

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

**Yeyuan Wang** <sup>1\*</sup>, **Dehong Gao** <sup>2\*</sup>, **Lei Yi** <sup>3</sup>, **Linbo Jin** <sup>3</sup>, **Jinxia Zhang** <sup>4</sup>,  
**Libin Yang** <sup>2†</sup>, **Xiaoyan Cai** <sup>1†</sup>

<sup>1</sup>School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi, China

<sup>2</sup>School of Cybersecurity, Northwestern Polytechnical University, Xi'an, Shaanxi, China

<sup>3</sup>Alibaba Group, Hangzhou, Zhejiang, China

<sup>4</sup>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 multi-modal 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, 2023a; 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 (Fang et al. 2020; Wang et al. 2021), agriculture (Hou, Feng, and Wang 2017; Van Horn et al. 2018), and e-commerce (Pang et al. 2019; Bai et al. 2020). Therefore, it is necessary to conduct in-depth research to better understand and model the fine-grained attributes of both 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; Kim, Son, and Kim 2021; Zeng, Zhang, and Li 2022). For example, salient visual regions can be located with pretrained object detectors as visual region features (Li et al. 2020; Cho et al. 2021; 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

\*These authors contributed equally and †corresponding authors.  
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

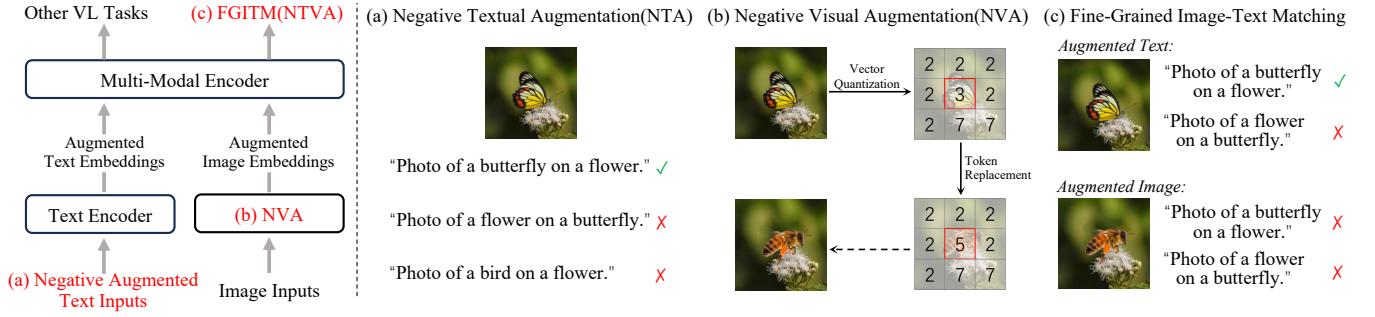


Figure 1: Fine-grained enhanced VLP architecture. NTA constructs the hard negative text samples for the language modality(a); We discrete the visual representation and construct the hard negative image samples for the visual modality(b); FGITM is proposed to leverage the fine-grained negative image and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

through Negative Textual Augmentation(NTA), which has made significant progress (Yuksekgonul et al. 2022; Zhang, Awal, and Agrawal 2023; Huang et al. 2024). As shown in Figure 1 (a), these NTA approaches employ either auxiliary models or syntax-based algorithms to generate negative text samples deliberately misaligned with corresponding images (Yuksekgonul et al. 2022; Doveh et al. 2023; Singh et al. 2023; 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) (Parcalabescu et al. 2022; 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 an innovative multi-modal model, called NAS. 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. The VD effectively quantifies continuous visual input into discrete tokens, easing fine-grained extraction and improving generalizability. Furthermore, we 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. To evaluate the effectiveness of our proposed model, we conducted experiments on downstream fine-grained multi-modal tasks. The results demonstrate that our model outperforms existing VLP models significantly. 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 AL-BEF 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

Visual Dictionary is widely used in VLP for transforming continuous visual features into discrete, higher-level representations (Huang et al. 2021; Chen et al. 2023; Zheng et al. 2024). This transformation aids in alleviating semantic gap and capturing more nuanced visual semantic concepts. Within the landscape of multi-modal research, the use of VD has begun to play a pivotal role (Li et al. 2022b; Chen et al. 2023). 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. The FDT framework (Chen et al. 2023) quantizes multi-modal features through a unified VD, further reinforcing the alignment across modalities. Moreover, IL-CLIP (Zheng et al. 2024) introduces an iterated learning algorithm based on the unified VD, which improves compositionality in large vision-language models.

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.

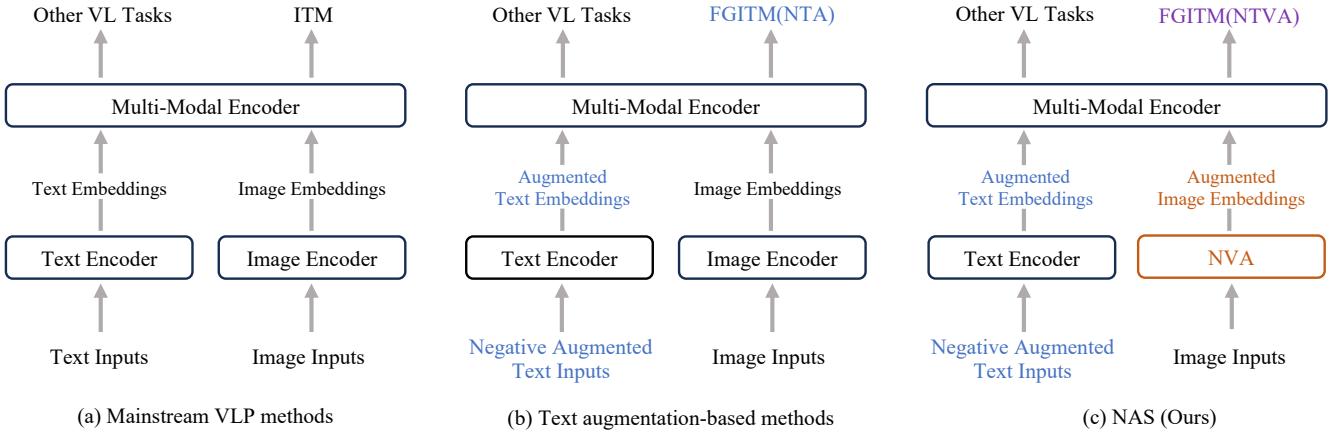


Figure 2: Comparison of our fine-grained NAS to other VL frameworks. Mainstream VLP methods utilize two "Dual Tower" encoders and use a multi-modal encoder for deep fusion of multi-modal features(e.g., ALBEF (Li et al. 2021) and METER (Dou et al. 2022))(a), NTA-based methods construct augmented negative 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. 2023b))(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).

## 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. 2022d). Recent studies employ NTA to construct fine-grained hard negative sentences with similar structures but different semantics (Yuksekgonul et al. 2022; Huang et al. 2024; Momeni et al. 2023). 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.

## 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  $v_{\text{cls}}, v_1, \dots, v_N$ , where  $v_{\text{cls}}$  is the [CLS] token’s embedding, and  $N$  denotes the number of image patches. The text  $T$  is similarly transformed into embeddings  $t_{\text{cls}}, t_1, \dots, t_M$ , with  $t_{\text{cls}}$  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:

$$h_i = \operatorname{argmin}_{d_j \in \mathcal{D}} \|v_i - d_j\|_2, \quad (1)$$

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

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

with  $\hat{d}_j$  representing the updated vector,  $\gamma$  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:

$$\hat{v}_i = \operatorname{sg}[d_{h_i} - v_i] + v_i, \quad (3)$$

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

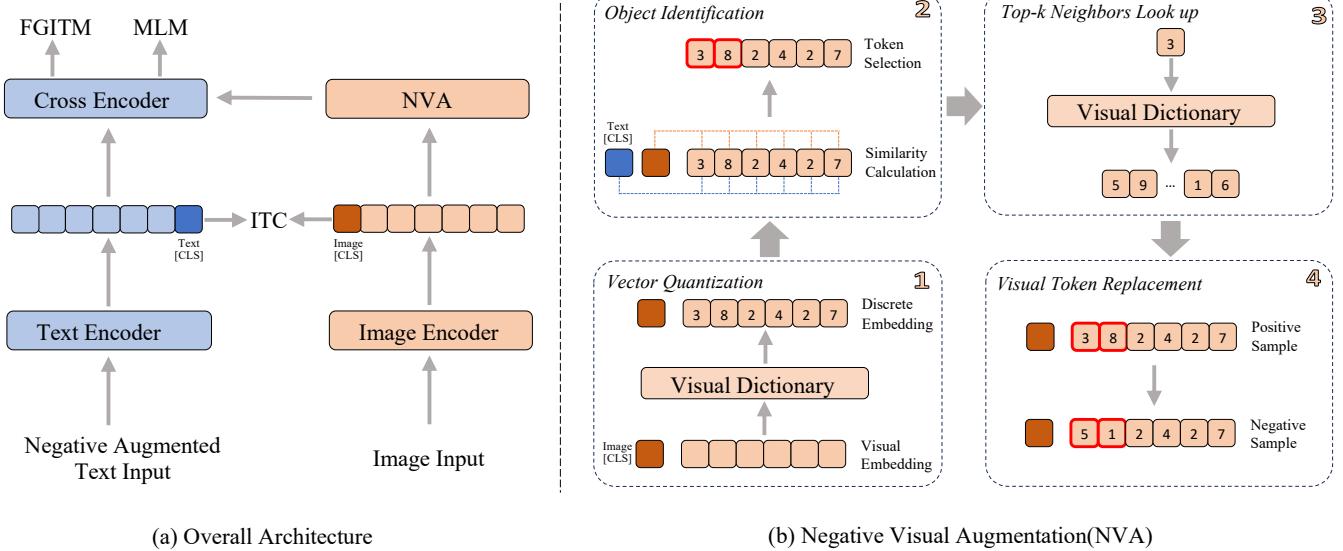


Figure 3: (a) The framework of the proposed end-to-end pretraining model NAS. (b) Illustration of our NVA. The continuous visual embedding encoded by the image encoder is firstly quantified into discrete embedding and then identifies the object in the image embedding based on the similarity between the 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.]

For the generation of negative image samples, we utilize the textual global feature  $T_{\text{cls}}$ , the visual global feature  $V_{\text{cls}}$ , and the visual local features  $v_i$ . Departing from conventional methods that rely solely on either  $T_{\text{cls}}$  or  $V_{\text{cls}}$  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  $T_{\text{cls}}$  and all local visual features  $v_i$ , and  $S_v$  between  $V_{\text{cls}}$  and  $v_i$ :

$$S_t = \text{cosine}\langle T_{\text{cls}}, v_i \rangle, \quad i = 1, \dots, N \quad (4)$$

$$S_v = \text{cosine}\langle V_{\text{cls}}, v_i \rangle, \quad i = 1, \dots, N \quad (5)$$

$$S = \lambda S_t + (1 - \lambda) S_v \quad (6)$$

The hyper parameter  $\lambda$  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 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  $p^{\text{itm}}$ . The FGITM loss is a cross-entropy loss:

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

where  $y^{\text{itm}}$  is a one-hot vector representing the ground-truth label.  $(I', T')$  includes  $(I, T)$ ,  $(I, T^{\text{neg}})$ ,  $(I, T^{\text{fg}})$ ,  $(I^{\text{neg}}, T)$  and  $(I^{\text{fg}}, T)$ , where  $I^{\text{neg}}/T^{\text{neg}}$  is the negative image/text selected in every training batch and  $I^{\text{fg}}/T^{\text{fg}}$  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  $v_{\text{cls}}$  and  $t_{\text{cls}}$  to normalized vectors in a reduced dimensional space. Two queues are maintained to cache the most recently ob-

Table 1: Statistics of the pretraining datasets.

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

tained  $M$  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 functions 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:

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

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

where  $\tau$  is a learnable temperature parameter.

Momentum distillation leverages the momentum model to distill the current training model, which is adapted to learn from pseudo-targets generated by the momentum model. The final targets are:

$$y^{\text{i2t}}(I) = (1 - \alpha)y_{\text{one-hot}}^{\text{i2t}}(I) + \alpha p^{\text{i2t}}(I_m) \quad (10)$$

$$y^{\text{t2i}}(T) = (1 - \alpha)y_{\text{one-hot}}^{\text{t2i}}(T) + \alpha p^{\text{t2i}}(T_m) \quad (11)$$

where  $y_{\text{one-hot}}^{\text{i2t}}(I)$  and  $y_{\text{one-hot}}^{\text{t2i}}(T)$  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  $y$ :

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

**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  $p^{\text{msk}}(I, \hat{T})$  denotes the predicted probability for a masked token. MLM minimizes a cross-entropy loss:

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

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

The full pretraining objective of NAS is:

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}} \quad (14)$$

## 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, currently there are **only** 2.9 million available images and 4 million image-text pairs. Detailed statistics are presented in Table 1. Our architecture leverages the initial six layers of BERT<sub>base</sub> 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  $\lambda$  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 (Loshchilov and Hutter 2019) 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  $256 \times 256$  resolution, and RandAugment (Cubuk et al. 2020) is adopted (we remove color changes from RandAugment because the text often contains color information). During the fine-tuning stage, the resolution of an image is up-scaled to  $384 \times 384$ , and the positional encoding of the image patches is interpolated (Dosovitskiy et al. 2021). The momentum parameter for updating the momentum model is 0.995, and the queue length of cached features for image-text contrastive learning is set as 65,536. We linearly ramp-up the distillation weight  $\alpha$  from 0 to 0.4 within the 1st epoch. All experiments are performed on 8 NVIDIA A800 GPUs and take around 2 days to train.

**Benchmarks** To test the effectiveness of our proposed NVA module, we have evaluated on three benchmarks. ARO (Yuksekgonul et al. 2022) is a large dataset 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** Our approach is benchmarked against several SOTA models. We primarily compare our model with multi-modal models and large language models. 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), XVLML (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 lan-

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	Relation	ARO		Winoground			Avg.
			Attribute	Avg.	Text	Image	Group	
Random Chance	-		50		25.0	25.0	16.7	22.2
UNITER	4M	-	-	-	32.3	13.3	10.0	18.5
ViT(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
ALBEF <sub>COCO</sub>	4M	60.5	88.5	74.5	27.5	15.8	11.0	18.1
<i>Large language models</i>								
BART	-	81.1	73.6	77.4	-	-	-	-
FLAN-T5	-	84.4	76.5	80.5	-	-	-	-
<i>Large Multi-modal models</i>								
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
<i>VD based models</i>								
FDT	3M	49.8	54.6	52.2	17.3	3.5	1.5	7.4
<i>Hard Negative based models</i>								
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	-	-	-	-
<b>NAS(NTA)coco</b>	2.9M	93.1	91.7	92.4	32.3	17.3	13.0	20.8
<b>NAS(NVA)coco</b>	2.9M	67.8	89.8	78.8	34.5	19.0	14.0	22.5
<b>NAS(NTVA)coco</b>	2.9M	<b>93.2</b>	<b>93.4</b>	<b>93.3</b>	<b>35.3</b>	<b>22.0</b>	<b>18.5</b>	<b>25.3</b>

Table 3: Results on the VALSE benchmark.

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
ALBEF <sub>COCO</sub>	2.9M	75.4	76.5	65.8	74.4	67.5	48.0	92.6	71.5
XVLM <sub>COCO</sub>	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
<b>NAS(NTA)coco</b>	2.9M	85.5	75.9	66.8	71.6	75.5	45.4	93.7	73.5
<b>NAS(NVA)coco</b>	2.9M	85.1	77.6	66.7	72.1	72.7	48.8	94.2	73.9
<b>NAS(NTVA)coco</b>	2.9M	87.3	77.6	70.1	69.9	75.8	46.7	93.2	<b>74.4</b>

Table 4: Ablation study of VD and NVA.

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

guage 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).

#### Detailed information about the benchmarks and baselines will be shown in Supplementary.



Figure 4: Cases on the VALSE benchmark. The first, second and third rows are the results on the existence, counting and actions(actant swap) test respectively. More examples are seen in Supplementary.

### 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 NAScoco. 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 and then fine-tuned on the COCO dataset.

Notably, even with less training data, NAS outperforms existing methods across these benchmarks (over the ALBEF baseline: an increase of 2.9% on VALSE, 7.2% on Winoground, and a remarkable 18.8% on ARO). Such gains validate our approach’s proficiency in leveraging hard negatives in both images and text. For NVA, our method significantly improves the fine-grained capability, demonstrating the effectiveness of our approach. In comparison, NTVA outperforms both NTA and NVA. It proves that our NVA module can be compatible with the NTA methods.

The NTA in our model performs particularly well on the ARO benchmark, attributing its success to the dataset curated for specific hard text negatives. Conversely, the NVA approach, through visual token replacement, ensures that our model attentively considers image details, thereby showing excellence in the VALSE and Winoground benchmark. By introducing the NTVA method, the fine-grained feature extraction and fine-grained modality alignment of the model are significantly enhanced. As depicted in Figure 4, our model adeptly discerns subtle image nuances (for instance, discerning “sheep standing”, “no dogs” and identifying “5 birds”), showcasing refined fine-grained feature extraction skills. In conclusion, NTVA consistently delivers the most

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

VD size	Winoground			VALSE	Avg.
	Text	Image	Group		
<b>1024</b>	31.3	<b>23.5</b>	14.8	72.5	35.3
<b>2048</b>	<b>35.3</b>	22.0	<b>18.5</b>	<b>74.4</b>	<b>37.6</b>
<b>4096</b>	29.8	23.3	15.3	71.7	35.0
<b>8192</b>	27.8	22.0	12.8	71.4	33.5

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

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

Table 7: Ablation study of the balance parameter  $\lambda$ .

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

compelling results across all benchmarks, cementing the superiority of integrating NVA with NTA methodologies. We also conducted a comparison on the broader retrieval task on COCO dataset. Despite utilizing less training data, our method demonstrates superior performance. Additionally, the use of exponential moving average mechanism to update VD results in the introduction of less than 1% of the total parameters. To validate the generalizability of our method, we applied it to the BLIP model (Li et al. 2022a), which also led to improved performance. These results can be found in Supplementary.

### Ablations

We conducted ablation studies to assess the effectiveness of the NVA module and its synergy with the NTA methods. The effectiveness of VD and NVA is assessed in Table 4. Additionally, we analyze the vector size  $m$  of the VD in Table 5. Further, we conducted ablations on the replacement ratio of visual tokens, and the results are shown in Table 6. Lastly, the effects of the balance parameter  $\lambda$  are detailed in Table 7.

The introduction of the VD resulted in overall performance enhancements, which supports the role of the VD in enhancing fine-grained feature extraction. The integration of NTA further refined the alignment of fine-grained modality, leading to significant improvements. Finally, the introduction of NVA improved the model’s performance, demonstrating 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 in our experimental setting. 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  $\lambda$ , 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.

## 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 our 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 tend to integrate image segmentation approaches to recognize accurate and complete semantic regions for NVA. Meanwhile, we will investigate the co-quantize approach to align multi-modal information earlier and deeper.

## 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.
- Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, 1931–1942.
- 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.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.
- 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.
- Doveh, S.; Arbelle, A.; Harary, S.; Schwartz, E.; Herzig, R.; Giryes, R.; Feris, R.; Panda, R.; Ullman, S.; and Karlinsky, L. 2023. Teaching structured vision & language concepts to vision & language models. In *Conference on Computer Vision and Pattern Recognition*, 2657–2668.
- Fang, Z.; Kuang, K.; Lin, Y.; Wu, F.; and Yao, Y.-F. 2020. Concept-based Explanation for Fine-grained Images and Its Application in Infectious Keratitis Classification. In *International conference on Multimedia*, 700–708. Seattle, WA, USA.
- 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.
- Hou, S.; Feng, Y.; and Wang, Z. 2017. VegFru: A Domain-Specific Dataset for Fine-Grained Visual Categorization. In *International Conference on Computer Vision*, 541–549. Venice, Italy.
- 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. StructureCLIP: 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.
- Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition*, 6700–6709.
- 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, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2022c. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. arXiv:2012.15409.
- Li, X. 2022. Multimodal Cognitive Computing. *SCIENTIA SINICA Informationis*, 53.
- Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.
- Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022d. 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.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.
- 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.
- Momeni, L.; Caron, M.; Nagrani, A.; Zisserman, A.; and Schmid, C. 2023. Verbs in Action: Improving Verb Understanding in Video-Language Models. In *International Conference on Computer Vision*, 15579–15591.
- 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.
- Pang, K.; Li, K.; Yang, Y.; Zhang, H.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2019. Generalising Fine-Grained Sketch-Based Image Retrieval. In *Conference on Computer Vision and Pattern Recognition*, 677–686. Long Beach, CA, USA.
- 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.

- Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision*, 2641–2649.
- Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; and Sacheti, A. 2020. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. arXiv:2001.07966.
- 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.
- Singh, H.; Zhang, P.; Wang, Q.; Wang, M.; Xiong, W.; Du, J.; and Chen, Y. 2023. Coarse-to-Fine Contrastive Learning in Image-Text-Graph Space for Improved Vision-Language Compositionality. In *Conference on Empirical Methods in Natural Language Processing*.
- 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.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Neural Information Processing Systems*, 30.
- 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.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.; Khan; Singhal, S.; Som, S.; and Wei, F. 2023a. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. In *Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada.
- Wang, W.; Yang, Z.; Xu, B.; Li, J.; and Sun, Y. 2023b. 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.
- Zhang, L.; Awal, R.; and Agrawal, A. 2023. Contrasting Intra-Modal and Ranking Cross-Modal Hard Negatives to Enhance Visio-Linguistic Compositional Understanding. arXiv:2306.08832.
- 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.

## Supplementary for Benchmarks

To test the effectiveness of our proposed Negative Visual Augmentation (NVA) module, we have evaluated on three benchmarks. The statistics of benchmarks we use are shown in Table 8.

**ARO** (Yuksekgonul et al. 2022) – or the Attribution, Relation, and Order benchmark, is a large dataset designed to evaluate the ability of Vision-Language Pretraining (VLP) models to understand four different types of skills. It consists of Visual Genome Attribution and Visual Genome Relation, which leverages the Visual Genome (Krishna et al. 2017) dataset along with the GQA (Hudson and Manning 2019) annotations to test the understanding of properties and relational understanding of objects in complex natural scenes. VG-Relation includes 48 distinct relations with 23,937 test cases, and VG-Attribution includes 117 unique attribute pairs with 28,748 test cases. It also leverages the COCO (Lin et al. 2014) and Flickr30k (Plummer et al. 2015) datasets to evaluate the model sensitivity to select the right caption after applying four different shuffling perturbations (e.g., exchanging nouns and adjectives, or by shuffling trigrams). These tests are performed on the 5000 and the 1000 images from the respective COCO and Flickr30k test splits, respectively.

**Winoground** (Thrush et al. 2022) – is a small dataset that evaluates the ability of VLP models for compositional reasoning, specifically understanding the meaning of the sentence after changing the order of its words. The dataset has 400 samples, each comprised of two images and two texts. The texts have the same words in a different order, each text corresponding to one image in the sample. The Winoground metrics include (a) image score - percent of samples where the model picks the correct text for each image; (b) text score - percent of samples where the model picks the correct image for each text; (c) group score - percent of samples where both text and image score conditions are satisfied jointly.

**VALSE** (Parcalabescu et al. 2022) - is a novel benchmark designed for testing VLP models for their Vision-Language (VL) grounding capabilities on specific linguistic phenomena. The dataset offers a suite of six tests covering a wide spectrum of basic linguistic phenomena affecting the linguistic and visual modalities: existence, plurality, counting, spatial relations, actions, and entity coreference. The existence test challenges models to discern when entities are present or absent in images. The plurality test checks a model’s ability to distinguish singular from plural entities in images. The counting test requires models to accurately count entities in an image. The relations test examines a model’s understanding of spatial prepositions. The actions test evaluates if a model can match actions described in text to those depicted in images and identify the participants and their roles. Finally, the coreference test checks if models can resolve pronominal references with visual grounding, relating pronouns in text to entities or regions in images.

## Supplementary for Baselines

Our methodology is compared to various previous state-of-the-art (SOTA) models, summarized as follows.

Table 8: Overview of the fine-grained VL benchmarks, where “IT” represents “Image-Text”.

Benchmark	Task	# IT pairs
ARO-Relation	Relation	24k
ARO-Attribution	Attribution	28.7k
VALSE	Linguistic Phenomena	6.8k
Winoground	Compositional Reasoning	1.6k

## Multi-modal Models

- **LXMERT** (Tan and Bansal 2019) learns the relationship between vision and language using Transformer (Vaswani et al. 2017) encoders and a unique cross-modality encoder. It also enhances generalization through diverse pretraining tasks.
- **ViLBERT** (Lu et al. 2019) extends the BERT (Devlin et al. 2019) architecture to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers.
- **UNITER** (Chen et al. 2020) introduces a conditional masking strategy on top of region-based image features to improve the fine-grained capability of the model.
- **ViLT** (Kim, Son, and Kim 2021) assigns visual feature extraction to Transformer modules and avoids separate visual embedders, resulting in performance and efficiency improvements on VL tasks.
- **CLIP** (Radford et al. 2021) exhibits robust zero-shot capabilities, trained via contrastive learning on large web-scale datasets, rivalling supervised methods in specific tasks.
- **ALBEF** (Li et al. 2021) first aligns the uni-modal image representation and text representation before fusing them with a multi-modal encoder, and implements momentum distillation to mitigate data noise.
- **XVLM** (Zeng, Zhang, and Li 2022) performs “multi-grained vision language pretraining” to align texts with visual concepts in the images based on multi-granularity alignments.
- **FLAVA** (Singh et al. 2022) excels across vision, language, and combined tasks with a tailored pretraining objective set.
- **Neg-CLIP** (Yuksekgonul et al. 2022) presents composition-aware hard negative mining to generate textual hard negatives during model training to improve the compositional and order understanding of VLP models.
- **syn-CLIP** (Cascante-Bonilla et al. 2023) contributes a million-scale synthetic dataset and data generation pipeline using a 3D physics-based simulation platform to improve understanding and compositional reasoning of VLP models.
- **SPEC** (Peng et al. 2024) introduces an efficient pipeline to synthesize candidate images that exclusively differ in

Table 9: Zero-shot image-text retrieval results on Flickr30K and COCO.

Model	# Images	Flickr30K (1K test set)						MSCOCO (5K test set)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT	6M	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
UNITER <sub>Large</sub>	4M	83.6	95.7	97.7	68.7	89.2	93.9	64.1	87.7	93.3	48.8	76.7	85.8
ViLT	4M	73.2	93.6	96.5	55.0	82.5	89.8	56.5	82.6	89.6	40.4	70.0	81.1
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
UNIMO <sub>Base</sub>	4M	77.4	95.1	97.8	62.4	86.2	91.7	-	-	-	-	-	-
ALBEF	2.9M	88.1	98.3	99.3	74.4	92.0	95.3	68.0	89.5	94.3	49.2	76.0	84.6
UNIMO-2	19.8M	88.5	96.8	98.9	72.7	91.2	94.6	-	-	-	-	-	-
<b>NAS(NVA)</b>	2.9M	<b>90.8</b>	<b>98.9</b>	<b>99.7</b>	<b>76.8</b>	<b>93.2</b>	<b>96.4</b>	<b>70.9</b>	<b>90.6</b>	<b>95.0</b>	<b>52.7</b>	<b>79.0</b>	<b>86.8</b>

a specific visual attribute and created the SPEC benchmark to diagnose the comprehension proficiency of VLP models.

- **FDT** (Chen et al. 2023) embeds both images and texts using a set of shared learnable discrete tokens, reducing the granularity gap between the two modalities. The matched visual and semantic concepts are enforced to be represented by the same set of discrete tokens by a sparse activation constraint.
- **BLIP2** (Li et al. 2023) employs a lightweight Querying Transformer to bridge the modality gap. Pretrained in two stages for VL representation and vision-to-language generative learning. It demonstrates emerging capabilities in zero-shot instructed image-to-text generation.
- **MiniGPT-4** (Zhu et al. 2023) aligns a frozen visual encoder with a frozen LLM, Vicuna, using just one projection layer.
- **LLaVA** (Liu et al. 2024) uses language-only GPT-4 to generate multi-modal VL instruction-following data and connects a vision encoder and LLM for general purpose VL understanding.

### Large Language Models

- **BART** (Yuan, Neubig, and Liu 2021) is a generation-based baseline, using a pure LLM to compute text-only GPTScore.
- **FLAN-T5** (Chung et al. 2024) excels in generalization performance and can perform well on more than 1800 NLP tasks.

### Supplementary for Experiments

This section provides an analysis of the performance of our model on the Image-Text Retrieval (ITR) task, which includes image-to-text retrieval (TR) and text-to-image retrieval (IR) subtasks. To assess the effectiveness of our pretrained model, we conducted experiments on the Flickr30K (Plummer et al. 2015) and COCO (Lin et al. 2014) datasets in both fine-tuning and zero-shot settings. In the fine-tuning paradigm, the model, which was pretrained on a training set, underwent additional fine-tuning before evaluation on a separate validation/test set. Conversely, the

zero-shot setting entailed evaluating the pretrained model directly on the test set. We benchmarked zero-shot retrieval on Flickr30K by employing the evaluation protocol established by ALBEF (Li et al. 2021), which was fine-tuned on COCO. For evaluation on ITR, we adopted a fine-tuning loss function comprising Image-Text Contrastive learning (ITC) and Image-Text Matching (ITM), with a re-ranking mechanism applied during inference. Initially, images and texts were encoded independently, and their respective similarity matrices were computed to yield the top-k candidates. Subsequently, candidate representations were input into a multi-modal encoder for re-ranking.

We compare our model with ImageBERT (Qi et al. 2020), UNITER (Chen et al. 2020), ViLT (Kim, Son, and Kim 2021), SOHO (Huang et al. 2021), CLIP (Radford et al. 2021), ALBEF (Li et al. 2021), UNIMO (Li et al. 2022c), UNIMO-2 (Li et al. 2022b) and METER (Dou et al. 2022). As depicted in Table 9 and Table 10, our method demonstrates superior performance compared with existing works on both zero-shot and fine-tuned retrieval benchmarks. Our NVA method excels against current VLP models, notably on the COCO dataset. Specifically, in the zero-shot scenario, our NAS (NVA) model surpasses ALBEF by 2.2% in TR and 2.6% in IR regarding R@1 on COCO. For the fine-tuned retrieval, NAS (NVA) outperforms METER-Swin (Dou et al. 2022) by margins of 1.0% for TR and 2.2% for IR in terms of R@1 on COCO with less training data. It’s important to highlight that while both Negative Textual Augmentation (NTA) and Negative Textual and Visual Augmentation (NTVA) methods showed underwhelming performance in retrieval tasks, the NTA method specifically caused the model to lose its capacity for capturing global information due to its focus on textual details. Our end-to-end NVA methodology does not suffer from this limitation, underscoring its advantages.

**Extending method on other backbones** Our proposed method is model-agnostic and can be applied to other VLP models. To verify it, we have extended our experiments to include BLIP (Li et al. 2022a). For a fair comparison, we compare our method with models fine-tuned on the COCO dataset. The experimental results are presented in Table 11.

Firstly, we observe that the BLIP-based model exhibits similar performance trends as the ALBEF-based model.

Table 10: Fine-tuned image-text retrieval results on COCO.

Model	# Images	MSCOCO (5K test set)					
		Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT	6M	66.4	89.8	94.4	50.5	78.7	87.1
UNITER <sub>Large</sub>	4M	65.7	88.6	93.8	52.9	79.9	88.0
ViLT	4M	61.5	86.3	92.7	42.7	72.9	83.1
SOHO	200K	66.4	88.2	93.8	50.6	78.0	86.7
ALBEF	2.9M	72.6	91.5	95.7	55.2	80.8	<b>89.4</b>
METER-Swin	4M	73.0	92.0	96.3	54.9	81.4	89.3
<b>NAS(NTA)</b>	2.9M	62.7	88.0	94.0	54.5	80.4	88.1
<b>NAS(NVA)</b>	2.9M	<b>74.0</b>	<b>92.7</b>	<b>96.4</b>	<b>57.1</b>	<b>82.0</b>	89.2
<b>NAS(NTVA)</b>	2.9M	63.1	87.8	94.2	54.5	80.2	88.0

Table 11: Experiments on BLIP-based model exhibit similar performance trends.

BackBone	Methods	ARO		Winoground			VALSE Avg.
		Rel.	Att.	Text	Image	Group	
ALBEF	Fine-tune	60.5	88.5	27.5	15.8	11.0	72.1
	<b>NTVA</b>	<b>93.2</b>	<b>93.4</b>	<b>35.3</b>	<b>22.0</b>	<b>18.5</b>	<b>74.4</b>
BLIP	Fine-tune	59.0	88.0	48.0	23.8	19.5	73.3
	<b>NTVA</b>	<b>65.3</b>	<b>90.7</b>	<b>49.3</b>	<b>27.3</b>	<b>22.8</b>	<b>74.7</b>

Specifically, our proposed method results in a significant enhancement across different benchmarks. This demonstrates that our proposed method can effectively enhance fine-grained capability and is transferable to a wide range of multi-modal models. Moreover, by applying our proposed method to the superior base model BLIP, we achieve superior performance.

## Supplementary for Qualitative Examples

This section presents a comprehensive qualitative evaluation of our model’s performance on the VALSE and Winoground benchmarks.

Figure 5 provides a selection of qualitative comparisons between our model and the baseline ALBEF model (Li et al. 2021) using the VALSE benchmark (Parcalabescu et al. 2022). These examples showcase our model’s ability to understand abstract concepts, such as stylized facial features in artwork, and to identify subtle details, such as barely noticeable masts in images. Additionally, our model excels in tasks related to quantification and action understanding, demonstrating its sensitivity to subtle visual nuances and affirming the improvements in fine-grained feature extraction. Figure 6 offers additional insights where our model showcases enhanced performance over the baseline on the Winoground benchmark (Thrush et al. 2022). The consistent success of our model in accurately pairing subjects with their attributes across diverse contexts indicates its advancements in fine-grained modality alignment and compositional reasoning.

Overall, these examples illustrate the enhanced fine-grained capability of our model, confirming the effectiveness of our proposed method.

Existence:



Ours:  
There are faces in this picture

ALBEF:  
There are no faces in this picture



Ours:  
There is a mast on the closest boat

ALBEF:  
There is no mast on the closest boat



Ours:  
There are no people visible

ALBEF:  
There are people visible



Ours:  
There is a skateboarder in the picture

ALBEF:  
There is no skateboarder in the picture

Counting:



Ours:  
There are exactly 2 lamp post

ALBEF:  
There is exactly 1 lamp post



Ours:  
There are exactly 7 raspberries pictured

ALBEF:  
There are exactly 5 raspberries pictured



Ours:  
There are exactly 9 people in the picture

ALBEF:  
There are exactly 5 people in the picture



Ours:  
There are exactly 5 lights on the traffic light

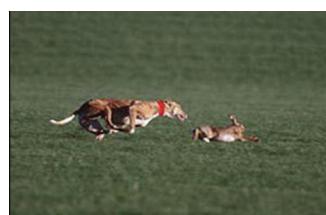
ALBEF:  
There are exactly 6 lights on the traffic light

Action(actant swap):



Ours:  
A man camps in a tent

ALBEF:  
A tent camps for a man



Ours:  
A dog hunts a rabbit

ALBEF:  
A rabbit hunts a dog



Ours:  
A man plows over a cow

ALBEF:  
A cow plows over a man



Ours:  
A woman decorates the wall

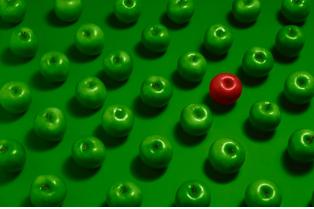
ALBEF:  
A wall decorates the woman

Figure 5: Examples on the VALSE benchmark. Sequentially from top to bottom, the panels display results for the existence test, counting test, and actions(actant swap) test respectively.

Winoground(*image*):



one green apple surrounded by a bunch of red apples



one red apple surrounded by a bunch of green apples

Ours ✓	ALBEF ✗
--------	---------



a car going into a hollow tube



a hollow tube going into a car

Ours ✓	ALBEF ✗
--------	---------

Winoground(*text*):



the dog is swimming and the person is standing



the dog is standing and the person is swimming

Ours ✓	ALBEF ✗
--------	---------



a person carrying more than one flotation device



more than one flotation device carrying a person

Ours ✓	ALBEF ✗
--------	---------



the red car is behind the blue car



the blue car is behind the red car

Ours ✓	ALBEF ✗
--------	---------



the red car is behind the blue car



the blue car is behind the red car

Ours ✓	ALBEF ✗
--------	---------



white land and black water



black land and white water

Ours ✓	ALBEF ✗
--------	---------



a large living thing in front of a large non living thing



a large non living thing in front of a large living thing

Ours ✓	ALBEF ✗
--------	---------

Figure 6: Examples on the Winoground benchmark. The first row illustrates the results of the image test, and the second row pertains to the text test.