Cross-Modal Masked Compositional Concept Modeling for Enhancing Visio-Linguistic Compositionality

Wei Li[1], **Zhen Huang**[2], **Xinmei Tian**[1†]
[1]MoE Key Laboratory of Brain-inspired
Intelligent Perception and Cognition, University of Science and Technology of China
[2]Independent Researcher
lwzkd@mail.ustc.edu.cn, xinmei@ustc.edu.cn

Abstract

Contrastively trained vision-language models like CLIP, have made remarkable progress in learning joint image-text representations, but still face challenges in compositional understanding. They often exhibit a "bag-of-words" behavior—struggling to capture the object relations, attribute-object bindings, and word order dependencies. This limitation arises not only from the reliance on global, single-vector representations for optimization, but also from the insufficient exploitation and modeling of the rich compositional information inherently present in paired image text data. In this work, we propose **MACCO** (**MA**sked **C**ompositional **C**oncept **MO**deling), a framework that masks compositional concepts in one modality and reconstructs them conditioned on the full contextual information from the other, enabling the model to capture and align cross-modal compositional structures more effectively. To facilitate this process, we introduce two auxiliary objectives that jointly align and regularize masked features both inter-modally and intra-modally. Extensive experiments on five compositional benchmarks, along with in-depth analyses, demonstrate that our approach not only significantly enhances compositionality in VLMs but also improves their ability to capture syntactic structure and linguistic information. Additionally, the improved compositionality also benefits text-to-image generation and multimodal large language model.

1 Introduction

Vision-language foundation models like CLIP (Radford et al., 2021) have significantly advanced multimodal learning by aligning images and texts in a shared semantic space via contrastive learning, and have been widely adopted in tasks such as image-text retrieval (Koukounas et al., 2024; Chen et al., 2023), VQA (Zhu et al., 2023; Liu et al., 2023), video understanding (Wasim et al., 2023), and text-to-image generation (Ramesh et al., 2022).

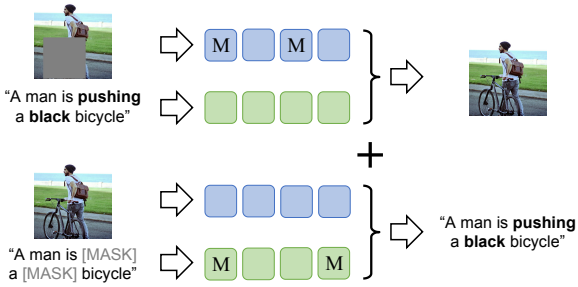

Figure 1: **The core idea of our method.** We mask compositional concepts in one modality and reconstruct them conditioned on the full information from the other.

However, compositional understanding remains a key limitation. These models often struggle with object relations, attribute-object bindings, and word order dependencies—frequently exhibiting "bag-of-words" behavior (Yuksekgonul et al., 2023; Thrush et al., 2022; Zhao et al., 2022; Hsieh et al., 2024). For instance, they tend to fails to distinguish between "the horse is eating grass" and "the grass is eating the horse" or between "a black dog with a white cat" and "a white dog with a black cat". Addressing this challenge is crucial for improving VLMs reasoning and facilitating their application in downstream tasks.

To enhance the compositional understanding capabilities of VLMs, most existing approaches focus on the careful construction of hard negative samples with subtle semantic variations, using rule-based templates (Yuksekgonul et al., 2023), LLM-generated captions (Doveh et al., 2023a), or synthetic scenes (Cascante-Bonilla et al., 2023). While effective, these methods are often costly, noisy, and lead the model to focus on superficial patterns specific to those negatives (Hsieh et al., 2024; Geirhos et al., 2020). Moreover, recent work (Kamath et al., 2024) shows that reliance on hard negatives may induce oversensitivity, causing models to rank semantically equivalent captions incorrectly. This

†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32284–32308

motivates an intriguing question: ***Beyond hard negative mining, can we improve compositionality of VLMs by designing a training framework that better exploits the rich aligned compositional information inherently present in existing image-text pairs?***

In this work, we introduce **MACCO** (**MA**sked **C**ompositional **C**oncept M**O**deling), a novel framework that enhances compositionality in VLMs without explicit hard negative construction. Our method masks compositional concepts in one modality and reconstructs it conditionally using full context from the other. As illustrated in Figure 1, the masked text and full image are used to reconstruct compositional concept words, while the masked image and full text are used to reconstruct image regions corresponding to compositional concepts. To better constrain global features during reconstruction and enrich local tokens with contextual global semantics, we introduce a parameter-free *global-to-local semantic injection* operation.

To facilitate this masked cross-modal reconstruction, we introduce two novel auxiliary objectives. First, the **Masked-augmented Cross-Modal Alignment Loss (MCA)** integrates global features of masked texts or masked images into the cross-modal contrastive learning process. Second, the **Masked-augmented Intra-Modal Regularization Loss (MIR)** regularizes the global features of masked instances within each modality to prevent representational collapse. Extensive experiments across five compositional benchmarks and four backbones demonstrate the effectiveness of our approach. In-depth analyses show that MACCO also enhances the model's ability to capture syntactic structure and semantic nuance. It produces more concept-aware embeddings, exhibits stronger robustness to semantically invariant perturbations, and better preserves fine-grained linguistic information. Moreover, MACCO can be integrated with hard negative mining methods to obtain additional gains. Finally, further experiments show that the improved compositionality also benefits text-to-image generation and multimodal large language models.

To summarize, our main contributions are:

1. We introduce a novel framework that improves vision-language compositionality in pre-trained VLMs without requiring explicit hard negative samples, and we show that the improved compositionality also benefits other multimodal tasks.
2. We propose two auxiliary objectives, MCA and MIR, to promote effective cross-modal reconstruction and alignment learning.
3. We validate the effectiveness of our approach through extensive experiments on five widely used vision-language compositional benchmarks, complemented by in-depth analyses. Our framework is also compatible with existing hard negative mining methods, yielding additional gains when integrated.

2 Related Works

Contrastive Vision-Language Models. Vision-language foundation models have achieved remarkable progress. Representative models such as CLIP (Radford et al., 2021), pretrained via contrastive learning on large-scale and noisy image-text datasets, exhibit impressive zero-shot transfer capabilities, leading to success across a wide range of tasks. Our motivation to focus on CLIP is twofold. First, contrastive learning has become a dominant and highly effective paradigm for multimodal representation learning. Second, CLIP-like models serve as the foundation of numerous applications, showcasing wide applicability across diverse domains. Enhancing CLIP is therefore of significant value, as improvements can benefits to a broader range of vision-language applications.

Vision-Language Compositionality. Despite significant progress, vision-language models such as CLIP still struggle with compositional reasoning—understanding fine-grained relations, attributes, and word order beyond object recognition (Yuksekgonul et al., 2023; Hsieh et al., 2024; Zhao et al., 2022; Thrush et al., 2022). To enhance compositionality, most prior work focuses on fine-tuning with hard negative samples, using strategies such as rule-based construction (Yuksekgonul et al., 2023), LLM-based synthesis (Doveh et al., 2023a), and negative image synthesis via diffusion models (Li and Li, 2025). Beyond these, SDS-CLIP (Basu et al., 2024) introduces a novel distillation loss from Stable Diffusion to improve the compositionality of CLIP. CLIP-CAE (Li et al., 2024) enhances the model's attention to compositional concepts by explicitly optimize internal attribution.

Masked Signal Modeling. Masked reconstruction is a widely adopted pretraining strategy. In NLP, BERT (Devlin et al., 2019) demonstrates the success of masked language modeling (MLM). Inspired by this, masked image modeling methods (MIM) like BEiT (Bao et al., 2021), MAE (He

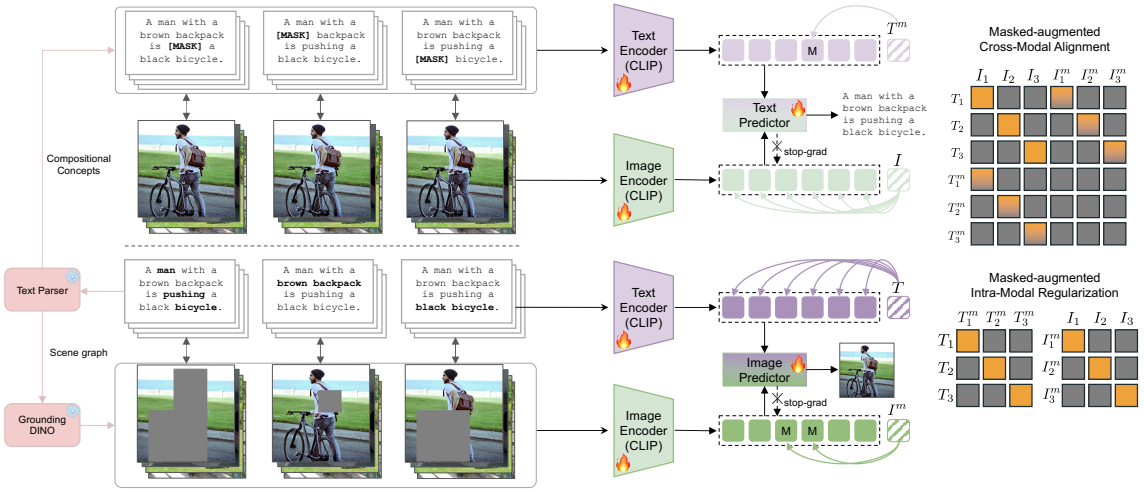

Figure 2: Our framework employs image and text predictors exclusively during training, removing them at inference time. The two image encoders share weights and function as a single encoder, as do the two text encoders.

et al., 2022), and SimMIM (Xie et al., 2022) train vision transformers to recover masked visual content. In VLMs, MaskVLM (Kwon et al., 2022) jointly reconstructs randomly masked image and text inputs, while Arici et al. (2021) explores MIM and MLM for structured catalog data to facilitate downstream vision tasks. These methods highlight the potential of masked modeling for cross-modal representations learning.

3 Method

3.1 Preliminaries of CLIP

CLIP consists of an image encoder E_I and a text encoder E_T, which project images and texts into a shared embedding space. The image encoder produces patch-level features and a global CLS token via full attention, while the text encoder generates token-level representations via causal attention, with the CLS token derived from the EOS token. Given a batch of paired samples $\mathcal{B} = (I_i, T_i)_{i=1}^{N}$, CLIP computes the similarity between global image and text embeddings using cosine similarity and is trained via a symmetric InfoNCE loss to align matching pairs and contrast mismatched ones. Detailed formulations are provided in Appendix A.

3.2 MACCO Framework

As illustrated in Figure 2, our framework enhances compositional understanding by masking compositional concepts in one modality and reconstructing them using the full features of the other modality as context. This design better exploits the aligned compositional signals inherent in paired image-text data. Specifically, masked texts are reconstructed using complete image features, and vice versa for masked images.

Prior studies (Kamath et al., 2023a; Dumpala et al., 2024) indicate that contrastive VLMs often struggle with compositional semantics due to limitations in the text encoder, particularly in capturing object relations and attribute bindings. Motivated by this, our framework emphasizes improving the text encoder's capacity to understand and represent compositional concepts. During reconstruction, we *stop the gradients* of the image features in both predictors, ensuring the loss focuses on optimizing text representations. We analyze this design choice and its impact in Section 4.5 and Appendix F.

Compositional Concept Extraction. To identify compositional concepts in both text and image modalities, we extract relation and attribute phrases from training examples. For textual inputs, we apply a scene graph parser (Wu et al., 2019) to obtain a mask M^T indicating the token positions of compositional phrases. For visual inputs, we use GroundingDINO (Liu et al., 2024b) to localize image regions corresponding to compositional phrases and map them to CLIP patch indices, producing a mask M^I over visual tokens. The full extraction pipeline and alignment details are provided in Appendix B and Algorithm 1.

Feature Extraction Formulation. Given a batch $\mathcal{B} = (I_i, T_i)_{i=1}^{N}$ of image-text pairs, we first obtain token-level compositional concept masks $\mathcal{M}^T \in \text{bool}^{N\times L}$ and $\mathcal{M}^I \in \text{bool}^{N\times(P+1)}$ for the text and image modalities, respectively, where L is the maximum text length and P is the number of image patches. A value of `True` indicates that the token is masked. For each modality, we initialize a shared learnable mask token (m_t for text and m_i for image). We then replace the tokens at masked

positions with the corresponding mask token and add positional embeddings PE^T and PE^I:

$$X^T = \text{Embed}(T) + PE^T, X_m^T = \text{Mask}(\text{Embed}(T)) + PE^T, \tag{1}$$

$$X^I = \text{Embed}(I) + PE^I, X_m^I = \text{Mask}(\text{Embed}(I)) + PE^I. \tag{2}$$

Then, we feed both the masked and unmasked sequences into the respective encoders. The masked and unmasked representations are then encoded as:

$$f^T = E_T(X^T), f_m^T = E_T(X_m^T), \tag{3}$$

$$f^I = E_I(X^I), f_m^I = E_I(X_m^I). \tag{4}$$

Here, f^T and f^I denote the full text features and full image features, while f_m^T and f_m^I correspond to masked variants.

Masked Textual Compositional Concept Modeling. To reconstruct masked text from cross-modal signals, we extract global features of the masked text ($f_m^{T|cls}$) and the full image ($f^{I|cls}$). Due to the causal nature of the text encoder, masked tokens lack sufficient future context. To mitigate this, we apply a simple ***global-to-local semantic injection*** operation, in which each masked token is enriched by integrating its representation with the global feature of the masked text, thereby enhancing contextual reasoning within the same modality.

For the image features, since CLIP's pretraining does not explicitly constrain local patch tokens, their alignment with text is weaker than that of the CLS token (Bica et al., 2024). Thus, we also inject global semantics into each image patch token to compensate for CLIP's weak local supervision and strengthen grounding during reconstruction. For simplicity, we formalize the *global-to-local semantic injection* operation as follows:

$$\bar{f_m^T} = \frac{1}{2}(f_m^T + f_m^{T|cls}), \bar{f}^I = \frac{1}{2}(f^I + f^{I|cls}). \tag{5}$$

The text predictor D^T uses two layers of cross-attention, attending from contextual masked text tokens to full image features, followed by a classification head, which is used to predict in the vocabulary space, following BERT (Devlin et al., 2019). The final loss for the masked modeling of texual compositional concepts is formulated as:

$$L_{MLM} = \mathbf{E}_{(T,I)\sim D}\mathcal{H}[D^T(\bar{f_m^T}, \texttt{stopgrad}(\bar{f}^I)), T], \tag{6}$$

where $\mathcal{H}$ denotes cross entropy loss. We compute the loss only on masked token.

Masked Visual Compositional Concept Modeling. For cross-modal image reconstruction, we also apply ***global-to-local semantic injection*** operation to enrich local tokens with contextual global semantics. Specifically, each local text token is fused with the global text feature $f^{T|cls}$ to obtain $\bar{f}^T$, and the global masked image feature $f_m^{I|cls}$ is similarly injected into local patch tokens to obtain $\bar{f_m^I}$:

$$\bar{f_m^I} = \frac{1}{2}(f_m^I + f_m^{I|cls}), \quad \bar{f}^T = \frac{1}{2}(f^T + f^{T|cls}). \tag{7}$$

The image predictor D^I employs masked image tokens as queries and full text features as keys/values in a three-layer cross-attention module. Following MAE (He et al., 2022), we use a decode embedding layer and 2D positional embedding prior to attention. The final prediction head reconstructs pixel values for each patch. The loss is the mean squared error (MSE) between reconstructed and original pixels:

$$L_{MIM} = \mathbf{E}_{(T,I)\sim D} \|[D^I(\texttt{stopgrad}(\bar{f_m^I}), \bar{f}^T), I\|^2, \tag{8}$$

we compute the loss only on the masked patches.

Masked-augmented Cross-Modal Alignment. We extend the standard contrastive learning framework in CLIP by incorporating the CLS token features of masked text or image inputs into the contrastive objective. Compared to their complete counterparts, masked inputs lack certain compositional concepts. For example, a masked text may retain only object-level information and thus can serve as a *soft negative sample* in image-to-text contrastive learning. Despite missing some details, the CLS token of masked input is still encouraged to encodes meaningful global semantics, as it facilitates reconstruction through *global-to-local semantic injection*. Thus, to better constrain the semantics, we introduce masked-augmented cross-modal alignment losses by computing contrastive losses between masked text and full images, and vice versa. During alignment, masked inputs are treated as *soft negatives* in the corresponding contrastive objectives. The masked-augmented image-to-text contrastive loss is formulated as follows:

$$\begin{aligned} L_{i2t}^{MCA} = &\sum_{i=1}^{N} \log \frac{\exp^{S(I_i,T_i)}}{\sum_{j=1}^{N}\exp^{S(I_i,T_j)} + \sum_{j=1}^{N}\exp^{S(I_i,T_j^m)}} \\ &+ \sum_{i=1}^{N} \log \frac{\exp^{S(I_i^m,T_i)}}{\sum_{j=1}^{N}\exp^{S(I_i^m,T_j)} + \sum_{j=1}^{N}\exp^{S(I_i^m,T_j^m)}}. \end{aligned} \tag{9}$$

Similarly, the masked-augmented text-to-image contrastive loss is formulated as:

$$\begin{aligned} L_{t2i}^{MCA} = &\sum_{i=1}^{N} \log \frac{\exp^{S(T_i,I_i)}}{\sum_{j=1}^{N}\exp^{S(T_i,I_j)} + \sum_{j=1}^{N}\exp^{S(T_i,I_j^m)}} \\ &+ \sum_{i=1}^{N} \log \frac{\exp^{S(T_i^m,I_i)})}{\sum_{j=1}^{N}\exp^{S(T_i^m,I_j)} + \sum_{j=1}^{N}\exp^{S(T_i^m,I_j^m)}}. \end{aligned} \tag{10}$$

Finally, the total masked-augmented cross modal contrastive loss is the sum of both:

$$L_{MCA} = -\frac{1}{2}(L_{i2t}^{MCA} + L_{t2i}^{MCA}). \quad (11)$$

Masked-augmented Intra-Modal Regularization. To prevent the masked text features or masked image features of different samples from collapsing into the same subspace and to constraint the deviation of the masked features from their corresponding full text or image features, contrastive loss is is well-suited for this regularization purpose. Additionally, contrasting single modality when performing cross-modal alignment is helpful for stable training (Zhang et al., 2024). Therefore, we introduce a new intra-modal regularization loss. Specifically, we apply intra-modal contrastive learning between the masked text features and the full text features, as well as between the masked image features and the full image features. The masked text to original text contrastive loss is formulated as follows:

$$L_{t2t}^{MIR} = \sum_{i=1}^{N} \left[\log \frac{\exp^{S(T_i^m, T_i)}}{\sum_{j=1}^{N} \exp^{S(T_i^m, T_j)}} + \log \frac{\exp^{S(T_i, T_i^m)}}{\sum_{j=1}^{N} \exp^{S(T_i, T_j^m)}} \right]. \quad (12)$$

Similarly, the masked image to original image contrastive loss is formulated as:

$$L_{i2i}^{MIR} = \sum_{i=1}^{N} \left[\log \frac{\exp^{S(I_i^m, I_i)}}{\sum_{j=1}^{N} \exp^{S(I_i^m, I_j)}} + \log \frac{\exp^{S(I_i, I_i^m)}}{\sum_{j=1}^{N} \exp^{S(I_i, I_j^m)}} \right]. \quad (13)$$

Finally, the total masked-augmented intra-modal contrastive loss is the sum of both:

$$L_{MIR} = -\frac{1}{2}(L_{t2t}^{MIR} + L_{i2i}^{MIR}). \quad (14)$$

Overall Training Objective. Our MACCO incorporates two masked modeling losses, L_{MLM} and L_{MIM}, as well as two masked-augmented auxiliary losses, L_{MCA} and L_{MIR}. The final loss is formulated as follows:

$$L_{total} = L_{MCA} + \lambda_1 L_{MIR} + \lambda_2 L_{MLM} + \lambda_3 L_{MIM}, \quad (15)$$

where λ_1, λ_2, and λ_3 are the weighting factors for the respective losses.

4 Experiments

Training Setup. Following (Li et al., 2024) and (Basu et al., 2024), we use approximately 110k high-quality image text pairs from MSCOCO (Lin et al., 2014) as the training set, and include additional experiments on CC3M in Appendix K. In the main experiments, we use the widely adopted OpenAI CLIP ViT/B-32 model, and provide supplementary results with ViT/B-16 and ViT/L-14 in Appendix I. We initialize both the text encoder and the image encoder with pretrained CLIP weights. The image and text predictors are trained from scratch. Following previous approaches (Basu et al., 2024; Li et al., 2024), we fine-tune the model for 5 epochs with a batch size of 256 and conduct a 50 steps warm up. The learning rate for the CLIP model is set to 5e-7, while the learning rate for the two predictors is set to 1e-3. We use AdamW as the optimizer with a weight decay of 0.2. Experiments are conducted on a single NVIDIA A100 GPU.

Evaluation Setup. At inference time, both predictors are removed, and the model's architecture remain the same as the pre-trained CLIP model. We perform a comprehensive evaluation using five widely used benchmarks for vision-language compositional understanding: ARO (Yuksekgonul et al., 2023), SugarCrepe (Hsieh et al., 2024), VL-Checklist (Zhao et al., 2022), VALSE (Parcalabescu et al., 2021) and What's-up (Kamath et al., 2023b). Detailed information about these datasets can be found in Appendix C.1.

For a fair comparison and comprehensive evaluation, we mainly selected three types of baselines: (i) the pre-trained CLIP model; (ii) the CLIP model fine-tuned on MSCOCO using only the contrastive loss (denoted as CLIP-FT); and (iii) CLIP-CAE (Li et al., 2024) and SDS-CLIP (Basu et al., 2024), which enhance the compositionality of CLIP-like model through fine-tuning on MSCOCO.

4.1 Main Results

As shown in Table 1, MACCO-CLIP achieves state-of-the-art performance across five widely-used benchmarks, significantly outperforming both the pretrained CLIP model and several fine-tuned variants, including CLIP-FT and CLIP-CAE. These results demonstrate MACCO-CLIP's strong advantages in relation understanding, attribute binding, and word order sensitivity.

Compared to CLIP, our model yields notable improvements, including 14.4% on ARO-Relation, 5.8% on ARO-Attribute, 21.9% on ARO-Order, and 8.3% average gain on Sugar-Crepe. Against CLIP-FT, we observe gains of 8.8% (ARO-Relation), 6.0% (Sugar-Crepe Relation), and 9.3% (VL-Checklist Relation). Notably, MACCO-CLIP achieves a 26.9% improvement on ARO-Order over CLIP-FT, significantly mitigating the well-documented insensitivity of CLIP to word order.

Model	ARO			Sugar-Crepe		VL-Checklist		VALSE	What's-up
	Relation	Attribute	Order	Relation	Attribute	Relation	Attribute	Relation	Relation
Random Chance	50.0	50.0	20.0	50.0	50.0	50.0	50.0	50.0	41.7
CLIP[1] (ViT-B/32)	58.7	62.7	<u>54.1</u>	68.8	70.8	63.6	67.7	<u>70.1</u>	41.8
CLIP-FT	64.3	<u>66.2</u>	49.1	71.1	77.7	60.9	67.4	69.3	41.4
IL-CLIP[2]	50.0	55.3	16.7	56.3	63.9	55.7	59.5	55.7	<u>42.3</u>
SDS-CLIP[3]	53.0	62.0	29.0	-	-	-	-	-	-
CLIP-CAE[4]	<u>69.5</u>	65.4	-	<u>73.0</u>	<u>78.9</u>	<u>65.4</u>	<u>68.6</u>	68.8	-
MACCO-CLIP (ours)	**73.1**	**68.5**	**76.0**	**77.1**	**79.1**	**70.2**	**68.7**	**75.3**	**43.2**

References: [1](Radford et al., 2021) [2](Zheng et al., 2024) [3](Basu et al., 2024) [4](Li et al., 2024)

Table 1: **Results on ARO, SugarCrepe, VL-Checklist, VALSE, and What's-up.** The best results are marked in **bold**, and the second-best results are <u>underlined</u>. Empty entries denote that the model's code has not been released. The result reported for CLIP-CAE are the average performance across its four model instances. The detailed results can be found in Appendix J.

MACCO-CLIP also consistently outperforms CLIP-CAE across all benchmarks, for example, with gains of 4.1% on Sugar-Crepe Relation and 4.8% on VL-Checklist Relation. While CLIP-CAE underperforms CLIP-FT on ARO-Attribute, our model improves over CLIP-FT by 2.3%. These gains may stem from a key design difference: although CLIP-CAE encourages models to focus on compositional concepts, it lacks an explicit mechanism for modeling dependencies between entities and their corresponding relations or attributes. In contrast, MACCO-CLIP incorporates a cross-modal masked modeling objective that explicitly encourage the model to capture such dependencies, resulting in semantically richer and more syntactically coherent representations. We also valid this effect in Section 4.4.

We note that both our MACCO-CLIP and prior work CLIP-CAE achieve smaller gains on attribute-focused tasks compared to relation-based tasks. This aligns with findings from prior work (Huang et al., 2023; Lewis et al., 2024), which suggests that attribute binding remains a more challenging aspect of compositional understanding and merits more investigation. Further discussion of this challenge can be found in Appendix S.

Finally, we also conduct additional experiments on three models with different scales and training paradigms, namely ViT-B/16, ViT-L/14, and SigLIP, and observe consistent and significant improvements across all of them. Detailed results are presented in Table 16 in Appendix I. Overall, these results highlight the effectiveness of our method.

4.2 Combined with Hard-Negative Samples

Since our framework is orthogonal to hard negative mining approaches, we further investigate whether it can be effectively integrated with hard negative samples. We consider two representative methods based on hard-negative samples: NegCLIP (Yuksekgonul et al., 2023), which incorporates hard negatives within a standard contrastive learning framework, and CE-CLIP (Zhang et al., 2024), which introduces two additional contrastive loss terms to better leverage the hard negatives.

For fair comparison, we use the same hard negative samples provided by NegCLIP across all experiments. As shown in Table 2, without bells and whistles, both models consistently benefit from the integration with MACCO, with the most notable improvements observed on VL-Checklist. For instance, MACCO enhances the performance of NegCLIP by 3.1% and CE-CLIP by 1.9% on the VL-Checklist Relation. These results further highlight the effectiveness of our method and its plug-and-play compatibility with existing hard negative mining methods. In Appendix N, we provide a more detailed discussion on our approach and hard-negative mining methods.

4.3 Downstream Tasks

In Table 3, we present the zero-shot classification accuracy and linear probing results on 11 widely used classification benchmarks. Details of the linear probing protocol settings are provided in the Appendix C.2. The results show that MACCO-CLIP incurs only a slight reduction in zero-shot and linear probe performance compared to the original CLIP model (a decrease of just 1.5% and 0.4% respectively). These results indicate that our method significantly improves compositional understanding while largely preserve the representation capacity of the original CLIP model. Nevertheless, simultaneously enhancing general representation while improving compositional understanding remains an open question and warrants further investigation.

Model	ARO			Sugar-Crepe		VL-Checklist		VALSE	What's-up
	Relation	Attribute	Order	Relation	Attribute	Relation	Attribute	Relation	Relation
Random Chance	50.0	50.0	20.0	50.0	50.0	50.0	50.0	50.0	41.7
CLIP[1] (ViT-B/32)	58.7	62.7	54.1	68.8	70.8	63.6	67.7	70.1	41.8
NegCLIP[2]	80.4	71.7	91.7	73.2	80.0	71.8	70.1	79.5	42.1
+ MACCO	80.2	71.6	**92.1**	**74.9**	79.7	**74.9**	**70.6**	78.7	**43.1**
CE-CLIP[3]	82.2	72.9	95.1	72.6	80.1	75.6	69.4	78.5	44.4
+ MACCO	**82.6**	**73.0**	**96.4**	**72.7**	80.0	**77.5**	**70.5**	**79.2**	**44.7**

References: [1](Radford et al., 2021) [2](Yuksekgonul et al., 2023) [3](Zhang et al., 2024)

Table 2: **Results on ARO, Sugar-Crepe, VL-Checklist, VALSE and What's-up when combined with hard negative samples**. Highlighted in **bold** denote an improvement over NegCLIP or CE-CLIP, while the underlined ones indicate a performance degradation compared to NegCLIP or CE-CLIP.

Model	Zero-Shot Avg.	Linear Probe Avg.	Comp. Avg.
CLIP	**59.5**	**80.1**	61.2
CLIP-FT	57.9	80.0	61.8
MACCO-CLIP (ours)	58.0	79.7	**67.7**

Table 3: **Zero-shot classification performance and linear probe results on 11 datasets.** The results in last column represent the average performance across five compositional understanding benchmarks.

Model	SICK-R		STS-Benchmark	
	Spearman	Pearson	Spearman	Pearson
CLIP	67.9	68.6	61.5	59.1
CLIP-FT	68.0	73.4	66.3	64.0
CLIP-CAE	69.3	71.6	**66.5**	**65.2**
MACCO-CLIP (ours)	**70.5**	**76.4**	65.3	64.9

Table 4: **Semantic textual similarity results on SICK-R and STS-Benchmark.**

4.4 Analysis

Semantic Textual Similarity. Following prior work CLIP-CAE (Li et al., 2024), we evaluate the text encoders of different models on two widely used STS benchmarks: STS-Benchmark (Cer et al., 2017) and SICK-R (Marelli et al., 2014). Details about the task can be found in Appendix C.3.

As shown in Table 4, our method achieves a notable improvement on SICK-R over CLIP-FT, and outperforms CLIP-CAE with a 4.8% gain in Pearson correlation. While slightly underperforming on STS-Benchmark, we attribute this to its domain heterogeneity and reliance on shallow lexical cues, where CLIP-CAE's keyword-focused optimization provides a slight advantage. In contrast, SICK-R demands deeper compositional reasoning and sensitivity to lexical-syntactic structure. These findings highlight that MACCO enhances the text encoder's ability to capture nuanced semantic and compositional relations, beyond surface similarity.

Model	Depth	TopConstituents	BigramShift	Tense	Avg.
CLIP	25.0	50.5	64.8	82.4	55.7
CLIP-FT	25.5	49.1	64.5	82.3	55.4
MACCO-CLIP (ours)	**26.2**	**50.7**	**65.7**	**84.0**	**56.6**

Table 5: **Probing results of linguistic information in text embedding.**

Linguistic Information Probing. The experiments on the STS task demonstrate that our model is more effective at capturing compositional semantic information within sentences. To further examine the linguistic properties encoded in the text embedding produced by different models, we perform a probing analysis using the SentEval toolkit (Conneau and Kiela, 2018) on the text encoders of different models. We use four representative tasks that evaluate the extent to which sentence embeddings encode latent structural and semantic information. As shown in Table 5, CLIP-FT shows performance degradation on three tasks relative to the original CLIP, whereas MACCO-CLIP consistently yields substantial accuracy gains across all tasks. These results further confirm that our method not only improves the model's ability to encode compositional concepts but also enhances the text encoder's capacity to capture syntactic structure and linguistic information.

Text Embedding Ingredients. Inspired by (Li et al., 2024), we follow the procedure outlined in their work to compute the similarity between the embedding of the full caption and that of the corresponding relation or attribute phrase in the ARO benchmarks. Figure 3 presents the similarity distributions for CLIP, CLIP-FT, and our MACCO-CLIP. As shown, the text encoder of MACCO-CLIP produces embeddings that exhibit significantly higher similarity to their corresponding compositional concept embeddings, compared to those generated by

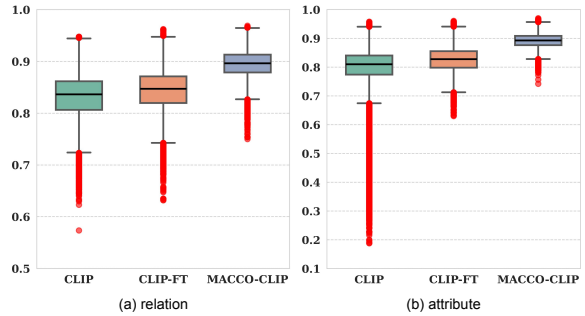

Figure 3: The similarity distribution between the embeddings of full captions and those of relation or attribute phrases extracted from the same text.

CLIP and CLIP-FT. This result further validates that our model more effectively captures compositional concepts in text, with its embeddings encapsulating richer semantic information.

Compositionality Robustness Evaluation. To assess robustness to semantically invariant perturbations, we evaluate MACCO-CLIP on the Hard Positive Compositional Benchmark (Kamath et al., 2024). The results are shown in Table 6. *Orig. Test Acc.* measures the ability to distinguish positives from meaning-altering hard negatives, while *Aug. Test Acc.* additionally requires recognizing semantically equivalent hard positives. For example, given an image I and captions "brown grass" (positive T), "blue grass" (hard negative T_n), and "chestnut grass" (hard positive T_p), the model must satisfy: $S(I,T) > S(I,T_n)$ and $S(I,T_p) > S(I,T_n)$.

The results show that MACCO-CLIP consistently outperforms CLIP-FT on both metrics, with notable gain on the SWAP subset. This subset presents a more challenging scenario that tests compositional understanding by constructing hard positives through the reordering of object-attribute phrases. On this subset, MACCO-CLIP achieves a 4.6% improvement over CLIP-FT, highlighting its superior capability in capturing object-attribute binding relationships. These results demonstrate that our method not only improves the model's sensitivity to meaning-altering perturbations but also enhances its robustness to semantically equivalent variations. This highlights our framework's strong robustness in visio-linguistics compositionality.

Robustness Analysis and Discussion About Detection Model. In Appendix M, we demonstrate that highly proficient off-the-shelf visual grounding model is not a strict requirement for our method and MACCO is resilient to noisy detection results. We further discuss the scenario generalizability and potential systemic biases of external pre-trained tools in Appendix P and Appendix Q. Due to space limitations, more detailed analysis and discussion can be found there.

Model	REPLACE		SWAP	
	Orig. Test Acc.	Aug. Test Acc.	Orig. Test Acc.	Aug. Test Acc.
CLIP	63.2	47.2	61.0	49.7
CLIP-FT	62.3	48.3	63.9	48.4
MACCO-CLIP (ours)	**68.3**	**48.5**	**66.1**	**53.8**

Table 6: **Results on Hard Positive Benchmark.**

4.5 Ablation

We conduct ablation studies to understand the effectiveness of each component in our framework (see Table 9 and Table 10 in Appendix D). We also provide a clearer ablation of mask strategies and auxiliary objectives in Table 12.

Cross-Modal Masked Modeling Losses. Experimental results show that incorporating reconstruction losses for both text and image improves model performance, regardless of whether auxiliary objectives are included. The best results are achieved when both losses are combined, highlighting the effectiveness of the masked modeling framework.

Auxiliary Losses. Even without masked modeling, introducing each auxiliary loss individually yields consistent performance gains, achieving the highest improvement when used together. When used with masked modeling, adding L_{MAC} brings a significant boost, and further incorporating L_{MIR} leads to the best performance. These results suggest that each auxiliary loss is beneficial on its own, and that compositional masked modeling synergizes with the auxiliary losses to enhance feature representation learning.

Global-to-local Semantic Injection. The ablation results in Table 10 and Table 11 shows that our method performs better with this strategy, confirming its effectiveness. We attribute this to its ability to enrich local tokens with global semantic context and to provide an additional constraint on the global representation.

Stop-Gradient and Masking Strategy. Blocking gradient flow from image features within the predictors yield better performance, due to a sharper focus on optimizing the text encoder. And our masking strategy targeting compositional concepts clearly outperforms random masking. We provide a more detailed discussion on masking strategies in Appendix L.

Model	BLIP-VQA ↑			Human-preference ↑		
	Color	Texture	Shape	Color	Texture	Shape
SD 1.5 (w/ vanilla CLIP text encoder)	0.3651	0.4135	0.3721	-0.4381	-0.4349	-0.3323
SD 1.5 (w/ MACCO-CLIP text encoder)	**0.3815**	**0.4236**	**0.3835**	**-0.3295**	**-0.3840**	**-0.2793**

Table 7: Experimental results on compositional text-to-image generation tasks.

Model	AMBER					MME
	Attribute	State	Number	Action	Relation	Perception
LLaVA-1.5-7B (w/ vanilla CLIP vision encoder)	75.8	73.9	78.2	81.1	68.4	1447.1
LLaVA-1.5-7B (w/ MACCO-CLIP vision encoder)	**76.5**	**74.8**	**78.2**	**81.7**	**69.3**	**1452.3**

Table 8: Experimental results when applying MACCO's vision encoder to multimodal large language models.

4.6 Application

Compositional Text-to-Image Generation. Text-to-image (T2I) diffusion models such as Stable Diffusion (Rombach et al., 2022) typically use pre-trained VLMs (*e.g.*, CLIP) as their text encoders. We therefore investigate whether MACCO, by improving the compositionality of VLMs, also enhances compositional generation in T2I models. As an extended application, we apply the text encoder trained under our MACCO framework to compositional text-to-image generation. Specifically, we replace the original text encoder (ViT-L/14) of Stable Diffusion v1.5 (SD 1.5) with the MACCO-CLIP (ViT-L/14) text encoder. We evaluate attribute binding on T2I-CompBench (Huang et al., 2023), which contains three subsets (color, texture, and shape), each with 300 text prompts. For each prompt, we generate 10 images with different random seeds and evaluate them using BLIP-VQA scores (Huang et al., 2023) and ImageReward (Xu et al., 2023) preference scores. As shown in Table 7, using the text encoder from MACCO-CLIP improves the attribute binding performance of the T2I model without additional fine-tuning. These results indicate that the stronger text representation backbone learned by MACCO also benefits text-to-image diffusion models and supports more accurate generation under compositional semantics.

Multimodal Large Language Models. Given that mainstream multimodal large language models (MLLMs) commonly adopt VLMs such as CLIP as visual backbone, we conduct transfer experiments based on the LLaVA-1.5-7B (Liu et al., 2024a) to further assess the potential of our method for improving MLLMs. We follow the two-stage training recipe of LLaVA (Liu et al., 2024a), and use LoRA for the instruction-tuning stage due to limited computational resources. We replace the visual encoder of LLaVA-1.5-7B with the vanilla CLIP ViT-L/14 and our MACCO-CLIP ViT-L/14 respectively, and perform the same two-stage training with identical data and hyperparameters. We then evaluate compositional perception performance on AMBER (Wang et al., 2023), a benchmark for multimodal hallucination, and on MME (Fu et al., 2025), a general multimodal benchmark. As shown in Table 8, using the visual encoder of MACCO-CLIP improves performance over the baseline across multiple AMBER dimensions, including attributes, states, actions, and relations, and also improves MME perception scores, indicating that the compositional gains from MACCO can also transfer to MLLMs. These results suggest that MACCO strengthens compositional semantic modeling through cross-modal compositional concept masked modeling, thereby enhancing the visual encoder's ability to capture fine-grained visual cues. In contrast to the compositional perception deficiencies of the original CLIP visual encoder (Yuksekgonul et al., 2023), MACCO produces visual representations with richer structural information.

5 Conclusion

Our work introduced MACCO, a framework that improves compositional understanding in VLMs like CLIP. By masking compositional concepts in one modality and reconstructing them from the other, MACCO better exploits the aligned compositional signals in paired image-text data. We further proposed two auxiliary objectives, MCA and MIR, to enhance cross-modal alignment and intra-modal regularization. Extensive experiments and in-depth analyses show that MACCO effectively improves compositional reasoning, enhances the model's encoding of syntactic structure and semantic nuance, and benefits other multimodal tasks.

6 Limitations

While MACCO introduces a novel and effective framework for enhancing the compositional understanding of vision-language models without relying on explicit hard negative construction, several limitations remain, each pointing to promising avenues for future research. **First**, although MACCO leverages naturally aligned image-text pairs for masked cross-modal reconstruction, it requires lightweight pre-processing to extract compositional concepts (*e.g.*, phrases or regions) from both modalities. This step, while minimal and compatible with standard tools, introduces a dependency that may limit flexibility in fully end-to-end pipelines. **Second**, MACCO adds two predictors and prediction heads during training, increasing the number of training-time parameters. These components are discarded at inference, but the approach still incurs greater training overhead than methods such as CLIP-CAE (Li et al., 2024). Improving the efficiency and transparency of MACCO's learned representations remains an important goal. **Third**, the current design targets contrastive vision-language models like CLIP. Its applicability to generative architectures such as BLIP (Li et al., 2022) has yet to be explored. Extending MACCO to generative objectives, especially those based on language modeling or captioning, is a natural and valuable direction for future work. **Lastly**, while MACCO enhances compositional robustness and alignment, it does not yet offer the interpretability of concept bottleneck models or attribution enhancement frameworks. Incorporating mechanisms such as concept probing or attributing tracing could yield deeper insights into model behavior. Despite these limitations, MACCO contributes a promising training paradigm for compositional reasoning in VLMs. In addition, MACCO leaves room for further improvement. For example, incorporating multi-granularity alignment as in X-VLM (Zeng et al., 2022) or adopting a stronger text encoder may further enhance MACCO. We provide detailed discussions in Appendix R and Appendix T.

7 Acknowledgements

This work was supported by the Natural Science Foundation of China under Grant 62571507. We sincerely thank the meta-reviewer and the anonymous reviewers for their constructive and valuable feedback.

References

Tarik Arici, Mehmet Saygin Seyfioglu, Tal Neiman, Yi Xu, Son Train, Trishul Chilimbi, Belinda Zeng, and Ismail Tutar. 2021. Mlim: Vision-and-language model pre-training with masked language and image modeling. *arXiv preprint arXiv:2109.12178*.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR.

Samyadeep Basu, Shell Xu Hu, Maziar Sanjabi, Daniela Massiceti, and Soheil Feizi. 2024. Distilling knowledge from text-to-image generative models improves visio-linguistic reasoning in clip. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6105–6113.

Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrovic. 2024. Improving fine-grained understanding in image-text pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 3974–3995. PMLR.

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, and Leonid Karlinsky. 2023. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20155–20165.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023. Altclip: Altering the language encoder in clip for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8666–8682.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*.

Dinh-Truong Do, Minh-Phuong Nguyen, and Le-Minh Nguyen. 2025. Enhancing zero-shot multilingual semantic parsing: A framework leveraging large language models for data augmentation and advanced prompting techniques. *Neurocomputing*, 618:129108.

Truong Dinh Do, Phuong Minh Nguyen, and Minh Nguyen. 2024. Zela: Advancing zero-shot multilingual semantic parsing with large language models and chain-of-thought strategies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17783–17794.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. 2023a. Dense and aligned captions (dac) promote compositional reasoning in vl models. In *Advances in Neural Information Processing Systems*, volume 36, pages 76137–76150.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023b. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

Sri Harsha Dumpala, David Arps, Sageev Oore, Laura Kallmeyer, and Hassan Sajjad. 2024. Seeing syntax: Uncovering syntactic learning limitations in vision-language models. *arXiv preprint arXiv:2412.08111*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023a. Text encoders bottleneck compositionality in contrastive vision-language models. *arXiv preprint arXiv:2305.14897*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023b. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. 2024. The hard positive truth about vision-language compositionality. In *European Conference on Computer Vision*, pages 37–54. Springer.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Jina clip: Your clip model is also your text retriever. *arXiv preprint arXiv:2405.20204*.

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. 2024. Does clip bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500.

Haoxin Li and Boyang Li. 2025. Enhancing vision-language compositional understanding with multimodal synthetic data. *arXiv preprint arXiv:2503.01167*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Wei Li, Zhen Huang, Xinmei Tian, Le Lu, Houqiang Li, Xu Shen, and Jieping Ye. 2024. Interpretable composition attribution enhancement for visio-linguistic compositional understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14616–14632.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007.

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. 2024. Preserving multimodal capabilities of pre-trained vlms for improving vision-linguistic compositionality. *arXiv preprint arXiv:2410.05210*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. In *Advances in Neural Information Processing Systems*, volume 37, pages 32731–32760.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. 2023. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13774–13784.

Le Zhang, Qian Yang, and Aishwarya Agrawal. 2025. Assessing and learning alignment of unimodal vision and language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14604–14614.

Xinyu Zhang, Jiahui Chen, Junkun Yuan, Qiang Chen, Jian Wang, Xiaodi Wang, Shumin Han, Xiaokang Chen, Jimin Pi, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Cae v2: Context autoencoder with clip target. *arXiv preprint arXiv:2211.09799*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13785–13795.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Preliminaries of CLIP

CLIP typically consists of two independent encoders: an image encoder E_I and a text encoder E_T. Consider a mini-batch $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^{N}$ of size N, consisting of image and text pairs (I_i, T_i). The image encoder first divides each image I_i into several image patches, which are embedded into a token sequence, and then positional encoding is added before feeding them into a transformer model. The output is a series of image tokens $V = \{v^{cls}, v^1, \ldots, v^P\} \in \mathbb{R}^{(P+1)\times d}$, where v^{cls} denotes the CLS token that encapsulates global information, v^i represents the patch embedding containing the local information of the image, and d denotes the feature dimension, while P denotes the number of patches. The image encoder employs full attention, meaning all patches and CLS token can attend to each other.

Similarly, the text encoder E_T tokenizes each text T_i, then pads it with padding token, adds positional embedding, and feeds it into the transformer model. The output is a series of text tokens $T = \{t^0, t^1, \ldots, t^{cls}, \ldots, t^L\} \in \mathbb{R}^{L\times d}$. t^{cls} is the text-side CLS token, initialized with the EOS token, which encapsulates the global information of the text. The text encoder uses causal attention, meaning that each token can only attend to itself and the previous tokens.

The image-text similarity is measured using the similarity of their global representations:

$$S(I_i, T_j) = \frac{v_i^{\text{cls}} \cdot t_j^{\text{cls}}}{\|v_i^{\text{cls}}\| \cdot \|t_j^{\text{cls}}\|} / \tau, \tag{16}$$

where τ is the temperature parameter.

CLIP maximizes the similarity between each matching image-text pair using an InfoNCE loss while minimizing the similarities with other non-matching image-text pairs. The Image-Text Contrastive (ITC) loss is formulated as follows:

$$\mathcal{L}_{ITC} = -\sum_{i=1}^{N} \left[\log \frac{\exp^{S(I_i,T_i)}}{\sum_{j=1}^{N} \exp^{S(I_i,T_j)}} + \log \frac{\exp^{S(T_i,I_i)}}{\sum_{j=1}^{N} \exp^{S(T_i,I_j)}} \right]. \tag{17}$$

B Details of Compositional Concept Extraction

Textual Compositional Concepts Extraction. For each text in the training set, we utilize a widely adopted text scene graph parser (Wu et al., 2019) to extract compositional concepts. This parser converts each text into a scene graph by identifying object-relation phrases and object-attribute phrases. It also provides the exact words in the text corresponding to each relation or attribute. For example, the sentence "A man with a brown backpack is pushing a black bicycle" will be parsed as: "man pushing bicycle", "brown backpack", "black bicycle" with compositional concept words being ["pushing", "brown", "black"]. In this way, we identify the compositional concepts (*i.e.*, relations and attributes) contained in each text. Finally, for each text, we can generate a binary text token mask M^T to indicate the positions of the compositional concepts within the text.

Visual Compositional Concepts Extraction. For each image in the training set, we leverage the scene graph annotations derived from its caption (*i.e.*, the object-relation and object-attribute phrases), along with a open-set object detector, GroundingDINO (Liu et al., 2024b). These open-world detection model does not require predefined categories and can detect regions in the image corresponding to the input textual description. Other open-world detection models are also applicable, and empirical results indicate that a reasonably capable detector can yield competitive performance, demonstrating robustness to the choice of detection model. In our main experiments, we use the base version of GroundingDINO and follow the official GitHub repository for inference setup.

Specifically, we input each relation or attribute phrase from the image caption into the detection model and obtain bounding box coordinates for the matching region. These coordinates are based on the unnormalized coordinates of the original image. We then apply a simple coordinate mapping algorithm to map these coordinates into the image coordinate space used by the CLIP model (since the images input into CLIP often undergo random cropping and resizing as part of data augmentation, we track the parameters of the image data augmentation process to facilitate the coordinate mapping). Next, we further map the region corresponding to these coordinates to specific image patches. Finally, we obtain the positions of the patches in the image corresponding to each relation or attribute phrase. For each relation or attribute phrase, we can generate a binary image token mask M^I to indicate the positions of the compositional concepts within the image corresponding to the phrase. The detailed extraction process is described in Algorithm 1.

C Details on Evaluation Benchmark

C.1 Compositionality Benchmark

To comprehensively assess the effectiveness of our method in improving compositional understanding, we conduct evaluations on five widely used compositional benchmarks, as well as a recently proposed benchmark that emphasizes robustness under semantically invariant perturbations. Below, we summarize the details of each dataset.

ARO (Yuksekgonul et al., 2023) systematically evaluates vision-language models on three core aspects of compositionality: relations, attributes, and word order. It comprises four major subsets: ARO-Relation, ARO-Attribute, and ARO-Order (which includes both COCO-Order and Flickr30k-Order). ARO-Relation spans 48 relation types and $23,937$ test samples, requiring models to accurately distinguish relational structures such as "a dog behind a tree" versus "a tree behind a dog". ARO-Attribute includes 117 attribute-object combinations across $28,748$ samples, challenging models to resolve attribute compositionality (*e.g.*, distinguishing between "a crouching cat and an open door" and "an open cat and a crouching door"). ARO-Order assesses sensitivity to word order by presenting four permuted versions of a caption, with the model tasked to identify the correct one. Performance is averaged over the COCO-Order and Flickr30k-Order subsets.

Sugar-Crepe (Hsieh et al., 2024) is a recently introduced benchmark focused on evaluating models with adversarially generated hard negatives. Lever-

Algorithm 1 Visual Compositional Concepts Extraction

Require: Training image I, corresponding scene graph phrases P, detector $\mathcal{G}$

Ensure: Image token masks of compositional concept $\{M^I(p)\}$ for $p \in P$

Initialize image preprocessor $\mathcal{C}$ (can track RandomResizedCrop parameter)
$\theta \leftarrow (i, j, h, w) \leftarrow \mathcal{C}(I_k)$ ▷ Record parameter
for each phrase $p \in P_k$ **do**
    $B_p \leftarrow \mathcal{G}(I_k, p)$ ▷ Raw detection
    for each object coordinate B_p^i in B_p **do**
        $\hat{B}_p^i \leftarrow \text{CoordinateMapper}(B_p^i, \theta)$ ▷ Map coordinates into space after apply image preprocessor
        $M^i(p) \leftarrow \text{GenerateMask}(\hat{B}_p^i)$
    end for
    Merge image token masks belonging to the sample phrase p to $M^I(p)$
end for

function CoordinateMapper(B_p, θ)
    Parse $\theta = (i, j, h, w)$
    Compute scaling factors $(s_w, s_h) \leftarrow (224/w, 224/h)$
    Transform coordinates:

$$\hat{B}_p = \begin{bmatrix} \max(B_p^{x_1} - j, 0) \cdot s_w \\ \max(B_p^{y_1} - i, 0) \cdot s_h \\ \min(B_p^{x_2} - j, w) \cdot s_w \\ \min(B_p^{y_2} - i, h) \cdot s_h \end{bmatrix}$$

    return $\hat{B}_p$
end function

function GenerateMask($\hat{B}_p$)
    return Image patch indices covered by $\hat{B}_p$
end function

aging large language models, it produces fluent and semantically plausible negative captions through targeted insertions, replacements, or rephrasings. Following (Li et al., 2024), we report accuracy on the relation and attribute subsets of SugarCrepe separately.

VL-Checklist (Zhao et al., 2022) is a large-scale compositionality evaluation dataset composed of over $410,000$ samples sourced from VG, SWIG, VAW, and HAKE. It covers a wide array of subcategories including color, material, size, action, and spatial relations. Consistent with prior work, we report average results for the relation and attribute categories.

VALSE (Parcalabescu et al., 2021) serves as a task-agnostic benchmark aimed at assessing the foundational visual-linguistic competence of general-purpose pretrained VLMs. It comprises six linguistic phenomena: existence, plurality, counting, spatial relations, actions, and entity coreference. We evaluate our method on the three subsets most relevant to visio-linguistic compositionality, in accordance with (Li et al., 2024).

What's-Up (Kamath et al., 2023b) is a spatial reasoning benchmark specifically designed to test VLMs' understanding of object spatial relation. It consists of three datasets: What'sUp (820 manually curated images), constructed with controlled object layouts to mitigate spatial priors. COCO-spatial $(2,687$ images), derived from the COCO dataset, pairs each image with two mutually exclusive captions differing in spatial expressions. GQA-spatial $(1,451 images)$, adapted from the GQA validation set, contains spatial questions with unambiguous object references and prominent object sizes. We report the average accuracy across all three datasets.

Hard Positive Benchmark (Kamath et al., 2024) is introduced to measure model robustness under semantic-preserving compositional perturbations. This benchmark comprises $56,191$ images, including $28,748$ swap-based and $27,443$ replacement-based hard positives.

C.2 Downstream Classification Benchmark

Due to computational constraints, we evaluate the model's performance under both zero-shot and linear probing settings across 11 widely used image classification datasets: CIFAR-10, CIFAR-100, Caltech-101, MNIST, VOC-2007, Aircraft, Hateful Memes, Rendered SST2, FER-2013, RESISC45, EuroSAT and FGVC-Aircraft. For linear probing, we adopt a full-shot training setup, training each model for 50 epochs using SGD optimizer with a learning rate of 0.1 and a weight decay of $1e{-}6$.

C.3 Semantic Textual Similarity Benchmark

STS-Benchmark (Cer et al., 2017) is a standard dataset for semantic similarity assessment, comprising sentence pairs drawn from diverse domains such as news headlines, image captions, and QA forums. Each pair is annotated with a continuous similarity score ranging from 0 to 5. In contrast, SICK-R (Marelli et al., 2014) is designed to assess

compositional semantics by systematically generating sentence pairs that reflect fine-grained semantic differences induced by lexical and syntactic variations. It places a greater emphasis on a model's ability to understand structured and compositional meaning.

D Detailed Ablation Study of Key Designs

In Table 9 and Table 10, we present ablation studies analyzing key components of our framework. Due to space limitations, detailed results are provided in here. The main text discusses the impact of cross-modal masked modeling losses, auxiliary objectives, global-to-local semantic injection, stop-gradient strategy, and the masking scheme

Our use of the global-to-local semantic injection strategy serves two purposes: first, to provide the key and value tokens with more contextual global information; and second, to make the reconstruction learning more effective in constraining the global representation. This is particularly important because our contrastive learning objective (whether CLIP or SigLIP) mainly supervises the global representation, and most downstream tasks also rely on global representations. Therefore, even though SigLIP adopts bidirectional text attention, our strategy should still offer benefits. To further validate this assumption, we conduct additional experiments on SigLIP ViT-B/16, with results shown in Table 11. The results further indicate that incorporating our global-to-local semantic injection strategy improves performance in SigLIP models, although the gain is smaller than in CLIP-based models. This suggests that our strategy remains beneficial even when applied to architectures like SigLIP that use bidirectional text attention.

E Further Simplified Ablation of Auxiliary Losses and Targeted Masking Strategy

To more clearly isolate the contributions of the auxiliary losses and the targeted masking strategy in our method, we summarize the key ablation results in Table 12. In light of these results, we draw the following two conclusions:

(1) Compositional masking outperforms random masking. A comparison between "Random Masking" (+1.4) and "Compositional Concept Masking" (+2.1) without auxiliary losses demonstrates that specifically targeting compositional concepts is more effective than random masking. Furthermore, contrasting "Random + Aux" (71.2) with MACCO (73.4) highlights that while both benefit from enhanced feature representation learning, MACCO achieves an additional performance gain of $+2.2\%$. This underscores the effectiveness of our masking strategy, as compositional concepts serve as the "structural glue" of a scene, masking these concepts forces the model to engage in higher-order vision-language reasoning and moving beyond simple token-level reconstruction.

(2) Compositional masking and enhanced feature representation learning work synergistically. The performance improvement achieved by MACCO (+7.3) significantly exceeds the additive contributions of "Auxiliary Losses" (+2.6) and "Compositional Masking" (+2.1) individually. This substantial synergy indicates that our masked-augmented losses (L_{MCA}&L_{MIR}) play a crucial role in effectively regularizing the feature space and facilitating the masked modeling process. This finding underscores the indispensable interplay between compositional masking and improved feature representation learning, as both components mutually reinforce each other to achieve notable performance gains.

F Freeze or Fire Image Encoder?

We conduct additional experiments with the vision encoder frozen, as shown in Table 13. Overall, the results indicate that finetuning the vision encoder leads to better performance, although freezing it yields marginal advantages on a few benchmarks.

This observation is reasonable regarding the findings from prior work (Zhai et al., 2022; Sung et al., 2022; Li et al., 2022), which suggest that jointly optimizing both modalities enhances the flexibility of the shared embedding space and enables more effective cross-modal alignment. Specifically, Zhai et al. (2022) suggests that while freezing the vision encoder may improve training efficiency and mitigate overfitting, it is generally more effective when the encoder is already strong, for instance, pretrained via self-supervised learning on large-scale image datasets such as JFT-300M. Sung et al. (2022) argues that the adaptability of the vision encoder's feature space is critical for downstream text-side tuning. If the vision encoder is fixed, its output space may be too rigid, limiting the text encoder's ability to capture cross-modal semantics. Li et al. (2022) shows that jointly optimizing both encoders leads to more stable and higher performance in both zero-shot and fine-tuning settings.

Model	$\mathcal{L}_{MLM}$	$\mathcal{L}_{MIM}$	$\mathcal{L}_{MAC}$	$\mathcal{L}_{MIR}$	ARO	Sugar-Crepe	VL-Checklist	Avg.
CLIP (ViT-B/32)	-	-	-	-	58.5	69.8	65.6	64.6
CLIP-FT	-	-	-	-	59.9	74.4	64.2	66.1
ablation of cross-modal masked modeling losses								
	✓	-	-	-	68.2	74.5	65.0	68.2 (+2.1)
	-	✓	-	-	65.3	75.4	64.7	68.5 (+2.3)
	-	-	✓	✓	65.0	75.1	66.2	68.7 (+2.6)
	✓	-	✓	✓	72.0	77.6	68.7	72.8 (+6.7)
	-	✓	✓	✓	69.9	74.9	66.8	70.5 (+4.4)
MACCO-CLIP	✓	✓	✓	✓	**72.5**	**78.1**	**69.5**	**73.4 (+7.3)**
ablation of two auxiliary losses								
	-	-	-	✓	58.2	75.4	64.0	65.9 (-0.2)
	-	-	✓	-	63.9	74.4	65.2	67.9 (+1.8)
	-	-	✓	✓	65.0	75.1	66.2	68.7 (+2.6)
	✓	✓	-	-	65.1	75.0	64.5	68.2 (+2.1)
	✓	✓	✓	-	71.5	77.9	68.3	72.6 (+6.5)
	✓	✓	-	✓	64.1	75.2	64.2	67.8 (+1.7)
MACCO-CLIP	✓	✓	✓	✓	**72.5**	**78.1**	**69.5**	**73.4 (+7.3)**

Table 9: **Ablation of different losses**. The numbers in parentheses indicate the performance gains relative to CLIP-FT.

Model	Global-to-local semantic injection	Stop-grad	Mask compositional concepts	ARO	Sugar-Crepe	VL-Checklist	Avg.
CLIP (ViT-B/32)	-	-	-	58.5	69.8	65.6	64.6
CLIP-FT	-	-	-	59.9	74.4	64.2	66.1
	-	✓	✓	72.2	76.3	67.6	72.0
	✓	-	✓	72.4	76.0	68.0	72.1
	✓	✓	-	71.1	75.4	67.0	71.2
MACCO-CLIP	✓	✓	✓	**72.5**	**78.1**	**69.5**	**73.4**

Table 10: **Ablation of global-to-local semantic injection operation, stop-gradient strategy and masking strategy**. We ablate the mask strategy with random mask, where random mask represents randomly masking image and text with a mask ratio of 75% and 15% following MAE (He et al., 2022) and BERT (Devlin et al., 2019).

Model	Global-to-local semantic injection	ARO	Sugar-Crepe	VL-Checklist	Avg.
SigLIP (ViT-B/16)	-	27.4	62.8	50.9	47.0
SigLIP-FT	-	49.9	79.0	65.3	64.7
MACCO-SigLIP	-	61.5	80.0	67.1	69.5
MACCO-SigLIP	✓	**62.5**	**80.2**	**67.5**	**70.1**

Table 11: **Ablation results of w and w/o global-to-local semantic injection on SigLIP**.

Method	L_{MCA} & L_{MIR} (Improved feature representation learning)	Masking Strategy	Avg. Compositional Performance
CLIP	×	None	64.6
CLIP-FT	×	None	66.1
CLIP + Auxiliary Losses	✓	None	68.7 (+2.6)
CLIP + Random Masking	×	Random	67.5 (+1.4)
CLIP + Random Masking + Auxiliary Losses	✓	Random	71.2 (+5.1)
CLIP + Compositional Concept Masking	×	Compositional	68.2 (+2.1)
MACCO-CLIP (ours)	✓	**Compositional**	**73.4 (+7.3)**

Table 12: **Simplified ablation of auxiliary losses and targeted masking strategy.**

Model	Fire Image Encoder	ARO	Sugar-Crepe	VL-Checklist	Avg.
MACCO-CLIP	-	**72.7**	75.2	67.3	71.7
MACCO-CLIP	✓	72.5	**78.1**	**69.5**	**73.4**

Table 13: **Ablation results of the choice not to freeze the image encoder.**

G Computational Budget

In Table 14, we present the model sizes along with the training and evaluation budgets for all models discussed in our paper. Compared to standard finetuning, our method does not introduce significant additional training cost, and the inference cost remains unchanged.

Model	#Params	Training Budge	Evaluation Budge
Backbone: CLIP ViT-B/32			
CLIP	151M	-	0.3h
CLIP-FT	151M	0.8h	0.3h
SDS-CLIP	151M	-	-
IL-CLIP	151M	-	0.3h
CLIP-CAE	151M	-	-
MACCO-CLIP	151M	1.0h	0.3h
Backbone: CLIP ViT-B/16			
CLIP	151M	-	0.4h
CLIP-FT	151M	2.5h	0.4h
MACCO-CLIP	151M	2.6h	0.4h
Backbone: CLIP ViT-L/14			
CLIP	427M	-	1.0h
CLIP-FT	427M	8.8h	2.2h
MACCO-CLIP	427M	9.0h	2.2h
Backbone: SigLIP ViT-B/16			
SigLIP	172M	-	1.5h
SigLIP-FT	172M	2.2h	1.5h
MACCO-SigLIP	172M	2.5h	1.5h

Table 14: **Model size and computational budge.**

H Error Bar

In Table 15, we report the mean and standard deviation of the model performance trained using four different random seeds.

I More Experiments on Other Model Scales

We conduct experiments on three models with different scales and training paradigms: ViT-B/16, ViT-L/14, and SigLIP ViT-B/16. The results are presented in Table 16. We compare our method with CLIP-FT or SigLIP-FT for a fair comparison.

Based on the experimental results, we have the following observations: **(1) Strong generalization:** Our method consistently improves compositional understanding across models with different scales and training paradigms (*e.g.*, InfoNCE loss vs. pairwise sigmoid loss), demonstrating strong generalization potential. **(2) Greater improvements on contrastively trained models:** Compared to SigLIP ViT-B/16, CLIP ViT-B/16 exhibits larger performance gains from our method. This may be due to the fact that SigLIP does not adopt explicit batch-level contrastive learning but instead relies on pairwise contrast, while our auxiliary losses are better aligned with batch-level contrastive learning paradigms.

J Detailed Version of Main Experimental Results

We present the detailed results on all benchmark subsets in Table 17, Table 18, Table 19 and Table 20. As shown, our method achieves the best performance on nearly all subsets across the five benchmarks.

K Experiments Beyond COCO Domain

To further validate the effectiveness of our method beyond the MSCOCO dataset, we conduct experiments in two respects: on the one hand, we evaluate it on out-of-distribution benchmarks outside the MSCOCO domain; on the other hand, we train it on non-COCO datasets.

Model	ARO			Sugar-Crepe		VL-Checklist		VALSE	What's-up
	Relation	Attribute	Order	Relation	Attribute	Relation	Attribute	Relation	Relation
Random Chance	50.0	50.0	20.0	50.0	50.0	50.0	50.0	50.0	41.7
CLIP (ViT-B/32)	58.7	62.7	54.1	68.8	70.8	63.6	67.7	70.1	41.8
CLIP-FT	64.4±0.40	66.2±0.05	48.8±0.28	71.1±0.26	77.4±0.22	60.6±0.22	67.4±0.05	69.4±0.30	41.3±0.16
MACCO-CLIP (ours)	**73.5**±0.60	**69.1**±0.64	**75.0**±1.14	**76.1**±0.95	**78.4**±0.50	**69.9**±1.03	**68.9**±0.62	**75.2**±0.44	**43.0**±0.72

Table 15: **Multiple Runs.** We report the mean and standard deviation over four training runs of CLIP-FT and our MACCO-CLIP with four different random seeds.

Model	ARO			Sugar-Crepe		VL-Checklist		VALSE	What's-up
	Relation	Attribute	Order	Relation	Attribute	Relation	Attribute	Relation	Relation
Random Chance	50.0	50.0	20.0	50.0	50.0	50.0	50.0	50.0	41.7
Backbone: CLIP ViT-B/16									
CLIP	59.9	62.0	54.1	66.3	70.5	61.7	68.8	68.8	41.9
CLIP-FT	61.2	62.3	39.9	71.6	78.5	56.9	68.5	67.7	**44.2**
MACCO-CLIP (ours)	**73.4**	**67.3**	**68.0**	**76.2**	**79.5**	**66.5**	**68.9**	**70.4**	41.7
Backbone: CLIP ViT-L/14									
CLIP	61.7	61.7	51.3	65.0	70.8	64.7	68.0	66.7	41.2
CLIP-FT	58.1	63.8	38.4	75.3	78.9	63.0	71.8	71.4	42.0
MACCO-CLIP (ours)	**72.6**	**65.7**	**59.8**	**77.3**	**79.3**	**72.5**	**72.1**	**74.0**	**42.4**
Backbone: SigLIP ViT-B/16									
SigLIP	26.6	44.6	11.1	57.9	67.6	42.0	59.8	53.5	**41.5**
SigLIP-FT	48.3	**67.5**	33.9	75.0	82.9	60.6	69.9	67.4	40.5
MACCO-SigLIP (ours)	**54.6**	67.0	**66.0**	**76.2**	**84.2**	**63.7**	**71.2**	**71.2**	40.8

Table 16: **Extended experimental results on other model scales.**

Model	ARO				
	Relation	Attribute	COCO-Order	Flicker-Order	Avg.
CLIP	58.7	62.7	48.0	60.2	57.4
CLIP-FT	64.3	66.2	43.3	54.8	57.2
IL-CLIP	50.0	55.3	16.8	16.6	34.7
SDS-CLIP	53.0	62.0	24.0	34.0	43.3
CLIP-CAE	69.5	65.4	-	-	-
MACCO-CLIP	**73.1**	**68.5**	**72.3**	**79.6**	**73.4**

Table 17: **Detailed results on ARO (Yuksekgonul et al., 2023).** In the main paper, we take the average performance of the model on the COCO-Order and Flickr-Order subsets as the performance of ARO-Order.

Model	Sugar-Crepe									
	REPLACE				SWAP			ADD		
	Relation	Attribte	Object	Avg.	Attribute	Object	Avg.	Attribute	Object	Avg.
CLIP	68.9	80.1	90.8	79.9	63.5	60.4	62.0	68.4	76.9	72.7
CLIP-FT	71.1	84.1	**92.9**	82.7	**70.3**	69.0	69.7	78.6	**87.2**	82.9
IL-CLIP	56.3	66.4	77.7	66.8	54.7	54.7	54.7	70.8	65.4	68.1
MACCO-CLIP	**77.1**	**84.6**	92.1	**84.6**	70.0	**72.2**	**71.1**	**82.8**	86.5	**84.7**

Table 18: **Detailed results on Sugar-Crepe (Hsieh et al., 2024).**

Model	VL-Checklist											
	Relation			Attribute						Object		
	Action	Spatial	Avg.	Action	Color	Material	Size	State	Avg.	Location	Size	Avg.
CLIP	71.3	55.7	63.6	73.3	69.3	66.5	63.6	65.4	67.7	77.4	76.2	76.8
CLIP-FT	70.7	51.1	60.9	74.1	73.1	66.9	60.4	62.4	67.4	79.8	78.3	79.1
IL-CLIP	62.3	49.0	55.7	64.8	65.0	61.9	48.9	57.1	59.5	71.1	67.3	69.2
MACCO-CLIP	**75.4**	**65.1**	**70.2**	**75.3**	**73.6**	**70.7**	**57.0**	**66.6**	**68.7**	**82.0**	**79.8**	**80.9**

Table 19: **Detailed results on VL-Checklist (Zhao et al., 2022).**

Model	VALSE			What's-Up			
	Action	Relation	Avg.	Whats'Up	COCO-spatial	GQA-spatial	Avg.
CLIP	74.8	65.4	70.1	31.1	47.4	46.9	41.8
CLIP-FT	73.5	65.1	69.3	30.7	46.9	46.5	41.4
IL-CLIP	58.5	52.9	55.7	26.0	**52.3**	**48.5**	42.3
MACCO-CLIP	**78.6**	**72.0**	**75.3**	**34.2**	47.1	48.3	**43.2**

Table 20: **Detailed results on VALSE (Parcalabescu et al., 2021) and What's-up (Kamath et al., 2023b).**

(1) Evaluation results on Winoground and MMVP. We conduct additional evaluations on two challenging out-of-distribution compositional reasoning benchmarks: Winoground (Thrush et al., 2022) and MMVP (Tong et al., 2024). Both benchmarks consist of image-text pairs that lie beyond the COCO domain, and are widely recognized as some of the most difficult benchmarks in the field, with the results presented in Table 21. As shown, our method achieves consistent improvements over the baseline on both benchmarks. The relatively smaller gains on Winoground may be attributed to the intrinsic difficulty of the benchmark (Diwan et al., 2022), and evaluation on Winoground has also been noted to present out-of-distribution challenges (Zhang et al., 2024; Li et al., 2024). Furthermore, the performance gains on MMVP are a promising signal. Although our method primarily aims to enhance the encoding capability of the text encoder, the gains on a vision-centric task like MMVP suggest that our approach may also benefit multimodal large language models that use CLIP as the vision encoder, such as LLaVA (Liu et al., 2023). In Section 4.6, we replace the vision encoder of LLaVA-1.5-7B (Liu et al., 2024a) with MACCO-CLIP (ViT-L/14) and vanilla CLIP (ViT-L/14) and train under identical settings. The results in Table 8 show that using vision encoder from our MACCO-CLIP yields better performance in mitigating compositional semantic hallucinations.

(2) Training beyond COCO. To further validate the generalization capability of our method beyond the COCO domain, we conduct experiments using a subset of CC3M released by the excellent work (Oh et al., 2024), which contains approximately 100k samples. We retrain both CLIP-FT and MACCO on this dataset and evaluate them on five compositional reasoning benchmarks. As shown in Table 22, MACCO significantly outperforms the baseline across all five benchmarks. This further substantiates the effectiveness of our method and demonstrates that its benefits are not limited to object-centric datasets like COCO.

This cross-domain effectiveness, combined with the OOD benchmark results, provides strong evidence that MACCO's benefits generalize beyond the specific structure of COCO.

Model	Winoground			MMVP
	Text score	Image score	Group score	Avg.
CLIP	31.6	11.1	9.4	14.8
CLIP-FT	32.2	8.8	5.9	20.7
MACCO-CLIP (ours)	32.2	11.1	8.2	21.5

Table 21: **Experiment results on Winoground and MMVP.**

L Discussion with Related Works About Masked Modeling

MaskVLM (Kwon et al., 2022) and CLIP-CAEv2 (Zhang et al., 2022) are two studies closely related to our work. While MaskVLM is an influential work in vision-language pretraining, our method differs from it in both the masking strategy and the training objective:

(1) Task focus and masking strategy is different. MaskVLM focuses on multimodal pretraining and is primarily designed for general multimodal

Model	Compositional Understanding Benchmarks				
	ARO	SugarCrepe	VL-Checklist	VALSE	What's-up
CLIP	58.5	69.8	65.7	70.1	41.8
CLIP-FT	64.5	75.7	67.5	70.6	41.2
MACCO-CLIP (ours)	**72.8**	**76.8**	**70.9**	**73.4**	**42.3**

Table 22: **Experimental results of models trained on CC3M.**

tasks, which makes the use of random masking appropriate. In contrast, our goal is to enhance fine-grained compositional understanding, which necessitates a more targeted masking strategy. Thus we introduce masking over compositional concepts spanning both text and image modalities.

(2) Training objective is different. As masked signal modeling primarily constrains local tokens, whereas both contrastive learning paradigms and downstream tasks rely on global tokens, thus designing compositional masking alone is not sufficient. Therefore, we incorporate global tokens into the masked modeling process to jointly optimize global representations and facilitate reconstruction. Specifically, we introduce a global-to-local semantic injection strategy. To ensure that the masked global tokens in global-to-local semantic injection carry meaningful semantics, we further propose two masked-augmented auxiliary losses to constrain the masked global tokens. And the masking strategy of CLIP-CAE v2 is also random (applied only to the image modality).

As pretraining methods can also be adapted for fine-tuning, we conduct two additional experiments to compare our method with settings that adopt the pretraining strategies to those used in MaskVLM and CLIP-CAEv2. The results are presented in Table 23. As shown, directly transferring the masked modeling strategies from these influential pretraining methods does not yield significant improvements, and their performance is consistently lower than ours across all benchmarks (with average gains of 1.4% and 2.8%, compared to our 7.3%). These results further validate the effectiveness of our method, which uses a more targeted masking strategy, along with two auxiliary losses and a global-to-local semantic injection strategy.

M Discussion About the Detection Model

Our method primarily leverages the object detection capability of advanced grounding models to identify the object regions corresponding to grounding phrases and apply masking over the full object area. We analyze the robustness of our method for the detection model from the following two aspects.

(1) Highly proficient visual grounding model is not a strict requirement. Since COCO captions typically describe the prominent objects in an image, the mentioned objects are generally easy to ground. We randomly sample 100 examples from the training set and found that only 7 of them contained objects described in the captions that are visually ambiguous and required fine-grained attribute reasoning for accurate localization. This suggests that our method does not heavily rely on strong visual grounding capabilities. To further validate this claim, we replaced GroundingDINO with OWLv2 (Minderer et al., 2023), a detection model that excels at open-set object recognition without explicit grounding training. The results are show in Table 24. As shown in the table, the model's performance using masks generated by OWLv2 is comparable to that with GroundingDINO, indicating that a highly proficient off-the-shelf visual grounding model is not a strict requirement, and a general open-set detector with strong object detection capabilities is sufficient.

(2) Robust to noisy detection results. To investigate the performance of our method under mild detection noise, we randomly replace the GroundingDINO's predictions with random rectangular bounding boxes with a 10% probability to simulate scenarios where some of GroundingDINO's outputs might contain noise. The experimental results are shown in Table 25. As shown in the results, even in the presence of some noise, our method still demonstrates significant improvement over the baseline, and even slightly outperforms the original data (by $+0.1\%$). This indicates that our approach exhibits a certain degree of robustness to potentially noisy grounding results, which further validates its reliability.

This result is reasonable because when the detection model introduces noise, the resulting mask becomes random block-wise masks. Since the im-

Model	ARO	Sugar-Crepe	VL-Checklist	Avg.
Baseline	59.9	74.4	64.2	66.1
MaskVLM	63.0	74.7	64.7	67.5 (+1.4%)
CLIP-CAEv2	66.0	75.3	65.3	68.9 (+2.8%)
MACCO-CLIP (ours)	**72.5**	**78.1**	**69.5**	**73.4 (+7.3%)**

Table 23: **Performance comparison with models utilizing the pretraining masking strategies of MaskVLM and CLIP-CAEv2.**

Model	Main Detection Strength	ARO		Sugar-Crepe		VL-Checklist	
		Relation	Attribute	Relation	Attribute	Relation	Attribute
CLIP	-	58.7	62.7	68.8	70.8	63.6	67.7
CLIP-FT	-	64.3	66.2	71.1	77.7	60.9	67.4
MACCO-CLIP (OWLv2)	category-level detection	72.0	68.3	76.9	**79.2**	69.6	**69.3**
MACCO-CLIP (Grounding DINO)	language-guided visual grounding, category-level detection	**73.1**	**68.5**	**77.1**	79.1	**70.2**	68.7

Table 24: **Experimental results when using an object detection model without explicit grounding training.**

age and text are paired, predicting random image regions in the image based on the complete text remains plausible. Moreover, this masking strategy is more challenging than patch-wise masking, requiring the model to more deeply understand the textual information and the alignment between the image and text. Additionally, both in the fields of computer vision and natural language processing, many studies have demonstrated the advantages of block-wise masked modeling over random masking. In the field of computer vision, block-wise masked modeling has been proven to be more effective than random masking, as shown in BEiT (Bao et al., 2021). Furthermore, in MAE (He et al., 2022), the authors performed ablation experiments on block-wise masked modeling (Table 1(f) in the MAE paper). The experimental results indicate that the performance of linear probing and fine-tuning with random block-wise masking pretraining is only 1.2% and 1.0% worse than that of random masking, respectively. This demonstrates that block-wise masking is indeed a highly effective strategy in the computer vision domain. And in the field of NLP, block-wise (or n-gram) masking is widely used in BERT-like models such as SpanBERT (Joshi et al., 2020) and UniLMv2 (Bao et al., 2020).

Model	ARO	Sugar-Crepe	VL-Checklist	Avg.
Baseline	59.9	74.4	64.2	66.1
MACCO-CLIP (with 10% noisy training data)	72.9	78.2	69.4	73.5 (+7.4%)
MACCO-CLIP	72.5	78.1	69.5	73.4 (+7.3%)

Table 25: **Experimental results under noisy outputs from GroundingDINO.**

N Discussion with Hard-negative Methods

Addressing compositional understanding in vision-language models is a critical challenge, and would like to discuss this issue from two points:

(1) Acknowledging contributions of prior work: We recognize and respect the pivotal contributions made by prior works in advancing compositional understanding in VLMs. Methods such as NegCLIP (Yuksekgonul et al., 2023), TSVLC (Doveh et al., 2023b), DAC (Doveh et al., 2023a), CE-CLIP (Zhang et al., 2024), syn-CLIP (Cascante-Bonilla et al., 2023), IL-CLIP (Zheng et al., 2024), FSC-CLIP (Oh et al., 2024), Triplet-CLIP (Patel et al., 2024), and recent research like CLIP-CAE (Li et al., 2024) and SDS-CLIP (Basu et al., 2024) have significantly progressed the field. These approaches predominantly focus on data-driven strategies, particularly through constructing hard negatives, with some methods (*e.g.*, SDS-CLIP and CLIP-CAE) improving the loss function. These innovations have markedly enhanced vision-language compositionality and multimodal research as a whole. Additionally, benchmarks like Sugar-Crepe have been instrumental in evaluating compositional understanding, playing a crucial role in identifying and addressing the limitations of VLMs.

(2) Our framework and its relationship to hard-negative mining: Improving compositional understanding in VLMs remains a significant challenge, especially when relying on standard image-text pairs. While many existing methods, such as the seminal NegCLIP, focus on hard-negative mining, our approach takes a different path by em-

phasizing the design of improved training frameworks and loss functions. These methods should be considered orthogonal from a machine learning perspective. For instance, in alignment with SDS-CLIP and CLIP-CAE, which do not compare directly against hard-negative mining strategies, our framework integrates seamlessly with hard-negative mining methods (*e.g.*, NegCLIP) to achieve additional gains, as demonstrated in Section 4.2 of the main text. This compatibility further validates the effectiveness of our design and highlights the complementary nature of these approaches.

O Discussion About Efficacy on Spatial Relationships

For spatial relationships, on the text side, we can accurately mask the spatial relationship expressions, which enables the model to learn to better understand spatial relations in the image through cross-modal masked modeling. On the image side, while spatial relations cannot be directly grounded to specific regions, we can mask the regions corresponding to the two related objects and use the full text to guide reconstruction. Our motivation is that if the model understands the spatial relationship described in the text (*e.g.*, "object A is to the left of object B"), it should be able to reconstruct the relative positions of the two objects. In this way, we aim to enhance the model's ability to interpret spatial relationships described in text. We present the performance of our method on several benchmark subsets specifically designed to evaluate spatial relation understanding in the Table 26. As shown, our method brings significant improvements in this type, indicating its effectiveness in enhancing the model's comprehension of spatial relationships.

P Discussion on Language Generalizability and Tool Dependency

Our framework leverages external tools for concept extraction, which may raise question about their availability and accuracy across diverse languages. We discuss the generalizability and robustness of our approach from two key perspectives:

(1) Modular flexibility for multilingual support. While our current implementation leverages a specific English scene graph parser (Wu et al., 2019), the MACCO framework is inherently modular and not restricted to any particular legacy tool. For non-English or low-resource languages, compositional concepts (objects, attributes, and relations) can be effectively extracted using modern open-source large language models (*i.e.*, Qwen) or multilingual NLP toolkits (*i.e.*, Stanza or spaCy). Recent studies (Do et al., 2024, 2025) demonstrate that LLMs are highly proficient in zero-shot compositional parsing across diverse languages, underscoring the broad applicability of our approach to multilingual scenarios.

(2) Resilience to imperfect tools. From a visual perspective, our in-depth analysis in Appendix M demonstrates that MACCO achieves robust performance even when the detection model is imperfect. From a textual perspective, if the parser fails to accurately identify compositional concepts in certain languages, our masking strategy degrades to a random masking approach at worst. As revealed in our ablation study (Section 4.5), even under random masking, our framework consistently outperforms the baseline, although masking compositional concepts yields the most substantial improvement. This empirical evidence highlights that MACCO is resilient to parsing inaccuracies, leveraging these tools as a guided prior that remains effective even in the presence of partial noise.

Q Discussion of Potential Biases from Pre-trained Tools

MACCO leverages external tools to perform compositional concept extraction. Although these tools are well-established in practice, they inevitably introduce certain systemic biases such as neglecting long tail patterns. We discuss how MACCO mitigates the potential impact resulting from these limitations from the following perspectives.

(1) Implicit constraints on rare compositional patterns: MACCO does not solely rely on the "labels" provided by parsers, it uses these labels to generate masking signals that guide cross-modal reconstruction. Even if pre-trained tools overlook rare compositional patterns, the global features from the full and masked signals still retain information about these patterns. Our proposed global-to-local semantic injection operation integrates global semantic features into local tokens during reconstruction. This mechanism introduces implicit constraints on rare compositional patterns, thereby mitigating the impact of parser limitations.

(2) Feature space regularization facilitating rare compositional pattern learning: The two masked-augmented losses we propose

Model	VL-Checklist (spatial relation subset)	VALSE (spatial relation subset)	What's-up (designed for evaluate spatial relation)
CLIP	55.7	65.4	41.8
CLIP-FT	51.1	65.1	41.4
MACCO-CLIP	**65.1 (+14)**	**72.0 (+6.9)**	**43.2 (+1.8)**

Table 26: **Performance on several benchmark subsets specifically designed to evaluate spatial relation understanding.**

(L_{MCA}&L_{MIR}) further regularize the feature space, acting as an additional implicit constraint. This ensures that the model's embedding space is grounded in the entire data distribution of natural image-text pairs, rather than being overfitted to the frequent patterns disproportionately emphasized by pre-trained tools. As a result, our approach prevents the model from losing sensitivity to infrequent but meaningful compositional patterns.

R Discussion with X-VLM

X-VLM (Zeng et al., 2022) is a pioneering work in multi-grained vision-language pre-training. In contrast, MACCO introduces a fundamental paradigm shift from "Explicit Alignment via Localization" to "Implicit Alignment via Reconstruction".

Visual Concept Localization vs. Compositional Concept Reconstruction: X-VLM relies on the model's ability to "point" to where a concept resides in the image using explicit bounding boxes. In contrast, MACCO emphasizes compositional reasoning. By strategically masking compositional anchors instead of random tokens, we encourage the model to infer missing structural dependencies through cross-modal context. This approach is inherently more demanding than localization, as it necessitates that the model internalizes how attributes and relations connect and bind to objects.

General MLM in Language vs. Targeted Masking in Vision and Language: While X-VLM includes a general MLM loss, MACCO adopts a more targeted approach by extracting and masking compositional concepts. Our approach addresses the "bag-of-words" bias in contrastive models by ensuring that the model cannot solve the reconstruction task without understanding the contextual interplay between concepts across modalities.

Future Paths for Enhancing MACCO with Multi-Grained Reasoning: Beyond these distinctions, we believe that X-VLM's multi-grained framework and MACCO's implicit reconstruction paradigm are highly complementary. For instance, X-VLM's multi-grained reasoning could naturally integrate with MACCO by introducing multi-grained reconstruction strategies, such as reconstructing compositional text while considering the global image or corresponding image regions. Furthermore, X-VLM's multi-grained contrastive loss could be adapted to anchor compositional concepts (*e.g.*, specific attribute-object pairs) to their associated visual regions.

S Discussion About the Attribute Binding Challenge

Attribute binding is indeed a persistent bottleneck in the field of multimodal learning. We would like to discuss this from two perspectives:

(1) Why attribute binding is fundamentally challenging. Attribute binding is inherently more difficult than relation modeling because attributes (*e.g.*, color, material, size) are often visually and semantically entangled with the objects they modify. In contrastive VLMs, models can exploit shortcuts and exhibit a "bag-of-words" behavior, detecting the presence of individual concepts (*e.g.*, "red", "car") without encoding the structural association that binds the attribute (*e.g.*, "red") to the correct object (*e.g.*, "car"). Consistent with findings in compositional text-to-image synthesis (Huang et al., 2025; Chefer et al., 2023; Rassin et al., 2023) and VLM research (Johnson et al., 2017), attribute binding remains a persistent challenge. Attribute cues are often spatially fused with object evidence in the same image regions, which makes them difficult to disentangle. By contrast, relations (*e.g.*, "on", "next to") typically have clearer geometric signatures.

(2) Preliminary ideas for future enhancements on attribute binding. First, *optimizing attention mechanisms.* During training, selectively blocking attention interactions across different attribute-phrase groups can reduce feature entanglement among attributes. Second, *token merging for attribute binding.* During inference, merging tokens that correspond to the same attribute phrase in the VLM encoder can encourage stronger

binding between attributes and their associated objects. Third, *synthetic data generation.* Leveraging powerful LLMs and text-to-image generation models to synthesize datasets enriched with diverse attribute combinations can improve generalization to rare attributes. Fourth, *enhanced text encoders.* Employing medium-sized LLMs as the text encoder can strengthen the model's ability to parse and understand complex attribute bindings in text.

T Discussion About Future Work Using Strong LLM-based Text Encoders

Zhang et al. (2025) highlights that the text encoder trained under the original CLIP paradigm has limited language understanding capabilities. In contrast, incorporating more powerful pretrained language models as the text encoder presents a promising direction for building robust foundation vision-language models. We discuss this from the following two perspectives:

Advantages of using LLM as text encoder: (1) Stronger language understanding and reasoning capabilities. LLMs possess richer syntactic, semantic, and contextual modeling abilities, enabling them to capture complex linguistic structures. This can significantly enhance a model's understanding of compositional relations and semantic nuances, which is particularly important for tasks that require complex reasoning. **(2) Better generalization ability with longer context.** LLMs are typically trained on large-scale, open-domain corpora, making them more robust to rare vocabulary, long-tail compositional patterns, and and more effective to understand long sentences containing complex linguistic structures. As a result, they tend to generalize better in zero-shot and open-world settings.

Challenges of using LLM as text encoder: (1) Higher computational cost and deployment complexity. Compared to CLIP's original lightweight text encoder, using an LLM significantly increases the number of parameters, training cost and inference latency, potentially limiting deployment in real-time scenarios. Efficient training strategy like SAIL proposed in the paper provides a promising solution to mitigate this issue. **(2) Potentially increased difficulty in cross-modal alignment.** Although LLMs produce powerful semantic representations, these may not align naturally with visual feature spaces. Especially without joint training or fine-tuning, there may be large semantic gap between modalities, which can degrade image-text alignment and contrastive learning efficiency.

In summary, replacing the CLIP-style text encoder with a strong LLM holds significant potential for improving language understanding and overall generalization in vision-language models, especially for tasks requiring complex compositional reasoning. It is a promising and increasingly recognized direction. Nonetheless, this approach also introduces new challenges related to cross-modal alignment and resource demands. In the future, we also hope to extend our framework to LLM-based CLIP-like models to further explore this direction.