southeast University, Nanjing, Jiangsu, China
{wangyeyuan, dehong.gdh, libiny, xiaoyanc}@nwpu.edu.cn
{yilei.yi, yuyi.jlb}@alibaba-inc.com
{jinxiazhang}@seu.edu.cn

Abstract

Existing Vision-Language Pretraining (VLP) methods have achieved remarkable improvements across a variety of vision-language tasks, confirming their effectiveness in capturing coarse-grained semantic correlations. However, their capability for fine-grained understanding, which is critical for many nuanced vision-language applications, remains limited. Prevailing VLP models often overlook the intricate distinctions in expressing different modal features and typically depend on the similarity of holistic features for cross-modal interactions. Moreover, these models directly align and integrate features from different modalities, focusing more on coarse-grained general representations, thus failing to capture the nuanced differences necessary for tasks demanding a more detailed perception. In response to these limitations, we introduce Negative Augmented Samples(NAS), a refined vision-language pretraining model that innovatively incorporates NAS to specifically address the challenge of fine-grained understanding. NAS utilizes a Visual Dictionary(VD) as a semantic bridge between visual and linguistic domains. Additionally, it employs a Negative Visual Augmentation(NVA) method based on the VD to generate challenging negative image samples. These samples deviate from positive samples exclusively at the token level, thereby necessitating that the model discerns the subtle disparities between positive and negative samples with greater precision. Comprehensive experiments validate the efficacy of NAS components and underscore its potential to enhance fine-grained vision-language comprehension.

Introduction

Multi-modal machine learning, which aims to process and relate information of multiple modalities, is a domain that has a significant impact on general artificial intelligence (Li 2022). Among these modalities, there has been a surge of interest in vision-language research, as the vision and language modalities are widely used and closely intertwined in human daily life (Wang et al. 2022a). However, current vision-language pretraining models mainly focus on capturing the overall relationship between vision and language (Wang et al. 2022b; Ji et al. 2023), often overlooking the more nuanced, local interactions (Gao et al. 2020; Wei et al. 2021). The ability of modeling the local relationship, which we refer to as fine-grained capability, is crucial in various artificial intelligence domains such as medicine (Wang et al. 2021), agriculture (Van Horn et al. 2018), and e-commerce (Bai et al. 2020). Therefore, it is necessary to conduct in-depth research to better understand and model the fine-grained attributes of vision and language modalities.

Achieving this goal relies on two pivotal advancements: **fine-grained feature extraction** (the accurate extraction of subtle information from each input modality), and **fine-grained modality alignment** (the precise calibration of multi-modal features). According to **fine-grained feature extraction**, the discrete tokens are frequently selected as the fine-grained text features (Devlin et al. 2019) for language modality, while various image features (from the single pixels to patches or region features) are selected for visual modality (Huang et al. 2020; Chen et al. 2020; Zeng, Zhang, and Li 2022). For example, salient visual regions can be located with pretrained object detectors as visual region features (Hu et al. 2022). The kaleidoscope-like patches are leveraged to represent multi-scale visual features (Zhuge et al. 2021). These image features have propelled the advancement of VLP in the general domain; however, researchers still struggle to achieve fine-grained capability due to the huge semantic gaps between those **discrete** language tokens and these **continuous** visual features (Zhao et al. 2023). According to **fine-grained modality alignment**, researchers apply contrastive learning, such as CLIP (Radford et al. 2021), to align images and sentences globally. The later extensions explored patch-token interactive methods to capture the fine-grained correlation (Yao et al. 2021). Recently, researchers attempted to improve the fine-grained capability through Negative Textual Augmentation(NTA) (Yuksekgonul et al. 2022; Huang et al. 2024). As shown in Figure 1 (a), these NTA approaches employ either auxiliary models

[*]These authors contributed equally and [†]corresponding authors.

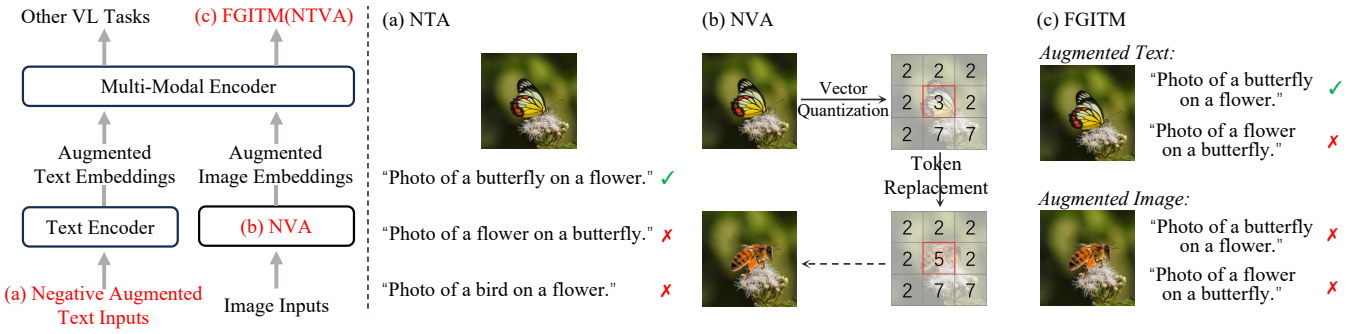


Figure 1: Fine-grained enhanced VLP architecture. NTA constructs the hard negative text samples for the language modality(a); We discrete the visual representation and construct the hard negative image samples for the visual modality(b); FGITM is proposed to leverage the fine-grained negative image and text samples to enhance the fine-grained capability(c).

or syntax-based algorithms to generate negative text samples misaligned with corresponding images (Yuksekgonul et al. 2022; Huang et al. 2024). Alongside the Image-Text Matching (ITM) task, these hard negative samples reinforce the VLP model to align the language and visual modalities fine-grainedly. Although the construction of hard negative samples for language modality is well-documented due to the sparsity of the text space and advancements in Large Language Models (LLMs) (Liu, Emerson, and Collier 2023), the construction for visual modality is hindered by the complexity of visual signals (Peng et al. 2024).

This paper addresses the issue of the Negative Textual and Visual Augmentation(NTVA) method, which creates hard negative samples for both language and visual modality. To tackle these challenges, we propose NAS, an innovative multi-modal model. As shown in Figure 1 (b), we integrate a VD into VLP model, which is regarded as the semantic abstraction of visual raw features, to bridge the semantic gaps between modalities. We further introduce a novel NVA approach, leveraging semantic-aware token replacement based on the VD to foster fine-grained alignment by constructing negative image samples. As shown in Figure 1 (b) and Figure 1 (c), our NAS creates negative image samples by altering tokens based on the global and local feature similarities of text and image inputs. These samples provoke the VLP models to pay more attention to detail alignment through FGITM task. Experimental results on downstream fine-grained multi-modal tasks demonstrate NAS's superior performance, significantly outperforming existing VLP models. In summary, our **contributions** are threefold:

- We first propose the NTVA method to simultaneously construct hard negative textual and visual samples, which can significantly improve the fine-grained capability together with the FGITM task. The NTVA method is a general data construction method that can be applied in related image fine-grained tasks.
- We introduce a novel VLP model named NAS, which applies the NTVA method to VLP models. Using the ALBEF structure as framework, NAS significantly improves the fine-grained capability of VLP models.
- Through comprehensive experiments on the ARO, Winoground, and VALSE datasets, we substantiate the efficacy of NAS. The results confirm that our proposed NTVA approach sets a new SOTA in these datasets.

Related Work

Visual Dictionary in Vision-Language

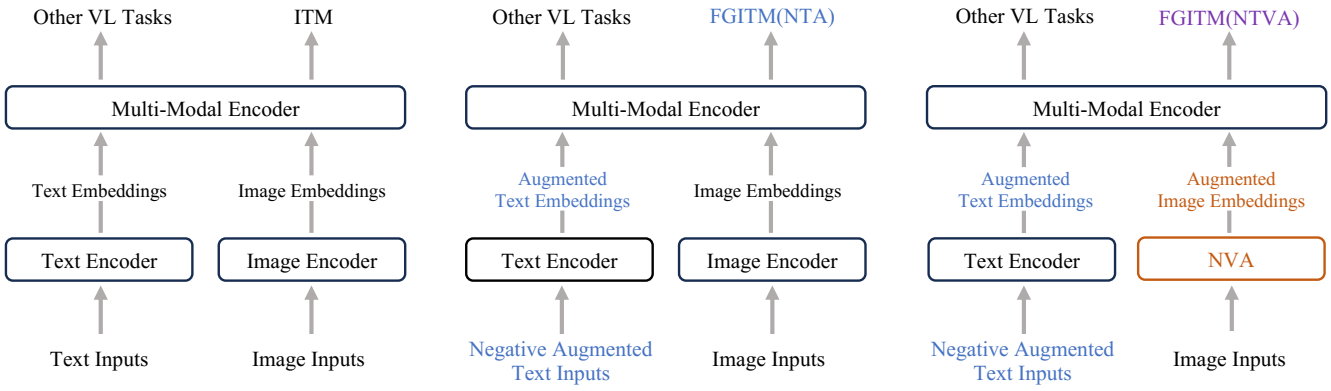
Within the landscape of multi-modal research, the use of Visual Dictionary (VD) has begun to play a pivotal role. SOHO (Huang et al. 2021) leverages the VD to address semantic discrepancies. Similarly, UNIMO2 (Li et al. 2022b) employs the VD as a cornerstone for modality alignment, effectively utilizing both uni-modal and multi-modal data streams. FDT (Chen et al. 2023) quantizes multi-modal features through a unified VD, reinforcing the alignment across modalities. Moreover, IL-CLIP (Zheng et al. 2024) introduces an iterated learning algorithm based on the unified VD to improve compositionality in VLP model.

Our approach employs the VD to bridge the semantic gap between different modal features. We address mode collapse in VD learning by updating the dictionary with an exponential moving average mechanism, which improves both VD learning and training stability and setting the stage for implementing our proposed NVA method.

Data Augmentation for Vision-Language

Data augmentation (DA) is widely applied in computer vision and has expanded into the realm of VLP (Mu et al. 2022; Li et al. 2022c). Recent studies employ NTA to construct fine-grained hard negative sentences with similar structures but different semantics (Yuksekgonul et al. 2022). In the visual domain, Syn-CLIP (Cascante-Bonilla et al. 2023) exploits 3D simulation engines to bolster conceptual understanding. SPEC (Peng et al. 2024) combines SAM (Kirillov et al. 2023) and stable diffusion (Rombach et al. 2022) to generate fine-grained negative image samples. However, these methods are hindered by the complexity of data generation and the potential for synthetic data to skew the consistency of the data distribution.

Our approach constructs negative image samples without relying on external models. As shown in Figure 2 (c), we capitalize on the VD embedded within our model to semantically modify input images in a novel end-to-end manner.



(a) Mainstream VLP methods (b) Text augmentation-based methods (c) NAS (Ours)

Figure 2: Comparison of our fine-grained NAS to other VL frameworks. Mainstream VLP methods utilize two "Dual Tower" encoders and a multi-modal encoder for deep fusion of multi-modal features(e.g., ALBEF (Li et al. 2021) and METER (Dou et al. 2022))(a), NTA-based methods construct augmented negative text samples to enhance VLP model's fine-grained ability with FGITM(e.g., VL-Match (Bi et al. 2023) and ViLTA (Wang et al. 2023))(b), our NAS introduces a NVA module to construct augmented negative image features, together with NTVA to enhance the VLP fine-grained capability with FGITM in an end-to-end manner(c).

Method

This section details our NAS architecture, the NVA module, and the FGITM pretraining task.

Model Architecture

Given an image-text pair (I, T), the image I is encoded into embeddings $v_{\text{cls}}, v_1, ..., v_N$, where v_{cls} is the `[CLS]` token's embedding, and N denotes the number of image patches. The text T is similarly transformed into embeddings $t_{\text{cls}}, t_1, ..., t_M$, with t_{cls} corresponding to the text's `[CLS]` token embedding, and M indicating the language encoder's maximum sequence length. For visual features, all embeddings, except for the `[CLS]` token, are quantized into discrete tokens based on the VD and then concatenated with the `[CLS]` token to form an enhanced visual representation. Our pretraining comprises two distinct stages. In the first stage, the quantized image embeddings are integrated with the encoded text embeddings through cross-attention mechanisms within the multi-modal encoder. In the second stage, the quantized image embeddings are employed to acquire token-level negative image samples via our NVA module. These samples, alongside positive image inputs, are fed into the multi-modal encoder. The multi-modal encoder's output serves to pretrain and fine-tune downstream tasks.

Negative Visual Augmentation Module

In this module, we introduce the VD as a fundamental component, which functions as a quantization framework to generate negative image samples. Formalized as a matrix $\mathcal{D} \in \mathbf{R}^{m \times c}$, it comprises m vectors, each of dimension c. Initially randomized, the dictionary is progressively refined through a moving average process over mini-batches. The process of associating each visual feature v_i with an embedding vector in the dictionary d_j is defined by:

$$h_i = \underset{d_j \in \mathcal{D}}{\text{argmin}} \quad \|v_i - d_j\|_2, \tag{1}$$

Updates to the VD within a mini-batch follow the equation:

$$\hat{d}_j = \gamma * d_j + (1 - \gamma) * \frac{\sum_{h_i = j} v_i}{n}, \tag{2}$$

with $\hat{d}_j$ representing the updated vector, γ functions as the momentum coefficient (ranging from $[0, 1]$), and n is the count of visual patches mapped to d_j within the current mini-batch—updating only when $n \neq 0$. Since the argmin operation is non-differentiable, we employ the stop-gradient operation to facilitate the visual encoder's training:

$$\hat{v}_i = \text{sg}[d_{h_i} - v_i] + v_i, \tag{3}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, h_i is an index in $\mathcal{D}$ and v_i is subsequently assigned the value of d_{h_i}.

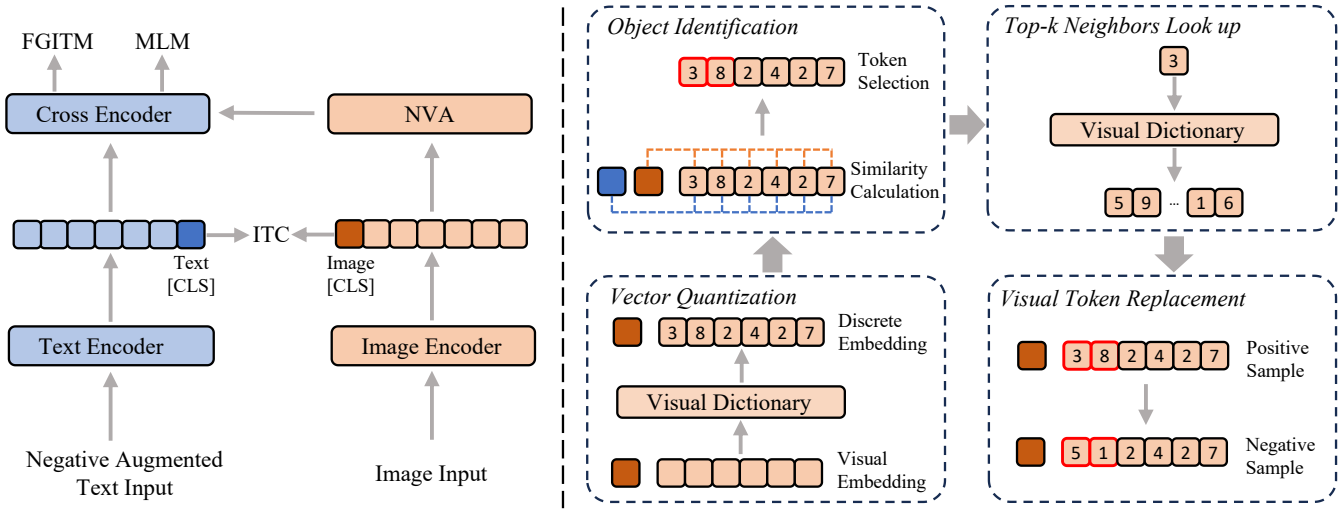
For the generation of negative image samples, we utilize the textual global feature T_{cls}, the visual global feature V_{cls}, and the visual local features v_i. Departing from conventional methods that rely solely on either T_{cls} or V_{cls} tokens for object identification (Liang et al. 2022; Jiang et al. 2022), our approach synthesizes both to improve accuracy. We identify the primary object in an image by calculating a weighted sum S of the cosine similarity S_{t} between T_{cls} and all local visual features v_i, and S_{v} between V_{cls} and v_i:

$$S_{\text{t}} = cosine\langle T_{\text{cls}}, v_i \rangle, \quad i = 1, \ldots, N \tag{4}$$

$$S_{\text{v}} = cosine\langle V_{\text{cls}}, v_i \rangle, \quad i = 1, \ldots, N \tag{5}$$

$$S = \lambda S_{\text{t}} + (1 - \lambda) S_{\text{v}} \tag{6}$$

The hyper parameter λ is used to control the weight of S_{t} and S_{v}. In the experiment, we set it to 0.5. The 30% tokens with the highest similarity score S are identified as the primary subject of the image. We then randomly replace these



(a) Overall Architecture

(b) Negative Visual Augmentation(NVA)

Figure 3: (a) The framework of the proposed end-to-end pretraining model NAS. (b) Illustration of our NVA. The continuous visual embedding encoded by the image encoder is firstly quantified into discrete embedding and then identifies the object in the image embedding based on the similarity between the global `[CLS]` embeddings and local discrete embeddings. We use the object embedding to search top-k neighbors in the dictionary and replace them with the neighbor tokens to construct negative image samples. [Best viewed in color.]

tokens with the $Top-k$ most similar tokens from the VD to construct token-level negative image samples. Specifically, we select the $Top-k$ embeddings in VD that are most similar to the current token(excluding itself), and randomly select one of them according to probability. Tokens with the same quantified index are replaced with the same embedding. In the experiment, we set k to 3. Additionally, we incorporate 2-D sinusoidal positional embeddings using a sine function to enhance the model's spatial context comprehension. These negative image samples compel the encoder to recognize and encode subtle nuances, crucial for tasks requiring granular visual-textual discernment.

Pretraining Tasks

Fine-Grained Image-Text Matching ITM predicts whether a given image-text pair is positive (matched) or negative (not matched), which is a binary classification task. Based on ITM, the proposed FGITM aims to capture fine-grained differences of image-text pairs. For each input image-text pair, we use two types of negative image samples: an in-batch negative image sample selected according to the similarity of images and texts in each mini-batch(the non-matching image with the highest similarity is selected as a negative image sample), and a token-level negative sample generated using the NVA module. We use the multi-modal encoder's output embedding of the `[CLS]` token as the joint representation of the image-text pair, and append a classification layer to predict the image-text matching probability p^{itm}. The FGITM loss is a cross-entropy loss:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T)\sim D}\mathcal{H}(y^{\text{itm}}, p^{\text{itm}}(I', T')) \tag{7}$$

where y^{itm} is a one-hot vector representing the ground-truth label. (I', T') includes (I, T), (I, T^{neg}), (I, T^{fg}), (I^{neg}, T) and (I^{fg}, T), where I^{neg}/T^{neg} is the negative image/text selected in every training batch and I^{fg}/T^{fg} is the fine-grained negative image/text.

Image-Text Contrastive Learning We follow the same settings of the ITC loss in ALBEF (Li et al. 2021). Specifically, the similarity between image and text is calculated by the similarity function $s(I,T) = l_v(v_{\text{cls}})^\top l_t(t_{\text{cls}})$, where l_v and l_t are linear transformations that consist of a linear layer and a normalization layer. These transformations map v_{cls} and t_{cls} to normalized vectors in a reduced dimensional space. Two queues are maintained to cache the most recently obtained M image representations I_m and M text representations T_m, which are calculated by a momentum text encoder and a momentum image encoder, respectively. The normalized features obtained from the momentum model are denoted as $l'_v(v'_{\text{cls}})$ and $l'_t(t'_{\text{cls}})$. $s(I,T_m) = l_v(v_{\text{cls}})^\top l'_t(t'_{\text{cls}})$ and $s(T,I_m) = l_t(t_{\text{cls}})^\top l'_v(v'_{\text{cls}})$ define the similarity between the positive representations from the pretraining encoders and the negative representations from the momentum encoders. For each image and text, we compute the softmax-normalized image-to-text and text-to-image similarities as:

$$p^{\text{i2t}}(I) = \frac{\exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(s(I,T_m)/\tau)} \tag{8}$$

$$p^{\text{t2i}}(T) = \frac{\exp(s(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(s(T,I_m)/\tau)} \tag{9}$$

where τ is a learnable temperature parameter.

Momentum distillation leverages the momentum model to generate pseudo-targets and guides the training model to

Dataset	MSCOCO	VG	SBU	CC-3M	**Sum.**
# images	113K	100K	843K	1.81M	2.87M
# texts	567K	769K	843K	1.81M	4.00M

Table 1: Statistics of the pretraining datasets.

learn from these targets. The final targets are:

$$y^{\mathrm{i2t}}(I) = (1-\alpha)y^{\mathrm{i2t}}_{\mathrm{one-hot}}(I) + \alpha p^{\mathrm{i2t}}(I_m) \tag{10}$$

$$y^{\mathrm{t2i}}(T) = (1-\alpha)y^{\mathrm{t2i}}_{\mathrm{one-hot}}(T) + \alpha p^{\mathrm{t2i}}(T_m) \tag{11}$$

where $y^{\mathrm{i2t}}_{\mathrm{one-hot}}(I)$ and $y^{\mathrm{t2i}}_{\mathrm{one-hot}}(T)$ denote the ground-truth one-hot similarity.

The ITC loss over the pretraining dataset D is defined as the cross-entropy $\mathcal{H}$ between p and y:

$$\mathcal{L}_{\mathrm{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\big[\mathcal{H}(y^{\mathrm{i2t}}(I), p^{\mathrm{i2t}}(I)) + \mathcal{H}(y^{\mathrm{t2i}}(T), p^{\mathrm{t2i}}(T))\big] \tag{12}$$

Mask Language Modeling Masked Language Modeling (MLM) utilizes both the image and the contextual text to predict the masked words. We randomly mask out the input tokens with a probability of 15% and replace them with the special token `[MASK]` (following BERT, the replacements are 10% random tokens, 10% unchanged, and 80% `[MASK]`). Let $\hat{T}$ denotes a masked text, and $p^{\mathrm{msk}}(I, \hat{T})$ denotes the predicted probability for a masked token. MLM minimizes a cross-entropy loss:

$$\mathcal{L}_{\mathrm{mlm}} = \mathbb{E}_{(I,\hat{T})\sim D}\mathcal{H}(y^{\mathrm{msk}}, p^{\mathrm{msk}}(I, \hat{T})) \tag{13}$$

where y^{msk} is the one-hot vocabulary distribution.

The full pretraining objective of NAS is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{itc}} + \mathcal{L}_{\mathrm{mlm}} + \mathcal{L}_{\mathrm{itm}} \tag{14}$$

Experiments

Pretraining Setup and Baselines

Pretraining Setup We use COCO (Lin et al. 2014), Visual Genome (VG) (Krishna et al. 2017), Conceptual Captions (CC) (Sharma et al. 2018), and SBU Captions (Ordonez, Kulkarni, and Berg 2011) as our pretraining datasets, which have a total of 4 million unique images and 5.1 million image-text pairs. However, only 2.9 million images and 4 million image-text pairs are currently available, as presented in Table 1. We leverage the initial six layers of $\mathrm{BERT}_{\mathrm{base}}$ to initialize the text encoder, the subsequent six layers to initialize the multi-modal encoder, and DEiT-224/16 to initialize the image encoder. The number of VD elements is set to 2,048. In the NVA module, we set the balance parameter λ to 0.5 and the parameter k to 3. For NTA and NTVA, to verify that our approach can work synergistically with existing methods, we fine-tune our model on the text-augmented COCO dataset (Yuksekgonul et al. 2022). Pretraining unfolds over 29 epochs in the first stage and a single epoch in the second stage, utilizing a batch size of 512. We adopt the AdamW optimizer with a weight decay of 0.02. In the first 1000 iterations, the learning rate is warmed-up to $1e^{-4}$, and decayed to $1e^{-5}$ following a cosine schedule. Each image is randomly cropped to 256×256 resolution, and RandAugment (Cubuk et al. 2020) is adopted. During the fine-tuning stage, the resolution of an image is up-scaled to 384×384, and the positional encoding of the image patches is interpolated. The momentum parameter for updating the momentum model is 0.995, and the queue length of cached features for ITC task is set as 65, 536. We linearly ramp-up the distillation weight α from 0 to 0.4 within the 1st epoch. All experiments are performed on 8 NVIDIA A800 GPUs.

Benchmarks We evaluated our method on three benchmarks. ARO (Yuksekgonul et al. 2022) is designed for evaluating VLP models' object relational understanding and sensitivity to perturbations. Winoground (Thrush et al. 2022) is a small dataset for evaluating compositional reasoning. VALSE (Parcalabescu et al. 2022) is designed for testing VLP models' visio-linguistic grounding capabilities.

Baselines For multi-modal models, we evaluate LXMERT (Tan and Bansal 2019), ViLBERT (Lu et al. 2019), UNITER (Chen et al. 2020), ViLT (Kim, Son, and Kim 2021), CLIP (Radford et al. 2021), ALBEF (Li et al. 2021), XVLM (Zeng, Zhang, and Li 2022), FLAVA (Singh et al. 2022), NegCLIP (Yuksekgonul et al. 2022), syn-CLIP (Cascante-Bonilla et al. 2023), SPEC (Peng et al. 2024), FDT (Chen et al. 2023), BLIP2 (Li et al. 2023), MiniGPT-4 (Zhu et al. 2023) and LLaVA (Liu et al. 2024). Among these, BLIP2, MiniGPT-4 and LLaVA are currently the most prominent multi-modal large language models. NegCLIP, syn-CLIP, and SPEC are based on hard negatives, while FDT is a VD-based model. For large language models, we compare our model with BART (Yuan, Neubig, and Liu 2021) and FLAN-T5 (Chung et al. 2024).

Enhancement on Fine-grained Capability

The evaluation includes results for the ARO (Table 2), Winoground (Table 2), and VALSE benchmark (Table 3). For NVA, we present fine-tuning results on the COCO dataset, which is denoted as $\mathrm{NAS}_{\mathrm{COCO}}$. For NTA and NTVA, we fine-tune our model on the text-augmented COCO dataset (Yuksekgonul et al. 2022). For a fair comparison, we reproduce ALBEF on the training dataset we used.

Notably, despite using less training data, NAS surpasses existing methods across benchmarks, outperforming the ALBEF baseline by 2.9% on VALSE, 7.2% on Winoground, and 18.8% on ARO. These gains highlight the model's ability to leverage hard negatives in both images and text effectively. NVA significantly enhances fine-grained capability, while NTVA outperforms both NTA and NVA, demonstrating the compatibility between NVA and NTA methods.

NTA performs particularly well on the ARO benchmark due to the dataset is curated for specific hard text negatives. Conversely, NVA's visual token replacement enhances attention to image details, achieving strong results on VALSE and Winoground. By introducing NTVA, the fine-grained feature extraction and fine-grained modality alignment of the model are significantly enhanced. As shown in Figure 4, our model adeptly discerns subtle details (e.g., discerning "sheep standing", "no dogs" and identifying "5 birds"), showcasing re-

Model	#Images	ARO			Winoground			
		Relation	Attribute	**Avg.**	Text	Image	Group	**Avg.**
Random Chance	-		50		25.0	25.0	16.7	22.2
UNITER	4M	-	-	-	32.3	13.3	10.0	18.5
ViLT(ViT-B/32)	4M	39.5	20.3	29.9	34.8	14.0	9.3	19.3
CLIP	400M	59.0	62.0	60.5	30.8	10.5	8.0	16.4
FLAVA	60M	25.0	73.0	49.0	25.3	13.5	9.0	15.9
$\text{ALBEF}_{\text{COCO}}$	4M	60.5	88.5	74.5	27.5	15.8	11.0	18.1
Large language models								
BART	-	81.1	73.6	77.4	-	-	-	-
FLAN-T5	-	84.4	76.5	80.5	-	-	-	-
Large Multi-modal models								
BEIT3	35M	60.6	74.6	67.6	-	-	-	-
LLaVA-7B	400M	-	-	-	13.5	5.3	2.3	7.0
MiniGPT-4	500M	46.9	55.7	52.3	23.3	18.0	9.5	17.0
VD based models								
FDT	3M	49.8	54.6	52.2	17.3	3.5	1.5	7.4
Hard Negative based models								
NegCLIP	400M	80.2	70.5	75.4	29.5	10.5	8.0	16.0
syn-CLIP	401M	71.4	66.9	69.2	30.0	11.5	9.5	17.0
SPEC	400M	66.4	73.7	70.1	-	-	-	-
$\text{NAS(NTA)}_{\text{COCO}}$	2.9M	93.1	91.7	92.4	32.3	17.3	13.0	20.8
$\text{NAS(NVA)}_{\text{COCO}}$	2.9M	67.8	89.8	78.8	34.5	19.0	14.0	22.5
$\text{NAS(NTVA)}_{\text{COCO}}$	2.9M	**93.2**	**93.4**	**93.3**	**35.3**	**22.0**	**18.5**	**25.3**

Table 2: Results on the ARO and Winoground benchmark. The NTA method yields substantial improvements on the ARO benchmark since it adopts task-specific hard negative types.

Model	#Images	**Existence** quantifiers	**Plurality** number	**Counting**	**SP.rel.** relations	**Action**	**Coreference**	**Foil-it!**	**Avg.**
LXMERT	0.18M	78.6	64.4	58.0	60.2	50.3	45.5	87.1	63.5
ViLBERT	3.1M	65.5	61.2	65.1	57.2	69.5	47.7	86.9	64.7
CLIP	400M	66.9	56.2	60.7	64.3	72.1	50.9	88.8	65.7
$\text{ALBEF}_{\text{COCO}}$	2.9M	75.4	76.5	65.8	74.4	67.5	48.0	92.6	71.5
$\text{XVLM}_{\text{COCO}}$	4M	83.0	75.6	67.5	70.2	71.2	48.0	94.8	72.9
FDT	3M	64.0	56.8	51.2	51.8	61.5	47.3	79.6	58.9
BLIP2	500M	55.5	71.5	66.0	62.4	67.6	50.3	95.9	67.0
MiniGPT-4	500M	65.5	72.5	67.4	68.4	71.0	51.8	95.8	70.4
$\text{NAS(NTA)}_{\text{COCO}}$	2.9M	85.5	75.9	66.8	71.6	75.5	45.4	93.7	73.5
$\text{NAS(NVA)}_{\text{COCO}}$	2.9M	85.1	77.6	66.7	72.1	72.7	48.8	94.2	73.9
$\text{NAS(NTVA)}_{\text{COCO}}$	2.9M	87.3	77.6	70.1	69.9	75.8	46.7	93.2	**74.4**

Table 3: Results on the VALSE benchmark.

Model	VD	NTA	NVA	Winoground			ARO		VALSE	**Avg.**
				Text	Image	Group	Relation	Attribute		
NAS(wo/VD)				28.5	15.0	11.0	59.2	88.0	71.5	45.5
NAS(w/VD)	✓			34.8	15.3	13.8	64.7	88.7	72.2	48.3
NAS(NTA)	✓	✓		32.3	17.3	13.0	93.1	91.7	73.5	53.5
NAS(NTVA)	✓	✓	✓	**35.3**	**22.0**	**18.5**	**93.2**	**93.4**	**74.4**	**56.1**

Table 4: Ablation study of VD and NVA.

Figure 4: Cases on the VALSE benchmark. More examples are shown in Supplementary.

fined fine-grained feature extraction. In conclusion, NTVA consistently delivers the most compelling results across all benchmarks, cementing the superiority of integrating NVA with NTA. We also conducted a comparison on the broader retrieval task on COCO dataset. Utilizing less training data, our method demonstrates superior performance. Additionally, the exponential moving average update for VD results in the introduction of less than 1% of the total parameters. To validate the generalizability of our method, we applied it to BLIP (Li et al. 2022a), which also led to improved performance. Results are shown in Supplementary.

Ablations

We conducted ablation studies to assess the NVA module and its synergy with the NTA methods. The effectiveness of VD and NVA is assessed in Table 4, while the VD vector size m, visual token replacement ratio, and balance parameter λ are analyzed in Table 5, Table 6, and Table 7, respectively.

The introduction of the VD improved overall performance by enhancing fine-grained feature extraction. The integration of NTA further refined the fine-grained alignment, leading to significant improvements. Finally, the introduction of NVA demonstrated the effectiveness of the proposed method and verifying the synergy between our method and existing NTA methods. Our observations regarding the VD size revealed that a size of 2048 consistently achieved optimal results, aligning with the VD's design intent to consolidate similar visual semantics under unified image features. However, excessively granular semantic distinctions may hinder the extraction process of visual semantics and affect vision-language alignment. Conversely, a smaller VD size may impede fine-grained modality alignment. Empirically, a size of $m = 2048$ yielded the most favorable outcomes and has thus been adopted as the default configuration.

In the ablation experiment on the replacement ratio of visual tokens, a replacement ratio of 30% yielded the best results. We believe that a low visual token replacement ratio is not enough to constitute a negative visual sample, while a high replacement ratio will destroy the overall semantic information of the image and is not conducive to the convergence of the pretraining process. Furthermore, in the ablation study of the balance parameter λ, setting $\lambda = 0.5$ yielded optimal performance. This result suggests that simultaneously leveraging the global information from both the image and text is more effective for object identification within images. Our results demonstrate the effectiveness of the NVA module, its synergy with existing NTA methods, and its versatile applicability across a wide range of tasks.

Size	Winoground			VALSE	**Avg.**
	Text	Image	Group		
1024	31.3	**23.5**	14.8	72.5	35.3
2048	**35.3**	22.0	**18.5**	**74.4**	**37.6**
4096	29.8	23.3	15.3	71.7	35.0
8192	27.8	22.0	12.8	71.4	33.5

Table 5: Ablation study of the embedding vector size of VD.

Ratio	Winoground			VALSE	**Avg.**
	Text	Image	Group		
10%	30.3	22.8	16.5	72.6	35.6
30%	**35.3**	22.0	**18.5**	**74.4**	**37.6**
50%	34.0	**23.8**	17.5	73.4	37.2
70%	32.5	22.3	16.3	72.3	35.9

Table 6: Ablation of the ratio of visual token replacement.

Value	Winoground			VALSE	**Avg.**
	Text	Image	Group		
0.0	30.3	**24.0**	16.0	73.2	35.9
0.5	**35.3**	22.0	**18.5**	**74.4**	**37.6**
1.0	33.0	22.8	15.0	72.7	35.9

Table 7: Ablation study of the balance parameter λ.

Conclusions

In this paper, we propose the NTVA method to simultaneously construct hard negative textual and visual samples. The comprehensive experiments demonstrate the effectiveness of NAS and confirm that the NTVA method synergizes the hard negative samples, which greatly improves the fine-grained capability of NAS, setting a new SOTA in the field. The NTVA method is a general data construction method that can be applied in related image fine-grained tasks. In the future, we will investigate the co-quantize approach to align multi-modal information earlier and deeper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U20B2065, U22B2036, 62372380, and 62103374, National Key Research and Development Project under Grant 2022YFB3104005, and the Natural Science Basic Research Program of Shaanxi (Program No.2024JC-YBMS-513), and Key Research and Development Program of Zhejiang Province under Grants 2024C01025.

References

Bai, Y.; Chen, Y.; Yu, W.; Wang, L.; and Zhang, W. 2020. Products-10K: A Large-scale Product Recognition Dataset. arXiv:2008.10545.

Bi, J.; Cheng, D.; Yao, P.; Pang, B.; Zhan, Y.; Yang, C.; Wang, Y.; Sun, H.; Deng, W.; and Zhang, Q. 2023. VL-Match: Enhancing Vision-Language Pretraining with Token-Level and Instance-Level Matching. In *International Conference on Computer Vision*, 2584–2593.

Cascante-Bonilla, P.; Shehada, K.; Smith, J. S.; Doveh, S.; Kim, D.; Panda, R.; Varol, G.; Oliva, A.; Ordonez, V.; Feris, R.; et al. 2023. Going beyond nouns with vision & language models using synthetic data. In *International Conference on Computer Vision*, 20155–20165.

Chen, Y.; Yuan, J.; Tian, Y.; Geng, S.; Li, X.; Zhou, D.; Metaxas, D. N.; and Yang, H. 2023. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *International Conference on Computer Vision*, 15095–15104.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 104–120. Springer.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25: 1–53.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Conference on Computer Vision and Pattern Recognition*, 702–703.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition*, 18166–18176.

Gao, D.; Jin, L.; Chen, B.; Qiu, M.; Li, P.; Wei, Y.; Hu, Y.; and Wang, H. 2020. FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2251–2260. Xi'an, China.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *Conference on Computer Vision and Pattern Recognition*, 17980–17989.

Huang, Y.; Tang, J.; Chen, Z.; Zhang, R.; Zhang, X.; Chen, W.; Zhao, Z.; Zhao, Z.; Lv, T.; Hu, Z.; et al. 2024. Structure-CLIP: Towards Scene Graph Knowledge to Enhance Multi-Modal Structured Representations. In *AAAI Conference on Artificial Intelligence*, volume 38, 2417–2425.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Conference on Computer Vision and Pattern Recognition*, 12976–12985.

Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; and Fu, J. 2020. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. arXiv:2004.00849.

Ji, Y.; Tu, R.; Jiang, J.; Kong, W.; Cai, C.; Zhao, W.; Wang, H.; Yang, Y.; and Liu, W. 2023. Seeing what you miss: Vision-language pre-training with semantic completion learning. In *Conference on Computer Vision and Pattern Recognition*, 6789–6798.

Jiang, C.; Xu, H.; Li, C.; Yan, M.; Ye, W.; Zhang, S.; Bi, B.; and Huang, S. 2022. TRIPS: Efficient Vision-and-Language Pre-training with Text-Relevant Image Patch Selection. In *Conference on Empirical Methods in Natural Language Processing*, 4084–4096.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *International Conference on Computer Vision*, 4015–4026.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123: 32–73.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Neural Information Processing Systems*, 34: 9694–9705.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2022b. UNIMO-2: End-to-End Unified Vision-Language Grounded Learning. arXiv:2203.09067.

Li, X. 2022. Multimodal Cognitive Computing. *SCIENTIA SINICA Informationis*, 53.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022c. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. arXiv:2110.05208.

Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Xie, P.; et al. 2022. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Zurich, Switzerland: Springer.

Liu, F.; Emerson, G.; and Collier, N. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11: 635–651.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Neural Information Processing Systems*, 36.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Neural Information Processing Systems*, 32.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, 529–544. Springer.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. *Neural Information Processing Systems*, 24.

Parcalabescu, L.; Cafagna, M.; Muradjan, L.; Frank, A.; Calixto, I.; and Gatt, A. 2022. VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena. In *Annual Meeting of the Association for Computational Linguistics*, 8253–8280.

Peng, W.; Xie, S.; You, Z.; Lan, S.; and Wu, Z. 2024. Synthesize Diagnose and Optimize: Towards Fine-Grained Vision-Language Understanding. In *Conference on Computer Vision and Pattern Recognition*, 13279–13288.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2556–2565.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv:1908.07490.

Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The iNaturalist Species Classification and Detection Dataset. In *Conference on Computer Vision and Pattern Recognition*, 8769–8778. Salt Lake City, UT, USA.

Wang, L.; Hu, W.; Qiu, H.; Shang, C.; Zhao, T.; Qiu, B.; Ngan, K. N.; and Li, H. 2022a. A Survey of Vision and Language Related Multi-Modal Task. *CAAI Artificial Intelligence Research*, 1(2): 111–136.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022b. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *International Conference on Machine Learning*. Baltimore, Maryland, USA.

Wang, W.; Yang, Z.; Xu, B.; Li, J.; and Sun, Y. 2023. Vilta: Enhancing vision-language pre-training through textual augmentation. In *International Conference on Computer Vision*, 3158–3169.

Wang, X.; Lan, R.; Wang, H.; Liu, Z.; and Luo, X. 2021. Fine-Grained Correlation Analysis for Medical Image Retrieval. *Computers & Electrical Engineering*, 90: 106992.

Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2021. Fine-grained image analysis with deep learning: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8927–8948.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations*.

Yuan, W.; Neubig, G.; and Liu, P. 2021. Bartscore: Evaluating generated text as text generation. *Neural Information Processing Systems*, 34: 27263–27277.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

Zeng, Y.; Zhang, X.; and Li, H. 2022. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *International Conference on Machine Learning*, 25994–26009. PMLR.

Zhao, Z.; Guo, L.; He, X.; Shao, S.; Yuan, Z.; and Liu, J. 2023. MAMO: Fine-Grained Vision-Language Representations Learning with Masked Multimodal Modeling. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1528–1538.

Zheng, C.; Zhang, J.; Kembhavi, A.; and Krishna, R. 2024. Iterated learning improves compositionality in large vision-language models. In *Conference on Computer Vision and Pattern Recognition*, 13785–13795.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.

Zhuge, M.; Gao, D.; Fan, D.-P.; Jin, L.; Chen, B.; Zhou, H.; Qiu, M.; and Shao, L. 2021. Kaleido-bert: Vision-language pre-training on fashion domain. In *Conference on Computer Vision and Pattern Recognition*, 12647–12657.