Negation-Aware Test-Time Adaptation for Vision-Language Models

Haochen Han, Alex Jinpeng Wang, Fangming Liu, Jun Zhu, *Fellow, IEEE*

***Abstract*—In this paper, we study a practical but less-touched problem in Vision-Language Models (VLMs), *i.e.*, negation understanding. Specifically, many real-world applications require models to explicitly identify what is false or non-existent, *e.g.*, radiologists may search for images that exclude specific conditions. Despite the impressive transferability of VLMs through large-scale training, they suffer from a critical limitation that fails to handle negation. To address this challenge, existing methods attribute its root cause to the scarcity of negation training data and propose to fine-tune VLMs on massive data containing explicit negation. Undoubtedly, such data-centric solutions demand substantial data and computational resources, limiting their sustainable widespread adoption. To tackle negation in a low-carbon manner, we empirically observe that the key obstacle lies in the dual-concept shifts between the affirmation and negation distributions. Therefore, we propose a Negation-Aware Test-Time Adaptation (NEAT) method to efficiently adjust distribution-related parameters during inference. In brief, NEAT can reduce distribution shift in consistent semantics while eliminating false distributional consistency in unrelated semantics. Extensive experiments on the various negation understanding tasks verify the effectiveness of the proposed method. Remarkably, with less than 0.01% of trainable parameters, NEAT achieves comparable or superior performance to state-of-the-art post-training approaches. Our code is available at https://github.com/hhc1997/NEAT.**

***Index Terms*—Negation Understanding, multi-modal Learning, Vision-language Models, Test-time Adaptation.**

I. Introduction

Negation, a fundamental logical concept, benefits human perception and decision-making by encoding information about non-existent entities. Recent studies in cognitive science [8], [16], [40] show that humans first understand negation before developing broader world knowledge: even 18-month-olds can use negative sentences to constrain novel object meanings [8].

However, such cognitive processes in humans do not emerge in advanced multi-modal artificial intelligence, especially Vision-Language Models (VLMs). Conversely, despite learning rich open-world concepts from millions of image-text pairs, VLMs fail to comprehend negation [2], [34], severely hindering their applications in many real-world scenarios. For example, drones might query "a road without ice" during extreme weather rescue missions, or a radiologist may search for images showing "pulmonary nodules without malignant features". Therefore, understanding what is false or non-existent is crucial for VLMs to perform precisely.

Co-corresponding authors: Alex Jinpeng Wang and Fangming Liu.

Haochen Han, Alex Jinpeng Wang, and Fangming Liu are with the Department of AI Computing, Pengcheng Laboratory, Shenzhen, China. e-mail: hhc2077@outlook.com; jinpengwang@csu.edu.cn; fangminghk@gmail.com. Alex Jinpeng Wang is also with the School of Computer Science and Technology, Central South University, Changsha, China.

Jun Zhu is with the Department of Computer Science and Technology, Institute for AI, BNRist Center, THBI Lab, Tsinghua-Bosch Joint Center for ML, Tsinghua University, Beijing, China. e-mail: dcszj@tsinghua.edu.cn.

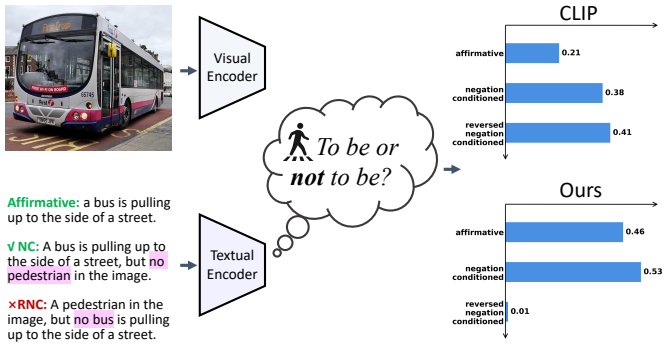


Fig. 1. A toy example to illustrate the **dual-concept shifts problem** in Vision-Language Models when understanding negation. Negation-conditioned (NC) texts negate absent elements and share consistent semantics with affirmative ones. Reversed negation-conditioned (RNC) texts wrongly affirm absent elements and negate present ones, which are semantically opposite to NC texts. Although CLIP learns diverse open-world knowledge, it exhibits a similarity gap between affirmative and NC texts, and high similarity between NC and RNC texts. We show how distributional adaptation can tackle this issue.

To achieve reliability against negation, existing methods [2], [34], [39] propose that the root cause lies in the scarcity of negation terms in VLMs' pre-training data, and resort to generating negation-inclusive data for post-training. Specifically, NegBench [2] introduces the large-scale synthetic dataset NegFull, which generates more than 70 million image-text pairs containing explicit negation based on CC12M [5]. The concurrent work NegationCLIP [34] proposes two data generation pipelines using Large Language Model (LLM) and Multi-modal Large Language Model (MLLM) to augment captions with negation. Despite the success, these data-centric approaches require substantial data and computational resources, limiting their sustainable widespread adoption. Thus, it is essential to endow VLMs with negation-aware capabilities in a cost-effective manner.

In this paper, we think outside the box of complex post-training and argue that—*the key obstacle to negation understanding is the dual-concept shifts between the affirmation and negation distributions*. Taking the prevailing CLIP [36] for example, as illustrated in Fig. 1, CLIP exhibits a significant similarity gap between the affirmative caption and its negation-conditioned caption, *e.g.*, specifying 'no pedestrian' for a more precise description, despite them sharing consistent semantics. On the other hand, CLIP shows incorrect similarity between the negation-conditioned caption and its semantically reversed counterpart, *e.g.*, negating existent entities (bus) and wrongly affirming absent ones (pedestrian). As a result, VLMs only interpret negation statements as bags of words and may even produce entirely incorrect judgments.

Based on the above discussions and observations, instead of the full-parameter fitting on negation data, we propose only

adjusting distribution-related parameters (*e.g.*, normalization layers) to tackle the dual-concept shifts during inference. To be specific, we propose a novel **Ne**gation-**A**ware **T**est-time Adaptation (dubbed NEAT) framework, which can rapidly learn from unlabeled multi-modal negation data and then make reliable predictions. Our NEAT encapsulates three key components. First, the negation-separated refinement module is adopted to obtain refined entropy between negated captions and visual content, where descriptions are decomposed into two affirmative parts for VLMs to handle adeptly. Second, to address the false distributional consistency in unrelated semantics, we propose to view the reversed negation-conditioned caption as the hardest negative information that shares opposite semantics to its corresponding visual sample. Theoretically, we show that this reversed contrastive objective only requires comparing against limited visual samples and is equivalent to a simple metric loss. Third, to achieve efficient adaptation, we suggest directly reducing the textual-distribution gap arising from dual-concept shifts. Empirically, NEAT establishes state-of-the-art performance on multiple negation benchmarks that span images, videos, and medical scenarios, outperforming the expensive post-trained baselines by a significant margin and observing strong generalization ability in adapted VLMs.

Our main contributions are summarized as follows:

- To the best of our knowledge, this work could be the first study to enhance negation understanding of VLMs via test-time adaptation. We reveal that the key to negation understanding is the dual-concept shifts problem between affirmation and negation distributions, which could be addressed by adapting lightweight normalization layers of models.
- We propose a novel test-time adaptation method named NEAT, which comprises three key components: the negation-separated entropy refinement to reduce distribution shift in consistent semantics, the reversed contrastive learning to eliminate the false distributional consistency in unrelated semantics, and the debiased textual learning to further achieve efficient adaptation.
- Extensive experiments verify the effectiveness of the proposed method. Remarkably, our NEAT achieves comparable or even superior performance to the best CLIP-based post-training baseline with significantly reduced resources—using less than 0.36‰ of the data (unlabeled) and 0.014‰ of the trainable parameters. Moreover, the adapted layers show strong transferability that can be directly generalized to unseen datasets and tasks.

II. Related Works

A. Negation Understanding in Vision-Language Models

Vision-language foundation models learned diverse knowledge from large-scale multi-modal data, which have attracted significant attention due to their powerful transfer capabilities [3], [51]. By supervising visual representations with natural language descriptions, pioneering works CLIP [36] and ALIGN [20] show great success in various downstream tasks. However, the web-crawled training data are mainly affirmative, which makes it difficult to comprehend the negation of the foundation VLMs. To comprehensively evaluate model performance in negation scenarios, recent works CREPE [30] and CC-Neg [39] proposed vision-language benchmarks that target compositional understanding with negation, but the reliance on linguistic templates limits the diversity of real negation queries. In comparison, NegBench [2] utilizes LLMs to generate more natural negated captions that span images, videos, and medical datasets. Despite being trained on billion-scale data, the unsatisfactory performance on these benchmarks reveals that VLMs struggle to understand negation. Recent work [53] shows that such problems persist even in powerful MLLMs like GPT-4o and Claude 3.5, which suffer from hallucinations when processing user-provided negation arguments.

To address this limitation, a series of data-centric approaches attempted to introduce negated captions as distractors in an additional post-training phase. For instance, NegCLIP [49] creates targeted negative captions to enhance compositional reasoning. However, these negative descriptions only focus on swapped relations and ignore negation attributes. To this end, CoNCLIP [39] modifies the contrastive loss to incorporate template-based examples that include negation words, *e.g.*, 'no', 'not', and 'without'. The recent advance NegBench [2] further fine-tunes VLMs on 70 million synthesized image-text pairs containing diverse negated statements. Concurrent work NegationCLIP [34] proposes two data generation pipelines that leverage LLMs and MLLMs to generate 229k negation-augmented image-text pairs for post-training.

Although these methods have achieved great success, almost all of them assume that the bottleneck of negation understanding lies in the scarcity of training examples with explicit negation. In contrast, this study pioneers an efficient solution from the perspective of the distribution shift. Considering the computational burden of large VLMs, our method can directly adapt VLMs to handle various negation understanding tasks with only a few parameter updates at test time.

B. Test-Time Adaptation

The distribution shift between training and test data poses a key challenge when transferring the zero-shot capabilities of foundation VLMs [14], [29], [41]. In practice, such shifts are inevitable due to environmental variations [22] or the encounter of unseen concepts [45]. To address the problem, considerable efforts have been devoted to developing training-time solutions: domain adaptation methods [15] that narrow gaps between source and target samples during training. Domain generalization methods [1] that directly learn domain-invariant representations through robust training strategies. However, these methods may require access to source domain data and cannot achieve adaptation in an online manner, limiting their practicality in real-time applications such as autonomous driving.

To this end, test-time adaptation methods have emerged to adapt pre-trained VLMs to test samples on the fly. The most prevalent application is to tackle out-of-distribution images [13], [21], [25], [38], [48], [50], [52] caused by corruptions or environments. For example, TDA [21] designs a training-free dynamic adapter to gradually refine pseudo labels of

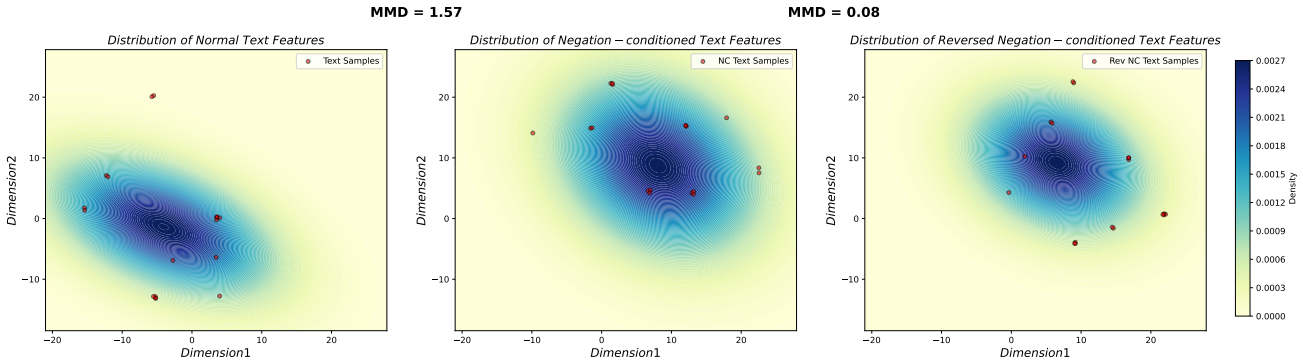


Fig. 2. **Dual-concept shifts** problem observed in pre-trained VLMs when understanding negation. Distributions of normal text, negation-conditioned text, and reversed negation-conditioned text of the test set of CIFAR10 are shown on the left, middle, and right, respectively. All embeddings are produced by the OpenAI CLIP and visualized with UMAP [31] for dimension reduction. The Maximum Mean Discrepancy (MMD) metrics are also provided above for quantitative analysis. The distributions of normal and negation-conditioned text show a strong shift (MMD=1.57) despite sharing the same semantics. Meanwhile, the distributions of negation-conditioned and its reversed text show high consistency (MMD=0.08) despite being semantically opposite.

test samples. TPT [38] learns adaptive prompts by enforcing entropy consistency across augmented views with confidence-based selection. DiffTPT [13] leverages pre-trained diffusion models to generate diverse and informative data augmentations to improve TPT. C-TPT [48] reduces the prediction uncertainty of TPT by establishing calibration error with text feature dispersion. RLCF [52] introduces a CLIP model to provide feedback that avoids the pitfall of the entropy minimization. RA-TTA [25] uses external knowledge obtained from a web-scale image database to prevent over-reliance on model predictions. DPE [50] evolves prototypes from both textual and visual modalities to capture more accurate multi-modal representations. Beyond these zero-shot image classification tasks, recent works have also shown the promise of TTA in more challenging vision-language retrieval tasks. For instance, to alleviate the potential bias from gender or race, PBM [23] generates fair retrieval subsets from the self-prediction of VLMs. TCR [26] manipulates both the modality uniformity and modality gap to achieve robust retrieval under query shift.

Different from these prior arts that mainly focus on covariate shifts like corruptions (*e.g.*, noise, blur, and weather) or style variations (*e.g.*, cartoon and sketch), this paper presents the first attempt to tackle concept shift problems arising from negation understanding. As negation can enable more precise queries in many real-world scenarios, it is reasonable to believe that this study could provide some novel and practical insights to the multi-modal community.

III. Preliminaries

A. VLMs Do Not Understand Negation

Let f_θ represent a VLM pre-trained on image-text pairs $\mathcal{D}_{train} = \{\mathcal{V}, \mathcal{T}\}$, where $v_i \in \mathcal{V}$ is the raw image and $t_i \in \mathcal{T}$ is the corresponding textual caption. One fundamental capability of VLMs is to make multiple data modalities comparable in the same embedding space, where visual embedding $\boldsymbol{v}_i \in \mathbb{R}^D$ and textual embedding $\boldsymbol{t}_i \in \mathbb{R}^D$ are derived by passing v_i and t_i through its visual encoder f_{θ_V} and textual encoder f_{θ_T}, respectively.

Despite the promising performance through large-scale training, the textual supervision $\mathcal{T}$ in data is mainly expressed affirmatively, which poses a natural limitation: vision-language foundation models fail to understand negation.

B. Test-Time Adaptation for Negation Understanding

For the pre-trained VLM, the key obstacle to negation understanding is the incongruence between the train and test distributions. Specifically, the text with negation conditions can be viewed as a concept shift [45] compared to the affirmative expressions, *i.e.*, $P(\mathcal{T}_{test}) \neq P(\mathcal{T})$, where $P(\cdot)$ denotes the distribution of the given textual data. As a result, VLMs suffer performance degradation [2] when facing real-world applications requiring comprehension of negation. To address such a generalization problem, TTA has emerged to boost the foundation model under data distribution shifts, which online updates only a minimal set of θ, *e.g.*, normalization layers [43] or prompt vectors [38], at test time before making a prediction.

IV. Methodology

A. Analysis of Concept Shifts in Negation Understanding

In this section, we present the analysis of where the shift in negation understanding comes from. Let $\hat{t}_i$ denote a text constrained by some negation conditions that maintains semantic consistency with its affirmative counterpart t_i. For instance, t_i typically describes the object in v_i or its associated attributes, *e.g.*, "a photo of a dog". By introducing some negative concepts, $\hat{t}_i$ can exclude specific elements for a more fine-grained description, *e.g.*, "a photo of a dog not on grass" or "a photo of a dog not running". To explore how such negation affects the understanding of VLMs, we experiment with the representative CLIP model on CIFAR10 and have several nontrivial findings.

Distribution shift within semantic consistency. For each image, we first form the normal caption "a photo of the [CLASS]" by inserting its class name into a fixed prompt template. Then, we synthesize a corresponding negation-conditioned caption "a photo of the [CLASS] but not of the [CLASS′]" by randomly incorporating a different class name. As each image in CIFAR10 is single-categorized, both descriptions are correct and ideally share consistent semantics.

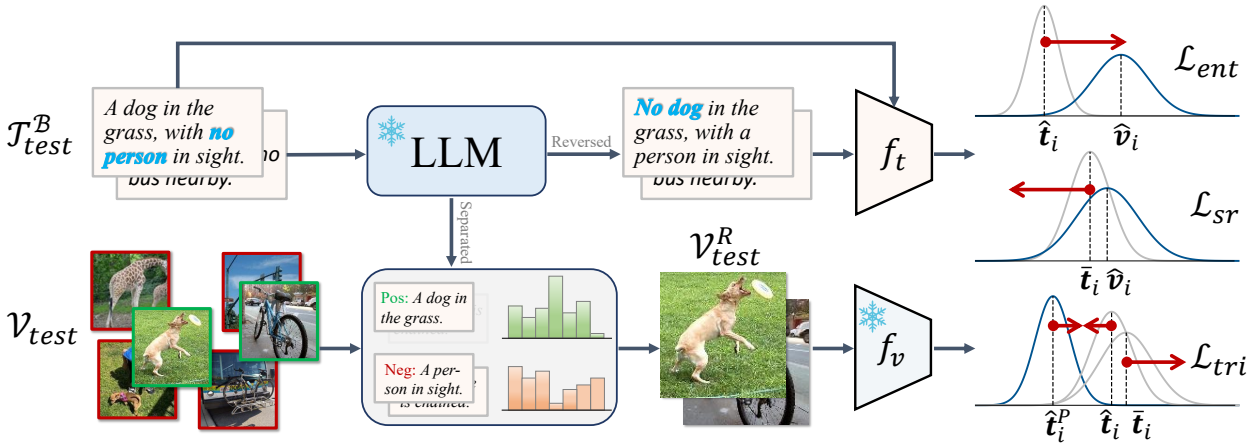


Fig. 3. Overview of the proposed NEAT. Given the test data with negation contexts, a LLM is first employed to separate the negation forms and generate semantically reversed counterparts. Then the refinement module is used to coarsely select candidate visual samples that are similar to positive entities while distant from negative entities. After that, three training objectives are adopted to reduce distribution shift in consistent visual-textual semantics ($\mathcal{L}_{ent}$), eliminate the false distributional consistency in unrelated visual-textual semantics ($\mathcal{L}_{sr}$), and directly debias the dual-concept shifts in the textual modality ($\mathcal{L}_{tri}$).

However, as shown in Fig. 2, the distributions of normal text (Fig. 2 left) and negation-conditioned text (Fig. 2 middle) show a significant shift in the embedding space.

Distribution consistency within semantic shift. For each negation-conditioned caption, we synthesize a semantically-reversed counterpart "a photo of the [CLASS′] but not of the [CLASS]" by simply swapping two class names. Obviously, such two descriptions convey entirely opposite semantics. However, we observed that the distributions of negation-conditioned text and their semantically reversed text (Fig. 2 right) show a substantial similarity in the embedding space.

Despite containing rich and diverse knowledge, these observations indicate that pre-trained VLMs only interpret text as bags of words [49]. Consequently, VLMs behave with the above dual-concept shifts when understanding negation.

B. Negation-Aware Test-Time Adaptation

Next we describe our **Ne**gation-**A**ware **T**est-time Adaptation (NEAT) framework in detail. As shown in Fig. 3, NEAT comprises three key components to bridge the above dual-concept shifts, *i.e.*, the refined entropy minimizing loss $\mathcal{L}_{ent}$, the semantics reversion loss $\mathcal{L}_{sr}$, and the textual debiasing loss $\mathcal{L}_{tri}$. For the given online batch with size $\mathcal{B}$, we update the pre-trained VLM in time by the following training objective:

$$\min_{\tilde{\theta}_T} \mathcal{L}(\mathcal{B}) = \mathcal{L}_{ent} + \mathcal{L}_{sr} + \mathcal{L}_{tri}, \tag{1}$$

where $\tilde{\theta}_T \subseteq \theta_T$ denotes the parameters of normalization layers to only adjust the data distribution. In the following, we will elaborate on each loss individually.

1) Entropy Refinement via Negation Separation: In online TTA, we have an unlabeled vision-language dataset $\mathcal{D}_{test} = \{\mathcal{V}_{test} = \{v_i\}_{i=1}^{N_V}, \mathcal{T}_{test} = \{\hat{t}_i\}_{i=1}^{N_T}\}$, where v_i depicts the natural visual concepts like in $\mathcal{D}_{train}$ but $\hat{t}_i$ may contain additional negative conditions that differ from $\mathcal{D}_{train}$. Our goal is to adapt the pre-trained model immediately to establish semantic correspondence between $\mathcal{V}_{test}$ and $\mathcal{T}_{test}$. To achieve this, one simple yet powerful solution is to fine-tune f_{θ_T} by minimizing the entropy of its predictions [43]. Formally, denote $\boldsymbol{T}_{test}^{\mathcal{B}} = [\hat{\boldsymbol{t}}_1, \ldots, \hat{\boldsymbol{t}}_{\mathcal{B}}]^\top \in \mathbb{R}^{\mathcal{B}\times D}$ the batched textual features generated by f_{θ_T}, and $\boldsymbol{V}_{test} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{N_T}]^\top \in \mathbb{R}^{N_T\times D}$ the fixed visual features generated by f_{θ_V}, the adaptation object of the online batch is

$$\min_{\tilde{\theta}_T} \mathcal{L}_{ent} = -\frac{1}{N_T}\sum_{i=1}^{N_T} \boldsymbol{P}_i \log \boldsymbol{P}_i, \tag{2}$$

where $\boldsymbol{P}_i = \text{Softmax}\left(\boldsymbol{v}_i \left(\boldsymbol{T}_{test}^{\mathcal{B}}\right)^\top / \tau_1\right) \in \mathbb{R}^{\mathcal{B}}$ is the model's output probability smoothed by a temperature τ_1. However, the large computational size would make entropy a sub-optimal confidence metric, leading to either model underfitting or overfitting [26].

To alleviate this, we follow the candidate selection strategy suggested in [26] to refine the prediction of the model. Specifically, let $\mathcal{N}(\cdot)$ denote a selection manner, and the corresponding visual candidate in the batch is obtained by

$$\boldsymbol{V}_{test}^{R} = [\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_{\mathcal{B}}]^\top \in \mathbb{R}^{\mathcal{B}\times D}, \hat{\boldsymbol{v}}_i = \mathcal{N}(\hat{\boldsymbol{t}}_i), \forall i \in [1, \mathcal{B}]. \tag{3}$$

The most straightforward selection method is the nearest neighborhood based on similarity [26]. However, as mentioned in Section IV-A, the strong shift within the training and test patterns makes the spatial relation of embeddings unreliable.

As a remedy, we propose to separate the negation part from $\hat{t}_i$ to avoid model understanding of negative semantics. To this end, we utilize the in-context learning ability of LLMs to decompose $\hat{t}_i$ into positive component $\hat{t}_i^P$ and negative one $\hat{t}_i^N$. For example, "a photo of a dog not on grass" would be split into "a photo of a dog" and "a photo of grass", where the negative element is also recaptioned affirmatively. Such simple parse tasks could be handled by lightweight LLMs, and we empirically find that Llama-3-8B [10] performs well for our goal. As the pre-trained VLM can readily match visual samples to affirmative captions, we can use the similarity

between v_j and $\hat{t}_i^N$ to penalize the correspondence from v_j to $\hat{t}_i^P$:

$$S(\hat{\boldsymbol{t}}_i, \boldsymbol{v}_j) = (\hat{\boldsymbol{t}}_i^P \boldsymbol{v}_j^\top)(1 - \alpha[\hat{\boldsymbol{t}}_i^N \boldsymbol{v}_j^\top]_+), \tag{4}$$

where α is a trade-off parameter to control the penalty and $[x]_+ = max(x, 0)$ is the hinge function. Intuitively, higher $S(\hat{\boldsymbol{t}}_i, \boldsymbol{v}_j)$ requires v_j depicts $\hat{t}_i^P$ well and remains unrelated to concepts in $\hat{t}_i^N$. Then we select the most similar sample by Eq.(4) to obtain the corresponding visual candidates $\boldsymbol{V}_{test}^R$. Consequently, the entropy minimizing objective for the online batch is

$$\begin{aligned} &\min_{\tilde{\theta}_T} \mathcal{L}_{ent} = -\frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \boldsymbol{P}_i^R \log \boldsymbol{P}_i^R, \\ &\text{where } \boldsymbol{P}_i^R = \text{Softmax}\left(\hat{\boldsymbol{v}}_1 \left(\boldsymbol{T}_{test}^{\mathcal{B}}\right)^\top / \tau_1\right) \in \mathbb{R}^{\mathcal{B}}. \end{aligned} \tag{5}$$

By excluding irrelevant samples, the refined entropy could prevent model underfitting and enhance negation understanding through narrowing undesired concept bias.

2) Semantics Reversion Learning: Section IV-B1 presents a solution to reduce distribution shift in consistent semantics, yet the incorrect distribution consistency still exists in unrelated semantics. To overcome this challenge, we propose to discriminate these entangled distributions using the hardest negation. Specifically, for each $\hat{t}_i$, we first employ the LLM to construct a semantically reversed caption $\bar{t}_i$ by swapping its positive and negative components. For example, "a photo of a dog not on grass" would be reversed as "a photo of grass but not of a dog". Ideally, on the one hand, $\bar{t}_i$ is completely irrelevant to its visual candidate $\hat{v}_i$ since it describes absent objects while negating present ones. On the other hand, $\bar{t}_i$ maintains certain similarity to the batched-unpaired sample $\hat{v}_j$, since it correctly excludes those absent objects. Such learning objectives naturally are equivalent to minimizing the mutual information between representations of paired $\bar{t}$ and $\hat{v}$, which could be approximated as the negative InfoNCE [33]:

$$\min_{\tilde{\theta}_T} \mathbb{E}_{p(\bar{t},\hat{v})} \left[\bar{\boldsymbol{t}}\hat{\boldsymbol{v}}^\top/\tau_2 - \log \mathbb{E}_{p(\hat{v}')} \exp(\bar{\boldsymbol{t}}\hat{\boldsymbol{v}}'^\top/\tau_2))\right], \tag{6}$$

where the pair $(\bar{t}, \hat{v})$ is sampled from the joint distribution $p(\bar{t}, \hat{v})$ and the pair $(\bar{t}, \hat{v}')$ is built by samples independently sampled from the marginal $p(\hat{v}')$, and τ_2 is a temperature parameter.

As illustrated in Fig. 4, although minimizing Eq.(6) can force the paired samples to be distant, bringing the reversed text closer to all unpaired visual samples would corrupt the learned semantic structure. To prevent the clustering of unrelated visual samples, we propose to sample only the hardest visual sample $\hat{v}_i^-$ that shares minimal similarity with the reversed text $\bar{t}_i$. As $\bar{t}_i$ provides partial supervision by specifying non-existent concepts, it is reasonable to align $\bar{t}_i$ with the most irrelevant visual content. Then the semantics

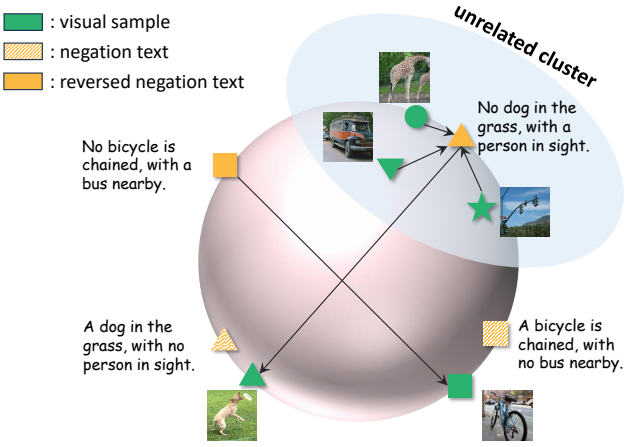


Fig. 4. A toy example to show the challenge of semantic-reversed contrastive learning. Although the negative InfoNCE pushes the reversed negation-conditioned text away from the target image, it relatively brings unrelated images closer together, thus corrupting the learned semantic structure.

reversion objective for the online batch is:

$$\begin{aligned} \min_{\tilde{\theta}_T} \mathcal{L}_{sr} &= \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \log \left[\frac{\exp(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^\top/\tau_2)}{\exp(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^\top/\tau_2) + \exp(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^{-\top}/\tau_2)} \right] \\ &= \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \log \left[\frac{1}{1 + \exp\left[(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^{-\top} - \bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^\top)/\tau_2\right]} \right] \\ &= \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} -\log\left[1 + \exp\left[(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^{-\top} - \bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^\top)/\tau_2\right]\right] \\ &= \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \Bigg[(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^\top - \bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^{-\top})/\tau_2 \\ &\qquad - \underbrace{\log\left[1 + \exp\left[(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^\top - \bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^{-\top})/\tau_2\right]\right]}_{\text{vanishing term}} \Bigg] \\ &\simeq \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} [(\bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^\top - \bar{\boldsymbol{t}}_i \hat{\boldsymbol{v}}_i^{-\top})/\tau_2]. \end{aligned} \tag{7}$$

For the reversed text $\bar{\boldsymbol{t}}_i$, our objective aims to decrease its similarity with the paired visual sample $\hat{\boldsymbol{v}}_i$, while increasing that with the hardest negative $\hat{\boldsymbol{v}}_i^-$. This gradually drives the vanishing term toward zero, allowing the objective to be approximated as a simple metric loss. Eq.(7) forces the reversed text to provide stronger negative supervision to its corresponding visual sample, which can eliminate the undesired distribution consistency.

3) Textual Dual-Concept Debiasing: The above training objectives mitigate the dual-concept shifts in cross-modal alignment. We further strengthen this debiasing in textual modality by directly regulating the triplet semantic gap:

$$\min_{\tilde{\theta}_T} \mathcal{L}_{tri} = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \left[\lambda \|\hat{\boldsymbol{t}}_i - \hat{\boldsymbol{t}}_i^P\|_2 + (2 - \|\hat{\boldsymbol{t}}_i - \bar{\boldsymbol{t}}_i\|_2)\right], \tag{8}$$

where λ is the balance hyperparameter. In Eq.(8), the first term brings the negation-conditioned caption closer to its affirmative, while the second term encourages maximum distance

between the negation-conditioned caption and its reversed counterpart.

V. EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed method through extensive experiments, comparing it with both state-of-the-art post-training and test-time adaptation methods. For a comprehensive comparison, our experiments are conducted across diverse visual domains, including images, videos, and medical imaging.

TABLE I
SUMMARY OF DATASETS AND TASKS FOR NEGATION UNDERSTANDING.

Dataset	Task	Size	Negation Type
COCO	Retrieval-Neg	5,000	LLM-Rephrased
	MCQ-Neg	5,914	LLM-Rephrased
VOC2007	MCQ-Neg	5,032	LLM-Rephrased
MSR-VTT	Retrieval-Neg	1000	LLM-Rephrased
	MCQ-Neg	1,000	LLM-Rephrased
CheXpert	Binary Classification-Aff	2,352	Templated
	Binary Classification-Neg	616	Templated

A. Experiment Setting

1) Datasets: Following the recent NegBench [2] evaluations, we test NEAT on four datasets: COCO, VOC2007, MSR-VTT, and CheXpert, each containing one or more well-designed negation understanding tasks. Specifically,

- **Retrieval-Neg task** evaluates whether VLMs can tackle real-world queries that mix affirmative and negative statements, such as 'a street without cars' or 'flowers that are not red'. This task challenges the model to not only identify the present elements but also exclude specific content, which could simulate more fine-grained retrieval in search engines and recommendation systems.
- **MCQ-Neg task** requires VLMs to select the correct description of the given image from multiple closely related choices. Based on the linguistic template, alternative descriptions could be categorized as Affirmation, Negation, and Hybrid. This task challenges the model to parse subtle yet critical differences among hard negatives.
- **Binary Classification-Neg task** tests whether VLMs can correctly interpret negation words by requiring models to distinguish the presence or absence of specific elements. This task is essential in medical diagnostics. For instance, 'The X-ray shows no evidence of pneumonia' is typical for ruling out lung pathologies.

For clarity, we summarize datasets and tasks for negation understanding in Table I, and we also show some cases in Fig. 7. In brief, the original validation set of MS-COCO [28] contains 5,000 images, where each is described by five captions. We adopt the LLM-rephrased variants from NegBench [2], comprising 5,000 Retrieval-Neg samples and 5,914 MCQ-Neg samples. VOC2007 [11] is an image dataset with 20 visual objects. We use the LLM-generated dataset from NegBench that contains 5,032 MCQ-Neg samples. The original test set of MSR-VTT [46] contains 2,990 video-caption pairs, where each video is captioned with 20 different descriptions. We adopt the LLM-rephrased variants from NegBench, comprising 1,000 Retrieval-Neg samples and 1,000 MCQ-Neg samples. Note that each video in Retrieval-Neg only has one corresponding caption. CheXpert [19] is a large dataset of chest X-rays, where the validation set has 234 images with expert radiologist annotations. Since each image contains multiple disease labels, we select 4 representative diseases, *i.e.*, Atelectasis, Cardiomegaly, Consolidation, and Lung Opacity, to create 616 negation classification pairs and 2,352 affirmative classification pairs as the base comparison. Moreover, we evaluate the zero-shot transfer ability for negation understanding on 9 widely used image datasets, including CIFAR10 [24], CIFAR100 [24], Caltch101 [12], OxfordPets [35], RESISC45 [6], EuroSAT [17], STL10 [7], SUN397 [44], and ImageNet1K [9].

2) Implementation Details: NEAT is a general TTA framework that could enhance most off-the-shelf pre-trained VLMs with negation understanding capabilities. Therefore, we select OpenAI CLIP [36], CoNCLIP [39], and NegCLIP [49] as the source models, where CoNCLIP and NegCLIP are two strong baselines designed to handle negation queries. To demonstrate the effectiveness of our method, we compare our NEAT with both post-training methods and state-of-the-art test-time adaptation methods. Following [26], NEAT updates the parameters within the Layer Normalization (LN) layers in the textual encoder f_{θ_T} using the AdamW optimizer. All TTA baselines are conducted with a batch size of 256. The temperatures are set as τ_1 = 0.03/0.07 and τ_2 = 0.07/1.0 for image/video tasks, respectively. The balanced parameter λ is set to 5 for all experiments. The trade-off parameter α is set to 1 for all experiments. To highlight practical adaptation during inference, we benchmark with offline and online updates to handle different scenarios.

B. Comparisons with State of the Arts

1) Compared to Post-training Methods: In this section, we compare our NEAT with post-training methods that fine-tune VLMs on labeled negation-enhanced datasets. Specifically, CC-Neg [39] is a synthetic dataset of 0.23 million image-caption pairs along with high-quality negated captions as distractors. NegFull [2] is a large-scale dataset of 70 million pairs generated from CC12M [5], where each image is captioned with multiple incorporated negated objects and hard negatives based annotations. For fair comparison with offline baselines, our NEAT applies offline adaptation where the model is first updated on all available data before proceeding to inference. In Table II, we present the cross-modal retrieval performance on MS-COCO and MSR-VTT datasets with negation queries. Note that our offline adaptation is conducted only on unlabeled retrieval-Neg data, and we also report the performance on normal retrieval tasks after TTA or post-training. From the results, we could observe the following conclusions:

- Despite being trained on massive data, CLIP suffers a 33.07% drop in terms of the sum in retrieval (*i.e.*, rSum) on MS-COCO and a 37.50% drop on MSR-VTT, showing that VLMs have difficulty in handling negation queries.

TABLE II
COMPARISONS WITH STATE-OF-THE-ART POST-TRAINING METHODS ON MS-COCO AND MSR-VTT DATASETS UNDER NEGATED QUERIES. ALL METHODS ARE BASED ON THE VIT-B/32 ARCHITECTURE FOR THE IMAGE ENCODER. THE BEST RESULTS ARE MARKED IN **BOLD**.

Method	Fine-tune Data	Retrieval-Neg Task							Retrieval Task						
		Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Evaluation Dataset: MS-COCO															
CLIP	None	45.30	69.78	79.10	24.96	47.93	59.40	326.47	49.98	75.02	83.30	30.37	54.76	66.11	359.54
	CC12M$\sim 12M$	47.66	72.60	81.28	29.00	53.71	64.72	348.97	**52.14**	75.38	**83.94**	**33.87**	**58.90**	**69.45**	**373.68**
	NegFull$\sim 70M$	44.48	69.32	78.40	28.17	51.91	62.98	335.26	48.60	72.86	81.08	30.07	54.19	65.32	352.12
+NEAT	None	48.98	**74.60**	**82.14**	30.00	54.63	65.66	356.01$\uparrow_{29.5}$	50.74	**76.02**	83.74	31.11	55.89	66.95	364.45
	CC12M$\sim 12M$	**49.04**	73.70	81.70	**31.91**	**56.70**	**67.72**	**360.77**$\uparrow_{11.8}$	50.96	74.54	82.88	33.25	58.22	69.08	368.93
ConCLIP	CC-Neg$\sim 0.23M$	43.04	66.64	76.12	24.30	48.22	59.83	318.15	45.72	68.44	76.16	27.20	51.93	63.50	332.95
NegCLIP	None	51.64	77.92	86.06	37.02	64.39	74.77	391.80	56.82	80.70	88.16	41.60	68.73	78.91	414.92
	CC12M$\sim 12M$	54.44	79.72	86.94	38.35	65.38	75.66	400.49	58.68	82.56	89.38	42.14	69.13	78.99	420.88
	NegFull$\sim 70M$	**56.26**	80.46	**88.46**	**40.29**	67.01	76.90	409.38	**59.68**	**83.36**	**90.06**	**42.66**	**69.05**	**79.06**	**423.87**
+NEAT	None	54.26	79.20	87.20	39.21	66.19	76.49	402.55$\uparrow_{10.8}$	56.26	80.26	88.04	40.59	67.81	78.08	411.04
	CC12M$\sim 12M$	**56.26**	**80.94**	88.38	40.21	**67.36**	**77.53**	**410.68**$\uparrow_{10.2}$	58.60	82.32	89.42	42.15	69.01	78.86	420.36
Evaluation Dataset: MSR-VTT															
CLIP	None	20.30	39.80	49.90	23.90	45.80	56.70	236.40	25.50	48.40	60.90	28.00	50.60	60.50	273.90
	CC12M$\sim 12M$	21.70	43.70	55.60	23.80	48.00	58.70	251.50	25.20	46.80	57.40	29.10	**52.30**	63.80	274.60
	NegFull$\sim 70M$	**25.20**	**50.10**	**59.50**	20.60	43.90	54.60	253.90	**29.00**	**52.00**	**63.90**	23.70	46.90	57.30	272.80
+NEAT	None	24.10	44.10	54.20	**24.80**	47.60	58.10	252.90$\uparrow_{16.5}$	26.10	48.10	59.50	**28.80**	50.00	58.80	271.30
	CC12M$\sim 12M$	22.70	45.30	56.90	24.50	**48.90**	**60.20**	**258.50**$\uparrow_{7.0}$	25.40	47.90	58.90	28.60	51.10	**62.80**	**274.70**
ConCLIP	CC-Neg$\sim 0.23M$	22.10	43.80	54.70	18.60	40.00	48.60	227.80	23.40	44.00	54.70	25.60	48.20	59.40	255.30
NegCLIP	None	19.80	41.00	52.90	25.10	51.00	61.70	251.50	24.00	48.20	59.30	29.80	53.70	64.80	279.80
	CC12M$\sim 12M$	23.30	46.50	58.50	27.00	53.60	**64.10**	273.00	27.80	51.80	63.80	30.00	**55.50**	**65.50**	294.40
	NegFull$\sim 70M$	**28.60**	**50.70**	**61.20**	**27.40**	51.60	63.00	**282.50**	**31.20**	**53.90**	**65.20**	**31.00**	54.30	64.70	**300.30**
+NEAT	None	23.10	46.20	55.70	26.60	50.80	61.10	263.50$\uparrow_{12.0}$	25.50	49.50	60.30	28.70	51.70	63.10	278.80
	CC12M$\sim 12M$	25.10	48.30	58.90	27.30	**54.10**	63.50	277.20$\uparrow_{4.2}$	28.40	53.10	63.90	29.20	54.00	65.20	293.80

TABLE III
COMPARISONS WITH STATE-OF-THE-ART ONLINE TTA METHODS ON MS-COCO AND MSR-VTT DATASETS UNDER NEGATED QUERIES. ALL METHODS ARE BASED ON THE VIT-B/32 CLIP. THE BEST RESULTS ARE MARKED IN **BOLD**.

Method	Retrieval-Neg Task							Zero-shot MCQ-Neg Task			
	Image-to-Text			Text-to-Image			rSum	MCQ Type			Total Acc
	R@1	R@5	R@10	R@1	R@5	R@10		Affirmed	Negated	Hybrid	
Evaluation Dataset: MS-COCO											
CLIP	45.30	69.78	79.10	24.96	47.93	59.40	326.47	69.09	6.84	39.21	39.25
*TENT	44.74	70.44	79.94	25.24	49.24	60.88	330.48$\uparrow_{4.01}$	45.23	15.72	17.40	26.43$\downarrow_{12.82}$
*SAR	45.00	70.28	79.86	25.19	48.93	60.63	329.89$\uparrow_{3.42}$	45.96	16.36	17.79	27.02$\downarrow_{12.23}$
*READ	43.44	68.04	77.54	24.93	47.61	59.23	320.79$\downarrow_{5.68}$	**68.85**	7.54	38.87	39.26$\uparrow_{0.01}$
*COME	44.80	69.42	79.28	25.47	48.56	60.40	327.93$\uparrow_{1.46}$	64.42	7.11	37.97	37.30$\downarrow_{1.95}$
*TCR	44.88	70.14	79.40	27.50	51.89	63.18	336.99$\uparrow_{10.52}$	45.37	15.56	21.17	27.71$\downarrow_{11.54}$
*NEAT	**46.58**	**72.20**	**80.80**	**28.12**	**52.08**	**63.69**	**343.47**$\uparrow_{17.00}$	64.94	**29.30**	**49.45**	**48.41**$\uparrow_{9.16}$
Evaluation Dataset: MSR-VTT											
CLIP	20.30	39.80	49.90	23.90	45.80	56.70	236.40	62.39	13.31	20.83	32.10
*TENT	19.90	42.50	52.60	25.00	46.20	57.30	243.50$\uparrow_{7.1}$	53.43	11.33	17.95	27.50$\downarrow_{4.6}$
*SAR	19.00	41.30	52.30	25.40	45.90	57.20	241.10$\uparrow_{4.7}$	50.45	11.33	17.63	26.40$\downarrow_{5.7}$
*READ	20.40	40.20	51.80	23.40	45.70	56.30	237.80$\uparrow_{1.4}$	63.28	13.31	19.87	32.10$\uparrow_{0.0}$
*COME	20.20	40.70	50.90	24.90	46.40	57.30	240.40$\uparrow_{4.0}$	56.72	13.03	18.59	29.40$\downarrow_{2.7}$
*TCR	19.70	41.90	51.10	**25.70**	46.80	**58.30**	243.50$\uparrow_{7.1}$	55.82	13.60	19.55	29.60$\downarrow_{2.5}$
*NEAT	**22.50**	**43.10**	**53.60**	24.90	**47.50**	57.30	**248.90**$\uparrow_{12.5}$	**73.43**	**15.86**	**31.41**	**40.00**$\uparrow_{7.9}$

- Our NEAT could significantly improve the negation comprehension ability of pre-trained VLMs, *e.g.*, it improves the rSum of CLIP by 11.80%$\sim$29.54% on MS-COCO.
- Our NEAT could generalize to different VLMs and modalities. For instance, it improves the rSum of NegCLIP and its CC12M fine-tuned model by 10.8% and 10.2% on MS-COCO, respectively. In addition, although only adapted on 1,000 video-text pairs, NEAT still improves the rSum of CLIP by 7.0%$\sim$16.5% on MSR-VTT.
- The normalization layers modified by NEAT do not impair VLMs' ability for normal vision-language comprehension on affirmative statements, with rSum performance remaining within stable and acceptable ranges, *i.e.* $-4.75\% \sim +4.91\%$.
- Remarkably, our NEAT achieves comparable or even superior performance to post-trained methods while avoiding the demand for massive negation data. Compared to the SOTA method that fine-tunes on 70 million well-designed negation-enriched pairs, *i.e.* NegFull, our NEAT surpasses its CLIP variant by a clear margin on MS-COCO, achieving much to 25.51% absolute improvement in rSum performance.

2) Compared to Test-time Adaptation Methods: In this section, we compare our NEAT with five SOTA TTA methods, *i.e.*, TENT [42], SAR [32], READ [47], COME, and TCR [26], under negated queries. Among the baseline methods, Tent, SAR, and COME are unimodal TTA approaches based on the entropy-minimizing objective or its variants, while READ and

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART ONLINE TTA METHODS ON MS-COCO AND MSR-VTT DATASETS UNDER NEGATED QUERIES. ALL METHODS ARE BASED ON THE VIT-B/32 NEGCLIP. THE BEST RESULTS ARE MARKED IN **BOLD**.

Method	Retrieval-Neg Task							Zero-shot MCQ-Neg Task			
	Image-to-Text			Text-to-Image			rSum	MCQ Type			Total Acc
	R@1	R@5	R@10	R@1	R@5	R@10		Affirmed	Negated	Hybrid	
Evaluation Dataset: MS-COCO											
NegCLIP	51.64	77.92	86.06	37.02	64.39	74.77	391.80	51.67	16.15	17.15	28.69
*TENT	53.74	79.70	88.24	38.65	66.55	77.20	$404.08\uparrow_{12.28}$	41.78	17.86	8.80	$23.00\downarrow_{5.69}$
*SAR	53.36	79.38	87.86	39.18	66.95	77.27	$404.00\uparrow_{12.20}$	41.34	19.30	8.75	$23.28\downarrow_{5.41}$
*READ	52.30	78.66	87.00	37.32	65.10	76.06	$396.44\uparrow_{4.64}$	51.82	16.15	17.25	$28.78\uparrow_{0.09}$
*COME	51.38	77.42	86.08	37.25	65.10	76.11	$393.34\uparrow_{1.54}$	44.34	15.40	12.28	$24.28\downarrow_{4.41}$
*TCR	**54.02**	**79.98**	**88.32**	39.01	**67.00**	77.54	**405.87**$\uparrow_{14.07}$	41.04	18.50	8.75	$22.93\downarrow_{5.76}$
*NEAT	53.68	79.40	88.02	**39.09**	66.99	**77.78**	$404.96\uparrow_{13.16}$	**52.12**	**28.45**	**34.79**	**38.74**$\uparrow_{10.05}$
Evaluation Dataset: MSR-VTT											
NegCLIP	19.80	41.00	52.90	25.10	51.00	61.70	251.50	51.94	12.46	17.63	27.30
*TENT	19.90	41.50	52.00	26.10	51.10	61.50	$252.10\uparrow_{0.6}$	41.19	10.20	11.54	$21.00\downarrow_{6.3}$
*SAR	20.40	43.00	53.70	25.90	51.30	61.90	$256.20\uparrow_{4.7}$	43.58	11.33	12.50	$22.50\downarrow_{4.8}$
*READ	21.40	42.90	54.00	25.00	51.20	61.40	$255.90\uparrow_{4.4}$	54.03	12.46	18.91	$28.40\uparrow_{1.1}$
*COME	21.30	42.50	54.40	25.10	51.30	62.10	$256.70\uparrow_{5.2}$	49.85	11.61	17.31	$26.20\downarrow_{1.1}$
*TCR	21.20	42.80	54.20	25.40	51.40	62.30	$257.30\uparrow_{5.8}$	48.66	12.75	17.31	$26.20\downarrow_{1.1}$
*NEAT	**23.80**	**45.30**	**56.80**	**27.20**	**52.50**	**62.80**	**268.40**$\uparrow_{16.9}$	**69.85**	**15.86**	**27.88**	**37.70**$\uparrow_{10.4}$

TABLE V
COMPARISONS WITH STATE-OF-THE-ART OFFLINE TTA METHODS ON MS-COCO DATASETS UNDER NEGATED QUERIES. ALL METHODS ARE BASED ON THE VIT-B/16 BLIP. THE BEST RESULTS ARE MARKED IN **BOLD**.

Method	Retrieval Neg Task							Zero-shot MCQ-Neg Task			
	Image-to-Text			Text-to-Image			rSum	MCQ Type			Total
	R@1	R@5	R@10	R@1	R@5	R@10		Affirmed	Negated	Hybrid	
Base model: ViT-B/16 BLIP											
BLIP	52.60	77.90	85.76	33.58	58.96	69.68	378.48	33.81	16.68	5.12	18.63
*TENT	37.96	66.98	79.50	40.54	67.16	78.33	$370.47\downarrow_{8.01}$	19.14	20.70	2.68	$14.03\downarrow_{4.60}$
*SAR	37.24	67.74	78.94	40.35	68.06	78.04	$370.37\downarrow_{8.11}$	11.71	23.53	1.09	$11.84\downarrow_{6.79}$
*READ	55.24	79.56	87.36	36.32	62.90	73.27	$394.65\uparrow_{16.17}$	35.73	16.52	5.62	$19.41\uparrow_{0.78}$
*COME	44.84	70.80	79.64	37.11	63.78	74.17	$370.34\downarrow_{8.14}$	31.74	17.59	3.68	$17.72\downarrow_{0.91}$
*TCR	48.38	76.10	85.34	**46.10**	**73.26**	**82.39**	$411.57\uparrow_{33.09}$	20.62	31.76	5.47	$18.99\uparrow_{0.36}$
*NEAT	**59.22**	**83.00**	**89.98**	41.12	68.40	78.47	**420.19**$\uparrow_{41.71}$	**77.21**	**28.34**	**49.95**	**52.49**$\uparrow_{33.86}$
Base model: ViT-B/16 BLIP Fine-tuned on MS-COCO Training Data											
BLIP	69.98	90.98	95.00	54.69	79.86	87.34	477.85	50.69	12.94	16.65	27.17
*TENT	68.14	89.46	94.10	57.44	82.02	88.85	$480.01\uparrow_{2.16}$	47.79	19.36	28.28	$32.16\uparrow_{4.99}$
*SAR	66.40	88.10	93.50	57.51	82.13	88.90	$476.54\downarrow_{1.31}$	47.29	19.36	29.97	$32.57\uparrow_{5.40}$
*READ	70.40	91.28	95.22	53.01	78.62	86.34	$474.87\downarrow_{2.98}$	53.00	11.76	17.20	$27.78\uparrow_{0.61}$
*COME	69.38	90.62	94.96	57.63	81.81	88.75	$483.15\uparrow_{5.30}$	55.81	11.12	16.90	$28.44\uparrow_{1.27}$
*TCR	68.58	89.66	94.12	**58.78**	**83.10**	**89.66**	$483.90\uparrow_{6.05}$	47.98	**19.52**	30.07	$32.89\uparrow_{5.72}$
*NEAT	**74.18**	**92.40**	**95.98**	56.40	81.11	88.47	**488.54**$\uparrow_{10.69}$	**58.27**	18.29	**46.22**	**41.53**$\uparrow_{14.36}$

TCR are multi-modal TTA approaches that further mitigate the modality bias. As entropy-based TTA methods are sensitive to the temperature parameter, we maintain the same τ_1 value as NEAT for a fair comparison. All methods are evaluated through online adaptation, where LN layers are updated on the current batch and then make a prediction. In detail, all methods are adapted on two unlabeled Retrieval-Neg datasets, *i.e.*, MS-COCO and MSR-VTT, and report the retrieval performance. To further investigate the generalization ability, we evaluate the classification accuracy of the adapted VLM on the corresponding MCQ-Neg task. As shown in Table III and Table IV, one could conclude that NEAT remarkably boosts the negation understanding of the baselines. More specifically,

- Most existing TTA methods only achieve marginal improvements over the base model, indicating that negated concepts pose challenging distribution shift problems.
- Although some multi-modal methods show considerable performance gains on Retrieval-Neg Task, *i.e.*, TCR, the accuracy on the corresponding MCQ-Neg task drops unexpectedly by 1.10% to 11.54%. This indicates that the observed improvements may be attributed to reducing modality gaps rather than addressing the concept shift problem arising from negation understanding.
- The extensions with NEAT achieve remarkably superior overall performance compared to all baselines under all settings. For example, on the MS-COCO dataset with the CLIP-based model, our NEAT surpasses the SOTA baseline by 6.48% on the Retrieval-Neg task and 20.7% on the more challenging zero-shot MCQ-Neg task. It also outperforms the SOTA baseline by 11.1% on Retrieval-Neg and 11.5% on MCQ-Neg for the MSR-VTT dataset with the NegCLIP-based model, indicating that our NEAT can generalize effectively across different scenarios.

In addition to the evaluation for CLIP-based VLMs, we also conduct experiments for BLIP-based VLMs [27], *i.e.*, a base model pretrained on 129 million data and its fine-tuned model on the MS-COCO dataset. From Table V, one could see that some TTA baselines show large performance

TABLE VI
ZERO-SHOT TRANSFER EVALUATION ON TWO TEMPLATE-BASED NEGATION CLASSIFICATION TASKS. THE BEST RESULTS ARE MARKED IN **BOLD**.

Method	CIFAR10	CIFAR100	Caltch101	OxfordPets	RESISC45	EuroSAT	STL10	SUN397	ImageNeg1K	AVERAGE
Negation-conditioned Zero-shot Classification: Top-1 Accuracy↑										
CLIP	64.44	49.30	75.00	61.59	44.75	25.13	77.41	47.56	45.97	54.57
NegCLIP	66.13	44.51	61.89	59.80	30.59	26.39	79.60	33.99	35.30	48.69
ConCLIP	74.95	19.84	22.00	26.79	15.05	15.04	89.79	26.42	18.06	34.22
NegCLIP+NegFull FT	**88.40**	56.68	**79.29**	67.70	**52.73**	**38.94**	95.16	**53.09**	**49.74**	**64.64**
CLIP+TCR	66.74	49.22	75.49	60.75	36.92	37.87	78.03	43.35	39.51	54.21
CLIP+NEAT	87.22	**57.96**	74.34	**71.90**	49.62	33.24	**95.59**	50.87	47.90	63.18
Reversed Negation-conditioned Zero-shot Classification: Top-1 Error Rate↓										
CLIP	22.83	8.29	16.07	5.70	12.13	9.80	19.13	3.81	5.01	11.42
NegCLIP	23.25	10.95	23.38	10.66	15.38	12.37	17.70	9.29	7.55	14.50
ConCLIP	9.97	1.72	2.86	4.42	1.62	1.54	0.53	0.40	0.28	2.59
NegCLIP+NegFull FT	2.47	1.31	1.53	0.74	4.22	8.94	0.34	0.19	0.17	2.21
CLIP+TCR	14.44	7.81	12.46	10.58	13.27	19.19	12.76	6.65	7.54	11.63
CLIP+NEAT	**2.24**	**0.44**	**1.60**	**0.22**	**0.43**	**6.85**	**0.33**	**0.07**	**0.04**	**1.36**

TABLE VII
COMPARISONS WITH MLLMS OF DIFFERENT SCALES ON MCQ-NEG.

Method	Zero-shot MCQ-Neg Task			
	Affirmation	Negation	Hybrid	Total Acc
Multimodal Large Language Model				
Qwen3VL-2B	83.56	16.15	43.84	48.73
Qwen3VL-4B	89.22	78.72	78.83	82.36
Qwen3VL-8B	91.44	81.71	86.03	86.52
Qwen3VL-32B	87.16	75.08	74.20	78.93
Multimodal Embedding Model				
CLIP+NEAT-0.15B	64.94	29.30	49.45	48.41
BLIP+NEAT-0.25B	77.21	28.34	49.95	52.49

TABLE VIII
ZERO-SHOT TRANSFER EVALUATION OF TTA METHODS ON THE VOC2007 MCQ-NEG TASK AFTER ADAPTATION ON UNLABELED MS-COCO RETRIEVAL-NEG DATA.

Method	MCQ-Neg Type			Total Acc
	Affirmed	Negated	Hybrid	
CLIP	82.70	3.30	58.93	38.70
*TENT	60.12	5.14	38.51	$27.37\downarrow_{11.33}$
*SAR	61.14	5.09	39.59	$27.97\downarrow_{10.73}$
*READ	80.79	3.87	58.08	$38.30\downarrow_{0.40}$
*COME	79.33	2.92	59.07	$38.14\downarrow_{0.56}$
*TCR	61.00	5.37	45.74	$30.79\downarrow_{7.91}$
*NEAT	**81.09**	**23.86**	**77.87**	**55.54**$\uparrow_{16.84}$
NegCLIP	75.37	5.79	40.31	30.47
*TENT	68.48	6.03	31.10	$25.60\downarrow_{4.87}$
*SAR	68.62	6.08	31.96	$26.02\downarrow_{4.45}$
*READ	76.25	5.70	41.07	$30.93\uparrow_{0.46}$
*COME	73.02	5.52	36.36	$28.32\downarrow_{2.15}$
*TCR	68.62	6.13	32.05	$26.08\downarrow_{4.39}$
*NEAT	**77.13**	**18.34**	**71.23**	**49.73**$\uparrow_{19.26}$

variations between image-to-text and text-to-image retrieval after adaptation. This may be because BLIP uses multiple pretraining objectives while adaptation only focuses on the image-text contrastive one. The phenomenon is significantly alleviated when the model is fine-tuned on corresponding image-text retrieval tasks. Moreover, like the observations on CLIP-based models, our NEAT remarkably boosts the effectiveness of the baselines. Specifically, NEAT improves BLIP by 41.71% (rSum) for cross-modal retrieval and 33.86% (accuracy) for multiple choice questions.

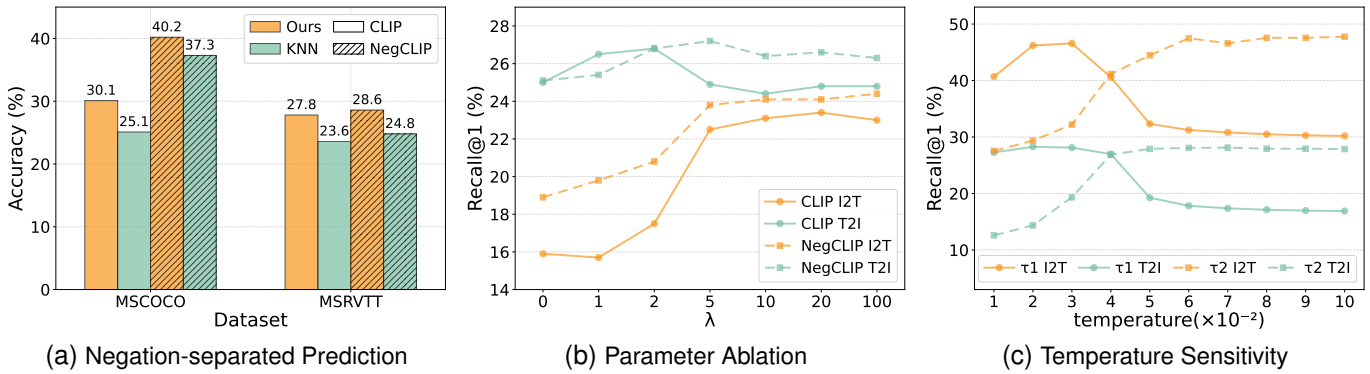
3) Compared to Multimodal Large Language Models: To further understand the limit of advanced multimodal systems in negation understanding, we compare our method with strong MLLMs on the zero-shot MCQ-Neg task. As shown in Table VII, MLLMs consistently achieve lower accuracy on the negation type than the affirmation type across all scales, with drops ranging from 9.3% (Qwen3VL-8b [4]) to 67.4% (Qwen3VL-2b). Notably, after applying NEAT on the Retrieval-Neg data, lightweight embedding VLMs can achieve comparable or superior performance to 2B-parameter MLLM. For example, CLIP with NEAT achieves a 13.15% improvement on the negation type despite being 10× smaller in model size. This highlights the practical value of our method in computation-intensive scenarios, such as drone-based rescue missions, where efficiently matching negation queries against thousands of images is necessary.

C. Generalizability Analysis

In this section, we investigate whether NEAT can learn generalizable negation patterns rather than simply fitting to the negation data. To this end, we conduct a series of studies of increasing difficulty to verify the performance of NEAT-adapted VLMs on unseen negation data and tasks. Unless otherwise stated, all TTA methods are adapted on the MS-COCO Retrieval-Neg dataset using the CLIP ViT-B/32 model.

1) Easy: Cross-Dataset Generalization: We first evaluate VLMs that were adapted by different TTA methods on the unseen VOC2007 MCQ-Neg task. As demonstrated in Table VIII, although some TTA baselines can adapt to negation data and enhance performance on such data, the adapted models fail to generalize to novel negation data, leading to substantial performance drops. In contrast, our NEAT remarkably improves CLIP from 38.70% to 55.54% and NegCLIP from 30.47% to 49.73% on the unseen voc2007 dataset.

2) Normal: Cross-Task Generalization: We further test NEAT-adapted VLMs across 9 widely used image classification datasets using two template-based negation understanding tasks. As elaborated in IV-A, we randomly introduce distrac-



(a) Negation-separated Prediction (b) Parameter Ablation (c) Temperature Sensitivity

Fig. 5. Finer-grained ablation studies. (a) The top-1 prediction accuracy on the vanilla KNN-based prediction and our negation-separated prediction. (b) The parameter analysis of λ on MSR-VTT. (c) The parameter analysis of τ_1 and τ_2 on MS-COCO.

TABLE IX
ZERO-SHOT TRANSFER EVALUATION ON CHEXPERT BY REPLACING LN LAYERS FROM DIFFERENT NEGATION-AWARE MODELS.

Method	Atelectasis		Cardiomegaly		Consolidation		Lung Opacity	
	Aff	Neg	Aff	Neg	Aff	Neg	Aff	Neg
PLIP	67.49	**84.38**	21.43	27.94	41.90	**90.91**	84.38	34.13
CONCLIP	66.67	83.75	20.85	27.21	41.43	**90.91**	89.00	41.67
PLIP NegFull	66.34	83.75	20.85	27.21	41.43	**90.91**	84.97	34.13
PLIP TCR	75.08	58.75	**44.40**	**44.12**	36.67	77.27	84.58	42.46
PLIP NEAT	**80.53**	**84.38**	33.20	31.62	**61.43**	89.39	**98.92**	**70.24**

TABLE X
ABLATION STUDY OF THE PROPOSED TRAINING OBJECTS.

$\mathcal{L}_{ent}$	$\mathcal{L}_{sr}$	$\mathcal{L}_{tri}$	I2T R@1	T2I R@1	MCQ
			45.30	24.96	39.25
✓			48.06	29.20	28.69
	✓	✓	7.12	4.67	42.78
✓		✓	49.74	28.95	27.26
✓	✓		44.96	**30.05**	45.91
✓	✓	✓	**48.98**	30.00	**48.41**

tor labels to construct the negation-conditioned caption, and swap its two class names to construct the reversed negation-conditioned caption. The comparison results are shown in Table VI. From the table, one could see that most negation-aware methods cannot generalize to downstream image classification tasks, performing even worse than the base model. Additionally, post-training on millions of well-designed negation data can endow VLMs with strong negation generalization capabilities, but the performance remains sub-optimal for the more difficult reversed negation-conditioned task. Remarkably, although only updating LN layers with minimal unlabeled data, our approach not only improves the accuracy of CLIP from 54.57% to 63.18% (average) for the negation-conditioned task, but also reduces the error rate of CLIP from 11.42% to 1.36% (average) for the reversed negation-conditioned task.

3) Hard: Cross-Normalization Layers Generalization: Beyond validating on complete models, we further investigate whether the adapted LN layers can be directly applied to domain-specific models. To this end, we use the medical foundation VLM, *i.e.*, PLIP [18], as the base model, replacing its textual LN layers with those from negation-aware VLMs to observe the performance impact. Following [2], we conduct evaluations on two classification settings. Specifically, the Binary Classification-Aff task requires the model to correctly associate each image with the two affirmative statements, such as 'This image shows Atelectasis' and 'This image shows Consolidation'. The Binary Classification-Neg task requires the model to distinguish each image with the presence or absence of a medical condition, such as 'This image shows Atelectasis' and 'This image does not show Atelectasis'. As shown in Table IX, one can observe that PLIP exhibits robust negation understanding abilities in most disease recognition tasks, *e.g.*, Atelectasis and Consolidation, suggesting that its training medical reports likely contain explicit negation expressions. Second, using the LN layers from post-training methods does not lead to significant performance changes, indicating that their negation understanding capabilities are more attributable to the encoder or decoder layers. Third, using the LN layers adapted by TCR brings performance gains in certain scenarios, but exhibits instability in others. Notably, our NEAT provides stable performance improvements overall, especially for the most frequent [2] condition Lung Opacity, *i.e.*, achieving 14.54% and 36.11% improvements for Binary Classification-Aff and Classification-Neg tasks, respectively.

D. Ablation Study

1) Impact of Each Component: To study the influence of specific components in our method, we first carry out ablation studies on the MS-COCO Retrieval-Neg data with different training objects. Table X reports the performance of the Retrieval-neg and the MCQ-Neg tasks. From the results, we observe the following conclusions: 1) $\mathcal{L}_{ent}$ boosts performance through the negation-separated prediction, *i.e.*, improving the R@1 of the base CLIP by 2.76% and 4.24% on text and image retrieval, respectively. However, naive entropy minimization fails to generalize the model to more difficult negation understanding scenarios, *e.g.*, performance drops significantly on the MCQ task. 2) $\mathcal{L}_{sr}$ is crucial for learning generalizable negation understanding capabilities, increasing the MCQ accuracy of the $\mathcal{L}_{ent}$ variant from 28.69% to 45.91%. 3) NEAT achieves overall optimal performance when all the loss terms are employed, showing that all three components make contributions. In addition, we ablate the Candidate Selection (CS) in NEAT to show its impact. Table XI reports

TABLE XI
ABLATION STUDY OF THE CANDIDATE SELECTION STRATEGY.

Method	Retrieval-Neg		Zero-shot MCQ-Neg			
	I2T R@1	T2I R@1	Aff	Neg	Hyb	Total
CLIP	45.3	25.0	**69.1**	6.8	39.2	39.3
NEAT w/o CS	**50.1**	**30.0**	57.2	24.0	38.4	40.3
NEAT	49.0	**30.0**	64.9	**29.3**	**49.5**	**48.4**
BLIP	52.6	33.6	33.8	16.7	5.1	18.6
NEAT w/o CS	49.8	36.4	25.2	18.7	3.9	15.9
NEAT	**59.2**	**41.1**	**77.2**	**28.3**	**50.0**	**52.5**

the offline performance based on CLIP and BLIP models. From the results, one could observe that entropy minimization on all samples leads to severe underfitting or overfitting. For instance, NEAT without candidate selection achieves comparable performance to NEAT on Retrieval-Neg, but drops by 8.1% on zero-shot MCQ (total) for CLIP and 36.6% for BLIP. We further compare different candidate selection strategies to verify the effectiveness of our negation-separated design. We further compare different candidate selection strategies to verify the effectiveness of our negation-separated method. Specifically, we compared with the KNN-based prediction [26] and report the zero-shot top-1 prediction accuracy. As shown in Fig. 5a, our refinement strategy substantially outperforms the KNN variant, which signifies that the simple feature-driven prediction is insufficient for handling negation.

2) Parameter Analysis: We next investigate the effect of the parameter λ by plotting the online Recall@1 scores with incremental values on MSR-VTT. As shown in Fig. 5b, we observe that: 1) when using smaller λ values, *e.g.*, 1 and 2, the second term in Eq.(8) would dominate the optimization objective, leading to sub-optimal model performance. 2) Our method can achieve stable performance in a relatively larger range, *i.e.* $5 \sim 20$, and shows no substantial performance variation even when using extremely large λ values.

3) Temperature Sensitivity: As TTA approaches are usually sensitive to temperature parameters, we further carry out experiments to investigate the influence of τ_1 and τ_2. The comparison results are shown in Fig. 5c. The figure shows that the adaptation process exhibits different sensitivities to τ_1 and τ_2. Specifically, the adapted model achieves optimal performance with smaller τ_1 values but with larger τ_2 values. Recall that our NEAT enhances negation understanding by reducing the dual-concept shifts, where τ_1 and τ_2 control the distribution shift in consistent and irrelevant semantics, respectively. In Eq.(5), sufficient contrastive targets enable us to utilize sharper distributions (smaller temperature) to enhance the discrimination of the model. However, in Eq.(7), we only sample one hardest contrastive candidate to avoid undesirable clustering. Thus, we need smoother distributions (larger temperature) to prevent overfitting to negative samples.

4) Analysis of Hardest Sampling: To further verify the effectiveness of the hardest sampling strategy in $\mathcal{L}_{sr}$, we vary the number of hardest negative samples from 1 to 256 and report the offline results on MS-COCO Retrieval-Neg and MCQ-Neg tasks. Specifically, we extend $\mathcal{L}_{sr}$ by selecting the top-K visual samples with the lowest similarity to the reversed text, and compute the loss by averaging over these K hardest negatives. As shown in Table XII, sampling only the hardest visual data achieves the best overall performance. Scaling K from one to all batch samples leads to a 1.5% drop in T2I Retrieval-Neg average and an 8.6% drop in MCQ-Neg total accuracy. Interestingly, the fine-grained results on MCQ reveal that the performance degradation is mainly attributed to the affirmation type, which drops from 64.9% to 56.6%. This observation aligns well with our analysis in Section IV-B2: contrasting against more visual samples would corrupt the learned semantic structure. In contrast, our proposed semantics reversion learning ($K = 1$) can avoid this issue while reducing computational overhead.

TABLE XII
EFFECT OF THE NUMBER OF HARDEST CONTRASTIVE SAMPLES IN THE SEMANTICS REVERSION LOSS.

Hardest	T2I Retrieval-Neg				Zero-shot MCQ-Neg			
Samples	R@1	R@5	R@10	Avg	Aff	Neg	Hyb	Total
1	**30.0**	**54.6**	**65.7**	**50.1**	**64.9**	29.3	**49.5**	**48.4**
2	**30.0**	54.5	65.6	50.0	64.0	30.1	48.2	47.9
4	29.9	54.4	65.5	49.9	63.1	30.9	46.6	47.3
8	29.8	54.2	65.5	49.8	62.4	31.1	44.7	46.5
16	29.7	54.1	65.3	49.7	61.3	**31.3**	43.1	45.6
32	29.5	53.8	65.0	49.4	60.7	31.1	40.9	44.6
64	29.2	53.6	64.9	49.2	60.1	30.8	38.5	43.5
128	28.9	53.2	64.5	48.9	58.8	31.0	36.1	42.3
256	28.5	52.8	64.1	48.5	56.6	29.6	32.5	39.8

E. Visualization and Analysis

1) Embedding Similarity Analysis: To visually investigate the performance of our NEAT against negation, we illustrate pairwise similarity distributions of image-text embeddings from different VLMs. Specifically, we compared our NEAT (adapted on unlabeled MS-COCO Retrieval-Neg) with CLIP, ConCLIP, NegFull-finetuned CLIP, and TCR (adapted on unlabeled MS-COCO Retrieval-Neg) in Fig. 6. For each image in the test set of CIFAR10, we contrast it with the normal text, *i.e.*, "a photo of the [CLASS]", the negation-conditioned text, *i.e.*, "a photo of the [CLASS] but not of the [CLASS′]", and the reversed negation-conditioned text, *i.e.*, "a photo of the [CLASS′] but not of the [CLASS]", where [CLASS] is the true label and [CLASS′] is a false one. As shown in Figs. 6a-6e, all VLMs could separate the positive and negative pairs apart enough. Although some methods are designed for negation understanding, *e.g.*, ConCLIP and NEAT, their ability to handle affirmative statements remains intact. When tasked with negation-conditioned understanding, Figs. 6f-6j demonstrate that some VLMs exhibit significant overlap between positive and negative distributions, eg, CLIP and ConCLIP. In contrast, our NEAT could discriminate the true and false negative pairs successfully. While for the more challenging reversed negation-conditioned text, Figs. 6k-6o show that CLIP and TCR wrongly provide higher similarity scores for positive (the most irrelevant) pairs. Although some negation-enhanced models, *e.g.*, ConCLIP and NegFull-finetuned CLIP, reduce similarity scores for positive pairs, they

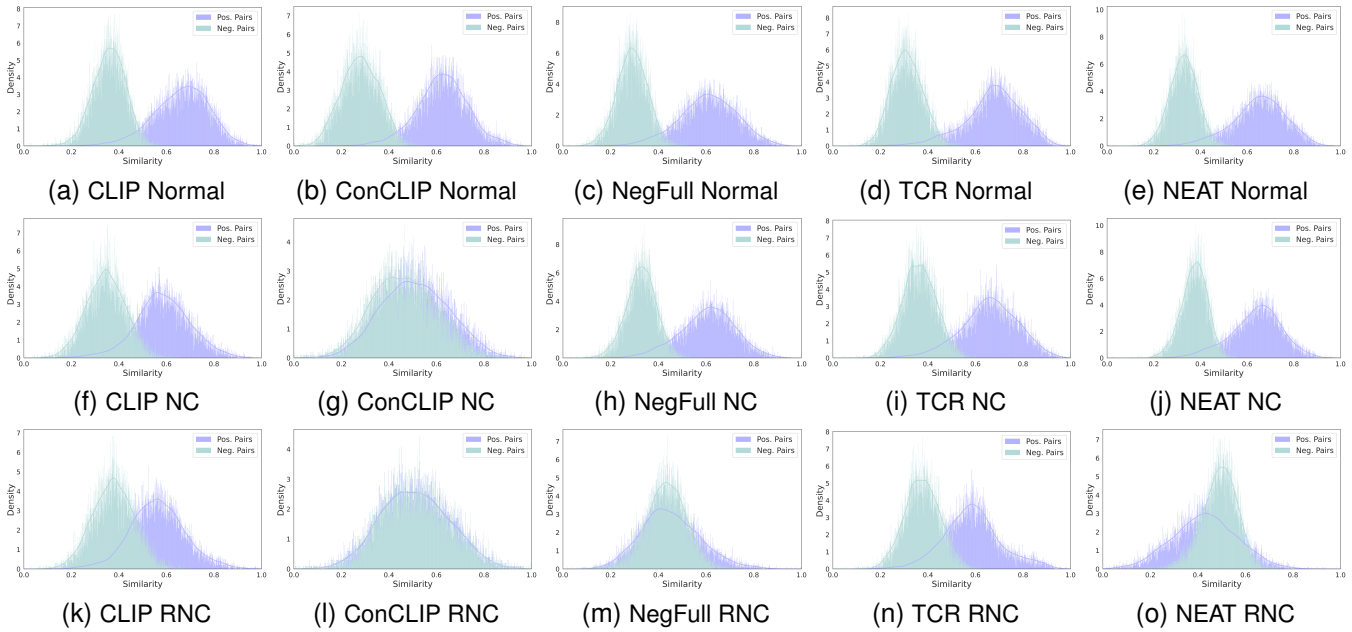


Fig. 6. Comparison of discrimination capability across different negation types. This figure shows the similarity distributions of positive and negative pairs (mean) from VLMs. Image-text similarity distributions for normal, negation-conditioned (NC), and reversed negation-conditioned text (RNC) on the CIFAR10 test set are shown in the top, middle, and bottom rows, respectively. Similarities are normalized to the 0-1 range for clearer presentation.

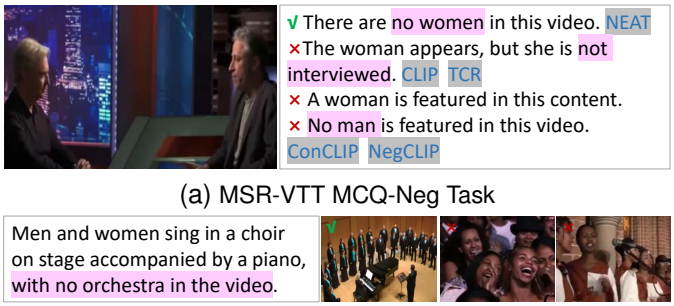


Fig. 7. Case studies of our NEAT in understanding negation. (a) Multiple choice questions with negated captions for a given video. (b) Top-3 retrieved videos for a given negated query. The negation parts are highlighted.

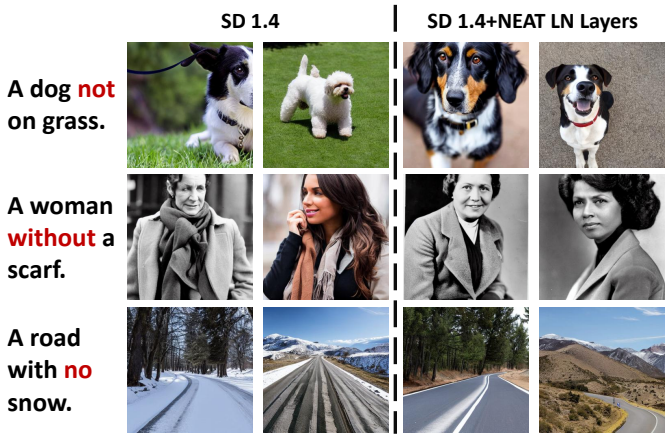


Fig. 8. Examples of text-to-image generation with negation prompt.

still maintain distributions comparable to those of negative pairs. Differently, our NEAT is the only approach that properly produces lower similarity scores for positives than negatives, which is consistent with the findings in Table VI.

2) Negation Understanding Examples: To visually illustrate the negation comprehension ability of our NEAT, we show some retrieved videos using negated queries and multiple choice questions with negated captions from MSR-VTT in Fig. 7, where TTA methods (TCR and NEAT) are adapted on unlabeled MSR-VTT Retrieval-Neg data. As shown in the MCQ results in Fig. 7a, most methods frequently misinterpret negation in either objects (*e.g.*, ConCLIP and NegCLIP) or actions (*e.g.*, CLIP and TCR). From the retrieved results in Fig. 7b, one could see that our NEAT is not misled by negation statements and successfully retrieves relevant samples. Overall, the above cases show that our NEAT can rapidly improve performance during test time.

3) Text-to-Image Generation with Negation: We further investigate whether NEAT can enhance negation understanding in generative multimodal models. Specifically, we replace the LN layers of the text encoder in Stable Diffusion v1.4 (SD1.4) [37] with those adapted by NEAT on the MS-COCO Retrieval-Neg data using CLIP ViT-L/14. As shown in Fig. 8, SD1.4 fails to exclude the negated content, while the NEAT-adapted variant generates correct images. This demonstrates that our NEAT adapted LN layers can generalize from discriminative to generative tasks, showing its potential to benefit broader multimodal systems built upon embedding VLMs.

VI. Conclusion

In this paper, we study enhancing the negation understanding ability of VLMs through test-time adaptation. The key idea is to address the dual-concept shifts problem between

affirmation and negation distributions. Specifically, we propose NEAT, a novel method that reduces distribution shift in consistent semantics while eliminating false distributional consistency in unrelated semantics. By adjusting only the lightweight normalization layers, our method could efficiently adapt VLMs to negation contexts during inference. Extensive experiments across multiple benchmarks spanning images, videos, and medical scenarios verify the effectiveness and generalization of our method. In future work, we plan to explore more negation understanding scenarios, *e.g.*, remote sensing image retrieval with negation-conditioned queries, and tasks, *e.g.*, visual question answering, broadening NEAT to handle these corresponding challenges.

ACKNOWLEDGMENTS

This work was supported in part by the Major Key Project of PCL under Grant PCL2025AS10 and PCL2024A06, and in part by the Shenzhen Science and Technology Program under Grant RCJC20231211085918010.

REFERENCES

[1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23922–23932, 2024.

[2] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29612–29622, 2025.

[3] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[7] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[8] Alex de Carvalho, Cécile Crimon, Axel Barrault, John Trueswell, and Anne Christophe. "look! it is not a bamoule!": 18-and 24-month-olds can use negative sentences to constrain their interpretation of novel word meanings. *Developmental science*, 24(4):e13085, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[12] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[13] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023.

[14] Meng-Hao Guo, Yi Zhang, Tai-Jiang Mu, Sharon X Huang, and Shi-Min Hu. Tuning vision-language models with multiple prototypes clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18962–18972, 2023.

[16] Uri Hasson and Sam Glucksberg. Does understanding negation entail affirmation?: An examination of negated metaphors. *Journal of Pragmatics*, 38(7):1015–1032, 2006.

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[18] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[21] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[22] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[23] Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. Mitigating test-time bias for fair image retrieval. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[25] Youngjun Lee, Doyoung Kim, Junhyeok Kang, Jihwan Bang, Hwanjun Song, and Jae-Gil Lee. Ra-tta: Retrieval-augmented test-time adaptation for vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[26] Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, Xiting Liu, and Mouxing Yang. Test-time adaptation for cross-modal retrieval with query shift. *arXiv preprint arXiv:2410.15624*, 2024.

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

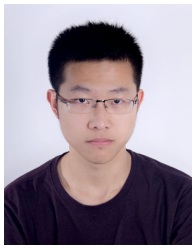
[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[29] Gen Luo, Yiyi Zhou, Minglang Huang, Tianhe Ren, Xiaoshuai Sun, and Rongrong Ji. Moil: Momentum imitation learning for efficient vision-language adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[30] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023.

[31] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[32] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023.
[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
[34] Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know" no" better: A data-driven approach for enhancing negation awareness in clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2825–2835, 2025.
[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.
[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
[38] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.
[39] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024.
[40] Eszter Szabó and Ágnes-Melinda Kovács. Infants' early understanding of different forms of negation. 2022.
[41] Alex Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. Enhancing visual grounding in vision-language pre-training with position-guided text prompts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3406–3421, 2023.
[42] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.
[43] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.
[44] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.
[45] Zehao Xiao, Jiayi Shen, Mohammad Mahdi Derakhshani, Shengcai Liao, and Cees GM Snoek. Any-shift prompting for generalization over distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13849–13860, 2024.
[46] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.
[47] Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*, 2024.
[48] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.
[49] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.
[50] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136, 2024.
[51] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.
[52] S Zhao, X Wang, L Zhu, and Y Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.
[53] Bin Zhu, Huiyan Qi, Yinxuan Gui, Jingjing Chen, Chong-Wah Ngo, and Ee-Peng Lim. Calling a spade a heart: Gaslighting multimodal large language models via negation. *arXiv preprint arXiv:2501.19017*, 2025.



Haochen Han received the B.S. degree from the School of Energy and Power Engineering, Chongqing University, in 2019, and the Ph. D. degree from the Department of Computer Science and Technology, Xi'an Jiaotong University, in 2024. He is currently a Research Associate in the Department of AI Computing, Pengcheng Laboratory. His research interests include robust multimodal learning and its applications, such as cross-modal retrieval and vision-language models.

Alex Jinpeng Wang is currently a Professor at Central South University. He received his Ph.D. from the National University of Singapore and his bachelor's and master's degrees from Sun Yat-Sen University (SYSU). His research interests include large-scale visual-language pre-training and data-centric AI. He has published extensively in top-tier conferences and journals, including CVPR, ICCV, AAAI, NeurIPS, ECCV, and IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). He also serves as a reviewer for CVPR, ICCV, ECCV, ICLR, NeurIPS, and TPAMI.

Fangming Liu received the B.Eng. degree from the Tsinghua University, Beijing, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong. He is currently a Full Professor with the Huazhong University of Science and Technology, Wuhan, China. His research interests include cloud computing and edge computing, datacenter and green computing, SDN/NFV/5G and applied ML/AI. He received the National Natural Science Fund (NSFC) for Excellent Young Scholars, and the National Program Special Support for Top-Notch Young Professionals. He is a recipient of the Best Paper Award of IEEE/ACM IWQoS 2019, ACM e-Energy 2018 and IEEE GLOBECOM 2011, the First Class Prize of Natural Science of Ministry of Education in China, as well as the Second Class Prize of National Natural Science Award in China.

Zhu Jun received his B.S. and Ph.D. degrees from the Department of Computer Science and Technology in Tsinghua University, where he is currently a Bosch AI professor. He was an adjunct faculty and postdoctoral fellow in the Machine Learning Department, Carnegie Mellon University. His research interest is primarily on developing machine learning methods to understand scientific and engineering data arising from various fields. He regularly serves as senior Area Chairs and Area Chairs at prestigious conferences, including ICML, NeurIPS, ICLR, IJCAI and AAAI. He was selected as "AI's 10 to Watch" by IEEE Intelligent Systems. He is a Fellow of the IEEE and an associate editor-in-chief of IEEE TPAMI.