UNDERSTANDING HARDNESS OF VISION-LANGUAGE COMPOSITIONALITY FROM A TOKEN-LEVEL CAUSAL LENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Language–Image Pre-training (CLIP) achieves striking cross-modal generalization by aligning images and texts in a shared embedding space, yet it persistently fails at compositional reasoning over objects, attributes, and relations—often behaving as a bag-of-words matcher. Existing causal accounts of CLIP largely model text as a single vector, obscuring token-level structure and leaving core phenomena—such as prompt sensitivity and failures on hard negatives—unexplained. We address this gap by developing a token-aware causal representation learning (CRL) framework grounded in a sequential, language-token SCM. Our theory extends block identifiability results to tokenized text, proving that CLIP's contrastive objective can recover the modal-invariant latent variable under both sentence-level and token-level SCMs. Crucially, the token granularity enables the first principled explanation of CLIP's compositional brittleness: composition nonidentifiability. We show that there exist pseudo-optimal text encoders that achieve perfect modal-invariant alignment yet are provably insensitive to SWAP, REPLACE, and ADD operations over the atomic concepts on objects, attributes, and relations, thereby failing to distinguish correct captions from hard negatives—despite optimizing the same training objective as true-optimal encoders. The analysis further connects language-side nonidentifiability with visual-side failures via the observed modality gap, and demonstrates how iterated composition operators compound hardness, suggesting improved negative mining strategies.

1 INTRODUCTION

Throughout the phylogeny of multimodal intelligence, Contrastive Language-Image Pre-training (CLIP, Radford et al. (2021)) emerged as a milestone for its exceptional ability to bridge vision and language. Trained on billions of image-text pairs, CLIP demonstrates remarkable robustness, evident in its out-of-distribution (OOD) generalization and zero-shot inference capabilities using textual prompts. From the lens of causal representation (Scholkopf et al. (2021); Yao et al. (2023)), the performance leap is largely attributed to learning a shared embedding space that achieves *modal-invariant alignment* between visual and textual features.

Despite these strengths, CLIP struggles with compositional reasoning across images and text, which arises from its weakness to isolate the hard negative structures composed of atomic concepts, *i.e.*, object, attribute, and relation (Yuksekgonul et al. (2023); Ma et al. (2023); Hsieh et al. (2023)). It often acts like a bag-of-words matcher, identifying concepts individually but failing to bind them to their specified order, attributes, or relationships derived from the images' correct descriptions, in other words, CLIP may confuse "a bulb in the grass" with "grass in a bulb," misinterpret attribute-noun pairings, or default to common co-occurrences instead of the specific composition described. These failures reveal that its embedding space unreliably encodes the compositional structure required for precise, human-like understanding in vision-language tasks.

This phenomenon has spurred a wave of empirical research to evaluate and remedy CLIP's compositional weaknesses. Although massive benchmarks and solutions (Hsieh et al. (2023); Patel et al. (2024)) were proposed, a rigorous theoretical explanation for why CLIP models falter remains elusive. Much of the existing theoretical work on CLIP simplifies the problem by modeling entire images

and text prompts as monolithic, fixed-length vectors. This abstraction, by its very nature, overlooks the compositional structure of atomic concepts, which presents as tokens at the heart of the issue analysis, leaving a critical gap in our ability to formally diagnose and understand these failures.

Motivated by this gap, our research aims for the first principled explanation to the difficulty behind vision-language compositionality. The breakthrough roots in a more granular causal representation theory to locate each token contribution to achieve the modal-invariant alignment. Specifically, our framework generalizes the existing SCMs of most multimodal CRL studies with our underlying text generation process defined by language-token sequence, enlighten by the memory-argumented Bayesian prior in the recent theoretic understanding of language generation (Wei et al. (2021)). The nuance refers to the causal representation with the consistent result in modal-invariant alignment in CLIP (Theorem.5, Corollary.6). While thanks to the token awareness in our practical premise, our framework provided new theoretical findings from a causal lens of understanding the image-text embedding space.

Our very first principled explanation for CLIP's compositional reasoning failures, which we termed "*composition nonidentifiability*" in the textual description. We formally prove (Theorems 7-9) with the existence of "pseudo-optimal" text encoders that achieve the same modal-invariant alignment as a "true" encoder during pre-training, however, the former fail to distinguish correct textual descriptions from hard negatives constructed through SWAP, REPLACE, and ADD operations considered as representative forms of hard negatives (Ma et al. (2023),Hsieh et al. (2023)). Since CLIP's training objective cannot differentiate between these "true-optimal" and "pseudo-optimal" solutions, the model is not guaranteed to learn the underlying compositional structure, which rigorously explains its vulnerability to confusing concepts and their relationships. This theoretical framework also extends to explain visual compositionality issues by combining the constant modality gap phenomenons (Zhang et al.; Chen et al. (2023)), and shows that iteratively applying these operations can generate more complex hard negatives, suggesting a path toward improving models via advanced negative mining.

2 PRELIMINARIES

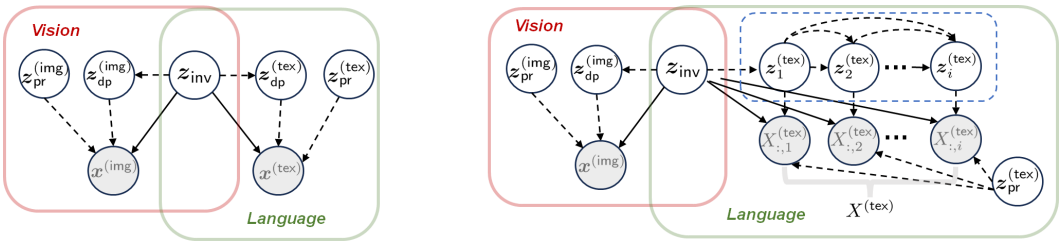
In this section, we briefly introduce Contrastive Language-Image Pre-training (CLIP), then go through its explainable theory derived from causal representation learning (CRL). A foundational introduction of CLIP-based research and structural causal models (SCMs) is helpful for understanding, and we recommend the readers access the background and related work in our Appendix.A.

2.1 CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING (CLIP)

The CLIP family Radford et al. (2021); Jia et al. (2021); Cherti et al. (2023) receives data coupled by image and text in mutual semantic through contrastive pre-training Oord et al. (2018); He et al. (2020). Suppose $\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle \sim p_{\mathsf{mm}}\big(\boldsymbol{x}^{(\mathsf{img})}, \boldsymbol{x}^{(\mathsf{tex})}\big)$ denotes an image-text pair drawn from a multimodal joint distribution p_{mm} (*i.e.* p_{mm}), the measure to indicate the mutual semantic across modalities. CLIP's image encoder $f(\cdot)$ and text encoder $g(\cdot)$ extract their normalized features $f(x^{(\mathsf{img})})$, $g(x^{(\mathsf{tex})})$ to construct InfoNCE objectives

$$
\begin{aligned}
&\min_{f,g} \mathbb{E}_{\mathcal{D}^{(K)} \sim p_{\mathsf{mm}}} \Big[\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) + \mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big)\Big] \\
\text{s.t. } &\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) = \sum_{i=1}^{K} -\log \frac{e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{j=1}^{K} e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_j^{(\mathsf{tex})})/\gamma\right)}}, \\
&\mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) = \sum_{i=1}^{K} -\log \frac{e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{j=1}^{K} e^{\left(f(x_j^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}
\end{aligned}
\tag{1}
$$

where $\mathcal{D}^{(K)} = \{\langle x_i^{(\mathsf{img})}, x_i^{(\mathsf{tex})} \rangle\}_{i=1}^{K}$ indicates the training batch composed of K image-text pairs, $\{x_i^{(\mathsf{img})}, x_i^{(\mathsf{tex})}\}_{i=1}^{K}$ indicates each training batch constructed by K image-text pairs drawn from the joint distribution p_{mm}, by which InfoNCE distinguishes the positive pairs sampled from p_{mm} against negative pairs sampled from the image and the text marginals derived from p_{mm}.



(a). Token-agnostic SCM for p_{mm} **(b).** Token-aware SCM for p_{mm}

Figure 1: Latent-variable SCMs that represents the multimodal image-text data generation processes from the sentence-level aspect (Assumption 1 (a)) and the token-level aspect (Assumption 4 (b)). The goal of causal representation learning seeks for the unsupervised recovery of the modal-shared latent variable $\boldsymbol{z}_{\mathsf{inv}}$ by CLIP, which were rigorously justified in Theorem.2, 5.

2.2 Conventional Causal Representation for Multimodal Contrastive Training

Under p_{mm} interpreted as the generative process defined by a SCM with some latent variable $\boldsymbol{z}_{\mathsf{inv}}$ shared across modalities, CRL demonstrates multimodal contrastive training (Eq.1) implicitly achieving the unsupervised recovery of the latent variable z_{inv} from $\boldsymbol{z}^{(\mathsf{inv})}$. To analyze CLIP, CRL demands the SCM assumption of multimodal data distribution to generate image-text training pairs:

Assumption 1. (**Token-agnostic SCM of image-text data generation**, Fig.1.a) The mutual semantics between image-text pairs are derived from the modal-invariant feature drawn from modal invariant density, *i.e.*, $z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}}$; given z_{inv}, we obtain image-dependent partition $z_{\mathsf{dp}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{img})}}(\cdot|z_{\mathsf{inv}})$ and text-dependent partition $z_{\mathsf{dp}}^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{tex})}}(\cdot|z_{\mathsf{inv}})$ specific to the image domain and text domain, respectively; and we also have the image-private partition $z_{\mathsf{pr}}^{(\mathsf{img})}$ and text-private partition $z_{\mathsf{pr}}^{(\mathsf{tex})}$ drawn from independent priors, *i.e.*, $z_{\mathsf{pr}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{img})}}$, $z_{\mathsf{pr}}^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{tex})}}$; then each image-text pair $\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle$ is generated through the nonlinear mixing functions $\mathbf{f}$,$\mathbf{g}$ to specify p_{mm}:

$$\begin{aligned} x^{(\mathsf{img})} &:= \mathbf{f}(z^{(\mathsf{img})}) = \mathbf{f}\big(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{img})}, z_{\mathsf{pr}}^{(\mathsf{img})}\big); \\ x^{(\mathsf{tex})} &:= \mathbf{g}(z^{(\mathsf{tex})}) = \mathbf{g}\big(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{tex})}, z_{\mathsf{pr}}^{(\mathsf{tex})}\big), \end{aligned} \tag{2}$$

where z_{inv}, $z_{\mathsf{dp}}^{(\mathsf{img})}$, $z_{\mathsf{pr}}^{(\mathsf{img})}$, $z_{\mathsf{dp}}^{(\mathsf{tex})}$, $z_{\mathsf{pr}}^{(\mathsf{tex})}$ denote real-value vectors drawn from the distributions with respect to $\boldsymbol{z}_{\mathsf{inv}}$, $\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{img})}$, $\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{img})}$, $\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{tex})}$, $\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{tex})}$ over the SCM generative process.

The assumption above is extended from the SCM defined in (Daunhawer et al. (2022)) to interpret the underlying causation in multimodal contrastive model, where their differences lie in the relation between $\boldsymbol{z}_{\mathsf{inv}}$ and $\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{tex})}$. Derived from the relaxed premise, CLIP still holds the alignment to identify the modal-invariant part of each image-text pair:

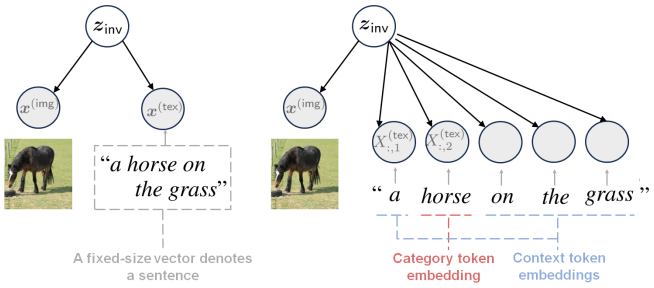
Theorem 2. ***(Block-Identified Modal-invariant Alignment (Token-agnostic))*** *Consider the image-text pair generated by Assumption.1. If their densities and mappings satisfy:* ***1)****.* $\mathbf{f}$, $\mathbf{g}$[1] *are diffeomorphisms;* ***2)****.* $\boldsymbol{z}^{(\mathsf{img})}$, $\boldsymbol{z}^{(\mathsf{tex})}$ *are smooth, with continuous distributions* $p_{\boldsymbol{z}^{(\mathsf{img})}} > 0$, $p_{\boldsymbol{z}^{(\mathsf{tex})}} > 0$ *almost everywhere. Consider the image encoder* $f : \mathcal{X}_{\mathsf{img}} \to (0,1)^{n_{\mathsf{inv}}}$ *and the text encoder* $g : \mathcal{X}_{\mathsf{tex}} \to (0,1)^{n_{\mathsf{inv}}}$ *as smooth functions that are trained to jointly minimize the functionals,*

$$\begin{aligned} \mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})} := \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}}} \Big[||f(x^{(\mathsf{img})}) - g(x^{(\mathsf{tex})})|| \Big] \\ - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{x}^{(\mathsf{tex})})) \end{aligned} \tag{3}$$

where $H(\cdot)$ *denotes the differential entropy of the random variables* $f(\boldsymbol{x}^{(\mathsf{img})})$ *and* $g(\boldsymbol{x}^{(\mathsf{tex})})$ *taking value in* $(0,1)^{n_{\mathsf{inv}}}$*. Then given the optimal image encoder* f^* *and the text encoder* g^**, there exist invertible functions* h_f *and* h_g *satisfying the following decompositions, respectively:*

$$f^* = h_f \circ \mathbf{f}_{1:n_{\mathsf{inv}}}^{-1}, \; g^* = h_g \circ \mathbf{g}_{1:n_{\mathsf{inv}}}^{-1} \tag{4}$$

[1]Ought to be regarded that we consider the output of $\mathbf{g}$ lies on a continuous space rather than discrete words and phrases. It allows for more feasible cases *e.g.*, soft prompts Zhou et al. (2022) for both Assumption.1 and 4 .



(a). The latent variable z_{inv} recovered by CLIP's encoders (token-agnostic CRL)

(b). The latent variable z_{inv} recovered by CLIP's encoders (token-aware CRL)

Figure 2: The comparison between **(a)** existing multimodal CRL theory (Daunhawer et al. (2022)) and **(b)** our CRL theory (Theorem.5 and Corollary.6). Our framework allows the analysis to CLIP with the word-and-phrase granularity, leading to our contributions to theoretically explain the CLIP weakness in compositional understanding (Section.4).

Corollary 3. *(Informal) The optimal encoders f^*, g^* in Theorem.2 are obtained if and only if* $(f^*, g^*) = \arg\min_{f,g} \mathcal{L}^{(\text{img}\to\text{tex})}_{\text{InfoNCE}} + \mathcal{L}^{(\text{tex}\to\text{img})}_{\text{InfoNCE}}$ *with infinite training pairs.*

Grounded in the principles of block identifiability (Von Kügelgen et al. (2021)), Theorem.2 demonstrates how optimal encoders can achieve modal invariance. It proves that under a mild assumption on the underlying data distribution of multimodal pairs, the optimal encoders (f^*, g^*) learn features that isolate a shared latent variable, $\boldsymbol{z}_{\text{inv}}$. This variable encapsulates all semantic information common to both the language and image modalities while simultaneously filtering out unshared, modality-specific information. This result provides a formal explanation for how CLIP's training objective leads to the cross-modal feature matching for the image and language representation.

3 Language-Token-Aware Causal Representation: Cornerstone to Interpret Compositional Reasoning Hardness

In this section, we generalize the statements of Theorem.2 as the inevitable path for interpreting the hardness of vision-language compositionality. In the pursuit of practical setup, we reconsider the assumption with the nonparametric functions that extend the text from a vector $x^{(\text{tex})} \sim p_{\boldsymbol{x}^{(\text{tex})}}$ to a k-column matrix $X^{(\text{tex,k})} \sim p_{\mathbf{X}^{(\text{tex,k})}}$, where $\forall \mathsf{k} \in \{1, \cdots, \mathsf{k}_{\max}\}$ indicates the sentence length and the i^{th} column $X^{(\text{tex,k})}_{:,i}$ indicates the i^{th} token embedding:

Assumption 4. (**Token-aware SCM of image-text data generation**, Fig.1.b) The mutual semantics between image-text pairs are derived via $z_{\text{inv}} \sim p_{\boldsymbol{z}_{\text{inv}}}$; given z_{inv}, the image-private partition $z^{(\text{img})}_{\text{pr}}$ and text-private partition $z^{(\text{tex})}_{\text{pr}}$ are drawn by $z^{(\text{img})}_{\text{pr}} \sim p_{\boldsymbol{z}^{(\text{img})}_{\text{pr}}}$, $z^{(\text{tex})}_{\text{pr}} \sim p_{\boldsymbol{z}^{(\text{tex})}_{\text{pr}}}$; and the image-dependent partition is obtained by $z^{(\text{img})}_{\text{dp}} \sim p_{\boldsymbol{z}^{(\text{img})}_{\text{dp}}}(\cdot|z_{\text{inv}})$. Suppose $z^{(\text{tex})}_i$ as the token-dependent partition of the i^{th} token, and each of them is recursively sampled via $z^{(\text{tex})}_i \sim p_{\boldsymbol{z}^{(\text{tex})}_i}(\cdot|z_{\text{inv}}, \{z^{(\text{tex})}_j\}_{j=1}^{i-1})$; then each image-text pair $\langle x^{(\text{img})}, X^{(\text{tex})} \rangle$ is generated through the nonlinear mixing functions $\mathbf{f}, \{\mathbf{g}_i\}_{i=1}^{\mathsf{k}_{\max}}$ to specify p_{mm}

$$
\begin{aligned}
x^{(\text{img})} &:= \mathbf{f}\big(z_{\text{inv}}, z^{(\text{img})}_{\text{dp}}, z^{(\text{img})}_{\text{pr}}\big); \\
X^{(\text{tex})}_{:,i} &:= \mathbf{g}_i\big(z_{\text{inv}}, \{z^{(\text{tex})}_j\}^{i}_{j=1}, z^{(\text{tex})}_{\text{pr}}\big).
\end{aligned} \tag{5}
$$

where the sampling stops at k^{th} step if $\mathsf{k} = \mathsf{k}_{\max}$ or $X^{(\text{tex})}_{:,k}$ reaches the embedding of [EOF].

Assumption.4 extends the image-language SCM definition in Assumption.1 by drawing the inspiration from the recent memory-argumented Bayesian LLM prior Wei et al. (2021). Derived from the token-level understanding to p_{mm}, we renew the block identifiability result to extend Them.2 from the sentence level to the token level:

Theorem 5. ***(Block-Identified Modal-invariant Alignment (Token-aware))*** *Consider the image-text pairs generated by Assumption.4 . If their densities and mappings meet:* ***1).*** $\mathbf{f}$ *and* $\mathbf{g}_i$ $(\forall i \in \{1, \cdots, \mathsf{k}_{\max}\})$ *are diffeomorphisms;* ***2).*** $\boldsymbol{z}^{(\text{img})}$, $\boldsymbol{z}^{(\text{tex})}_i$ $(\forall i \in \{1, \cdots, \mathsf{k}_{\max}\})$ *are smooth and with*

continuous distributions $p_{\boldsymbol{z}^{(\mathsf{img})}}>0$, $p_{\boldsymbol{z}_i^{(\mathsf{tex})}}>0$ *almost everywhere. Consider* $f:\mathcal{X}_{\mathsf{img}}\to(0,1)^{n_{\mathsf{inv}}}$ *and* $g:\cup_i^{\mathsf{k}_{\max}}\mathcal{X}_{\mathsf{tex}}^{(i)}\to(0,1)^{n_{\mathsf{inv}}}$ *as smooth functions that are trained to jointly minimize the functionals,*

$$\begin{aligned}\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})} :=& \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})}\rangle\\ \sim p_{\mathsf{mm}}}}\left[||f(x^{(\mathsf{img})})-g(X^{(\mathsf{tex})})||\right]\\ &-H(f(\boldsymbol{x}^{(\mathsf{img})}))\ -H(g(\boldsymbol{X}^{(\mathsf{tex})})),\end{aligned} \tag{6}$$

where $H(\cdot)$ *denotes the differential entropy of the random variables* $f(\boldsymbol{x}^{(\mathsf{img})})$ *and* $g(\boldsymbol{X}^{(\mathsf{tex})})$ *taking value in* $(0,1)^{n_{\mathsf{inv}}}$*. Then given the optimal image encoder* f^* *and the text encoder* g^**, there exist invertible functions* h_f *and* h_g *satisfying the following decompositions, respectively:*

$$f^*=h_f\circ\mathbf{f}_{1:n_{\mathsf{inv}}}^{-1},\ g^*=h_g\circ\mathbf{g}_{1:n_{\mathsf{inv}}}^{-1} \tag{7}$$

Corollary 6. *(Informal) The optimal encoders* f^**,* g^* *in Theorem.5 are obtained if and only if* $(f^*, g^*)=\arg\min_{f,g}\mathcal{L}_{\mathsf{InfoNCE}}^{(\mathsf{img}\to\mathsf{tex})}+\mathcal{L}_{\mathsf{InfoNCE}}^{(\mathsf{tex}\to\mathsf{img})}$ *with infinite training pairs.*

Theorem.5 and Corollary.6 mirror the insights of Theorem.2 and Corollary.3 that both recover the modal-invariant latent variable, $\boldsymbol{z}_{\mathsf{inv}}$, while the former do so under a token-aware SCM that assumes a textual description as a sequential composition process instead of a generated vector. This granular view provides the necessary foundation for our analysis. We will now use this framework to offer a principled explanation for CLIP's observed failures in compositional reasoning.

4 Composition Nonidentifibility in CLIP

As observed in existing research, CLIP is born vulnerably to identify the language compositional difference in an image-text pair. While such concrete definition could be shifted across specific literature. Our study focuses on the definition used to build CREPE (Ma et al. (2023)) and SUGARCREPE (Hsieh et al. (2023)): for an image-text pair $\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}\rangle$, they considered the tokenized word or phrase (*i.e.*, $X_{i,:}^{(\mathsf{tex})}$, a column of token-embedding matrix $X^{(\mathsf{tex})}$) as the *atomic concept* that represent a type of object (*i.e.*, OBJ), attribute (*i.e.*, ATT), or relation (*i.e.*, REL), then a hard negative textual description constructed from $X^{(\mathsf{tex})}$ can be categorized into three formats.

SWAP form. The hard negative **SWAP**$(X^{(\mathsf{tex})})$ is generated by exchanging two existing atomic concepts of the same type (object or attribute) within the text (*i.e.*, switching the column location between $X_{i,:}^{(\mathsf{tex})}$, $X_{j,:}^{(\mathsf{tex})}$, $\forall i\neq j$), without introducing anything new. Relationship swapping is omitted as it often produces nonsensical results, leaving the subcategories SWAP-OBJ and SWAP-ATT.

REPLACE form. The hard negative **REPLACE**$(X^{(\mathsf{tex})})$ is created by substituting a column $X_{i,:}^{(\mathsf{tex})}$ with regards to a single atomic concept (object, attribute, or relation) in the text $X^{(\mathsf{tex})}$ with a new-concept column (*i.e.*, $\mathsf{RF}(X_{:,j}^{(\mathsf{tex})})$ that denotes the "rephrased embedding" to this new atomic concept), which causes a mismatch with the visual scene. It literally can be subcategorized into REPLACE-OBJ, REPLACE-ATT, and REPLACE-REL according to the atomic concept type.

ADD form. The hard negative **ADD**$(X^{(\mathsf{tex})})$ is created by inserting a new atomic concept into the text (*i.e.*, adding a new-concept column $\mathsf{ADD}(X_{:,j}^{(\mathsf{tex})})$ into the position j) to create a mismatch with the scene. This is categorized as ADD-OBJ (adding an object) and ADD-ATT (adding an attribute); adding new relationships is avoided as it results in implausible text.

The aforementioned taxonomy of vision-language compositionality can summarize the cases in most other research using different definitions of vision-language compositionality.

Derived from the modal-invariant alignment in Theorem.5, we establish the theorems to question whether the vision-language compositionality can be achieved by **identifying the difference between an image's textual description and its hard negative in the recovered causal representation**, which are extracted from the pre-trained image and text encoders in CLIP (Eq.1). Specifically,

Theorem 7. ***(SWAP-form Composition Nonidentifibility)*** *Suppose image-text pairs generated by Assumption.4 with densities and mappings under the conditions in Theorem.5. If the optimal image encoder* f^* *and the optimal text encoder* g^* *satisfy Theorem.5 , thus*

$$\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f^*, g^*)\to 0 \tag{8}$$

with invertible functions h_{f^*} and h_{g^*} that fulfill $f^*{=}h_{f^*}\circ\mathbf{f}^{-1}_{1:n_{\mathsf{inv}}}$ and $g^*{=}h_{g^*}\circ\mathbf{g}^{-1}_{1:n_{\mathsf{inv}}}$, there exists a pseudo-optimal text encoder g^{**} derived from g^* that satisfy

$$\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*,g^{**})\to 0 \tag{9}$$

while if $g^{**}(X^{(\mathsf{tex})})$ equals to one of its column permutations, i.e.*, $\exists\pi(X^{(\mathsf{tex})})\in\Pi_k(\{1,\cdots,k\})$:*

$$g^{**}([X^{(\mathsf{tex})}_{:,1},X^{(\mathsf{tex})}_{:,2},\cdots,X^{(\mathsf{tex})}_{:,k}]) = \quad g^{**}([X^{(\mathsf{tex})}_{:,\pi(1)},X^{(\mathsf{tex})}_{:,\pi(2)},\cdots,X^{(\mathsf{tex})}_{:,\pi(k)}]), \tag{10}$$

it holds the SWAO-form hard negative $\boldsymbol{SWAP}(X^{(\mathsf{tex})})=\hat{\pi}(X^{(\mathsf{tex})})$ as the composition permuted by $\hat{\pi}$, so that $\forall\hat{\pi}(X^{(\mathsf{tex})})\in\Pi_k(\{1,\cdots,k\})\cap\big\{\{X^{(\mathsf{tex})}_{:,1},X^{(\mathsf{tex})}_{:,\pi(1)}\}\times\cdots\times\{X^{(\mathsf{tex})}_{:,k},X^{(\mathsf{tex})}_{:,\pi(k)}\}\big\}$,

$$g^{**}([X^{(\mathsf{tex})}_{:,1},X^{(\mathsf{tex})}_{:,2},\cdots,X^{(\mathsf{tex})}_{:,k}]) = \quad g^{**}([X^{(\mathsf{tex})}_{:,\hat{\pi}(1)},X^{(\mathsf{tex})}_{:,\hat{\pi}(2)},\cdots,X^{(\mathsf{tex})}_{:,\hat{\pi}(k)}]), \tag{11}$$

where $\Pi_k(\{1,\cdots,k\})$ indicates the set of arbitrary permutation orders of $\{1,\cdots,k\}$.

Theorem 8. ***(REPLACE-form Composition Nonidentifibility)*** *Given g^{**} defined by Theorem.7, if there is a token embedding $X^{(\mathsf{tex})}_{:,j}$ with its rephrase embedding $\mathsf{RF}(X^{(\mathsf{tex})}_{:,j})$ that satisfies*

$$g^{**}([X^{(\mathsf{tex})}_{:,1},\cdots,X^{(\mathsf{tex})}_{:,j},\cdots,X^{(\mathsf{tex})}_{:,k}]) = \quad g^{**}([X^{(\mathsf{tex})}_{:,\pi(1)},\cdots,\mathsf{RF}(X^{(\mathsf{tex})}_{:,j}),\cdots,X^{(\mathsf{tex})}_{:,\pi(k)}]), \tag{12}$$

with a column permutation $\pi(X^{(\mathsf{tex})})\in\Pi_{k-1}(\{1,\cdots,j{-}1,j{+}1,\cdots,k\})(j)$, it holds the REPLACE-form hard negative $\boldsymbol{REPLACE}(X^{(\mathsf{tex})})=\hat{\pi}(X^{(\mathsf{tex})})$ as the permutation with $\mathsf{RF}(X^{(\mathsf{tex})}_{:,j})$ that satisfy $\forall\hat{\pi}(X^{(\mathsf{tex})}_{:,-j})\in\Pi_{k-1}(\{1,\cdots,j-1,j+1,\cdots,k\})\bigcap\big\{\{X^{(\mathsf{tex})}_{:,1},X^{(\mathsf{tex})}_{:,\pi(1)}\}\times\cdots\{X^{(\mathsf{tex})}_{:,j-1},X^{(\mathsf{tex})}_{:,\pi(j-1)}\}$ $\times\{X^{(\mathsf{tex})}_{:,j+1},X^{(\mathsf{tex})}_{:,\pi(j+1)}\}\cdots\times\{X^{(\mathsf{tex})}_{:,k},X^{(\mathsf{tex})}_{:,\pi(k)}\}\big\}$ and $\forall\hat{X}^{(1)}_j,\hat{X}^{(2)}_j\in\{X^{(\mathsf{tex})}_{:,j},\mathsf{RF}(X^{(\mathsf{tex})}_{:,j})\}$,

$$g^{**}([X^{(\mathsf{tex})}_{:,1},\cdots,\hat{X}^{(1)}_j,\cdots,X^{(\mathsf{tex})}_{:,k}]) = \quad g^{**}([X^{(\mathsf{tex})}_{:,\hat{\pi}(1)},\cdots,\hat{X}^{(2)}_j,\cdots,X^{(\mathsf{tex})}_{:,\hat{\pi}(k)}]). \tag{13}$$

where $X^{(\mathsf{tex})}_{:,-j}$ indicates $X^{(\mathsf{tex})}$ without the j^{th} column.

Theorem 9. ***(ADD-form Composition Nonidentifibility)*** *Suppose image-text pairs generated by Assumption.4 with densities and mappings under the conditions in Theorem.5. If the optimal image encoder f^* and the optimal text encoder g^* satisfy Theorem.5 , thus*

$$\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*,g^*)\to 0 \tag{14}$$

with invertible functions h_{f^*} and h_{g^*} that fulfill $f^*{=}h_{f^*}\circ\mathbf{f}^{-1}_{1:n_{\mathsf{inv}}}$ and $g^*{=}h_{g^*}\circ\mathbf{g}^{-1}_{1:n_{\mathsf{inv}}}$, there exists a pseudo-optimal text encoder g^{**} derived from g^* that satisfy

$$\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*,g^{**})\to 0 \tag{15}$$

with the ADD-form hard negative $\boldsymbol{ADD}(X^{(\mathsf{tex})})=\hat{\pi}(X^{(\mathsf{tex})})$ as the permutation where $X^{(\mathsf{tex})}\in\mathcal{X}_{\mathsf{base}}$ and $\hat{\pi}(X^{(\mathsf{tex})})=([X^{(\mathsf{tex})}_{:,1},\cdots,X_j,\mathsf{ADD}(X^{(\mathsf{tex})}_{:,j}),\cdots,X^{(\mathsf{tex})}_{:,k}])\in\mathcal{X}_{\mathsf{ADD}}$, such that $\exists z^*_{\mathsf{inv}}\in\mathcal{C}_{\mathsf{inv}}$

$$z^*_{\mathsf{inv}}\in((g^*)^{(j)})^{-1}_{1:n_{\mathsf{inv}}}(\mathcal{X}_{\mathsf{base}})\ \cap\ ((g^*)^{(j+1)})^{-1}_{1:n_{\mathsf{inv}}}(\mathcal{X}_{\mathsf{ADD}}),$$

then it holds

$$g^{**}([X^{(\mathsf{tex})}_{:,1},\cdots,X^{(\mathsf{tex})}_{:,k}]) = g^{**}([X^{(\mathsf{tex})}_{:,1},\cdots,X^{(\mathsf{tex})}_{:,j},\mathsf{ADD}(X^{(\mathsf{tex})}_{:,j}),\cdots,X^{(\mathsf{tex})}_{:,k}]). \tag{16}$$

Interpretation. The statements and proof sketches in Theorems.7, 8, and 9 resemble the spirit of using Theorem. and Corollary.6 to construct a "pseudo-optimal" text encoder g^{**} that occur when the "true-optimal" text encoder g^* could be practically obtained by the causal representation of CLIP. In this situation, g^* and g^{**} can simultaneously achieve the modal-invariant alignment $\Big($*i.e.*, $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*,g^*)\simeq 0$ and $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*,g^{**})\simeq 0\Big)$ with the optimal image encoder f^* during pre-training. **Nevertheless, distinct from g^* that could perfectly distinguish arbitrary permutations from a text $X^{(\mathsf{tex})}$, g^{**} fails to identify some token sequences re-permuted from the columns of $X^{(\mathsf{tex})}$, according to the compositional rules in Theorem.7-9.** Since the encoders g^* and g^{**} share the same architecture and their parameters both achieve modal-invariant alignment during

Table 1: The correspondence between our theorems and the taxonomy of vision-language composition reasoning types. NEG and QUA denote negations and quantifiers.

	Atomic concepts	$X^{(\text{tex})}$	Pre-condition	Hard negative
Thm.7 (SWAP-form Composition Nonidentifibility)	OBJ,ATT	"*a white cat and a black dog* play"	"*a black dog and a white cat play*"	"*a white dog and a black cat play*"
Thm.8 (REPLACE-form Composition Nonidentifibility)	OBJ,ATT,REL,QUA	"*a horse* **on** *the grass*"	"*the grass* ***under*** *a horse*"	"*the grass* ***on*** *a horse*"
Thm.9 (ADD-form Composition Nonidentifibility)	OBJ,ATT,NEG,QUA	"*flowers*"	$g^*(X^{(\text{tex})})$=$g^*(\text{ADD}(X^{(\text{tex})}))$	"*no flowers*"

pre-training, **there are no evidences and solutions to identify which one in g^*, g^{**} would be learned in practice.**

It is noteworthy that Theorems.7-9 are **grammar-agnostic** so can flexibly transfer across a broad range of language as long as they can convey the consistent semantic. Besides, they are motivated by the "SWAP-REPLACE-ADD" taxonomy that covers the most cases of vision-language compositionality in other research with different definitions. To better understand the non-identified textual-token compositions in Theorem.7-9, we illustrated some instances with regards to embedding their language tokens by g^{**} in Table.1.

Extension to the hardness of vision compositionality. Theorems.7-9 are derived from the composition operators to describe the hardness in the language level, whereas the existing study argue that the hardness also happen to misunderstanding the visual concepts presented in images. Since the natural image generation process significantly differs from language in Assumption.4, it is impossible to derive the same causal analysis to explain the vision compositionality.

Instead, we resort to the constant modality gap phenomenon. Specifically, (Zhang et al.) observed that relevant image-text pairs extracted by CLIP's image and text encoders, show the consistent distance between their features. (Chen et al. (2023)) extend their results to justify that CLIP may not isolate two images when they share some mutually exclusive atomic concepts. It is obvious that when an image with its counterpart regenerated by modifying some atomic concepts via SWAP, REPLACE, or ADD forms, it definitely leads to the appearance of mutually exclusive atomic concepts between them. It explains the hardness of vision compositionality using CLIP.

The nonidentifiability with multiple atomic concepts. The hard negative in Theorem.7-9 focus on the text instances $X^{(\text{tex})}$ derived from after the modification with a single atomic concept. We now demonstrate that their can be combined and extend to the nonidentified image-text matching involved with multi-concept modification. In specific, given an image $x^{(\text{img})}$ and its hard negative description of $F(X^{(\text{tex})})$ ($F_1(\cdot)$ = **SWAP**$(\cdot)$,**REPLACE**$(\cdot)$, or **ADD**$(\cdot)$) using Theorem.7-9, we know the existence of <f^*, g^{**}> to generate the nonidentified image-text matching. For the image and its modified hard negative, <f^*, g^{**}> has no difference with <f^*, g^*>. To this, we may consider the second hard negative description $F_2(F_1(X^{(\text{tex})}))$ generated from $F_1(X^{(\text{tex})})$ ($F_2(\cdot)$ = **SWAP**$(\cdot)$,**REPLACE**$(\cdot)$, or **ADD**$(\cdot)$) using Theorem.7-9 on another atomic concept, and there must be some pseudo encoder pairs <f^*, g^{***}> with regards to <f^*, g^{**}> (*i.e.*, <f^*, g^{**}> was treated as the true encoder pairs since <$f^*(x^{(\text{img})})$, $g^*(X^{(\text{tex})})$> and <$f^*(x^{(\text{img})})$, $g^{**}(X^{(\text{tex})})$> in terms of our theorems).

In other words, it is possible to generate more complex hard-negative textual instances by stacking the compound nonidentified matching effects through iteratively using **SWAP**$(\cdot)$,**REPLACE**$(\cdot)$, or **ADD**$(\cdot)$. While the process can not be endless because each calling of **SWAP**$(\cdot)$, **REPLACE**$(\cdot)$, or **ADD**$(\cdot)$ will reduce the solution space of the hard negative derived from $X^{(\text{tex})}$. In practice, we found that the second calling is sufficient to generate more confusing hard negative cases of $X^{(\text{tex})}$.

5 Experiments

In this section, we provide some empirical studies to verify our theoretical results from three aspects. **First**, we attempt to verify whether Theorem.7-9 could be used to generate the practical hard negative instances covered by the existing vision-language compositional reasoning benchmarks, so that it literally suits the reality; **Second**, we aim to justify the existence of "pseudo-optimal" text encoders

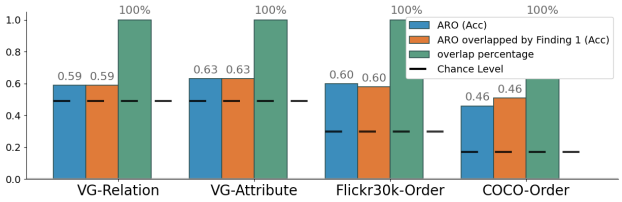


Figure 3: CLIP's accuracy (ACC) on the negative samples generated by ARO and our Algorithm1. The overlap percentage indicates how many negative samples in ARO belong to the cases in Theorem.7-9.

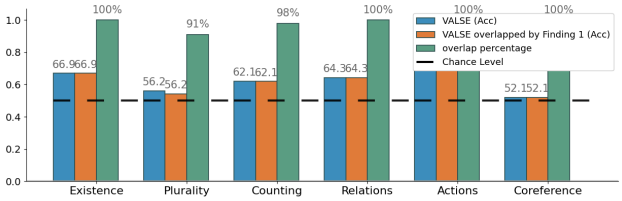


Figure 4: CLIP's accuracy (ACC) on the negative samples generated by VALSE and our Algorithm1. The percentage indicates how many negative samples in VALSE belong to the cases in Theorem.7-9.

induced by Theorem.7-9. **Finally**, we provide the experiments of CLIP-based models trained and evaluated with regular hard negative pairs and hard negative pairs generated by the second calling to **SWAP**$(\cdot)$,**REPLACE**$(\cdot)$, or **ADD**$(\cdot)$, which generate the more complex non-identified cases in the textual descriptions. The implementation of composition operators **SWAP**$(\cdot)$, **REPLACE**$(\cdot)$, and **ADD**$(\cdot)$ with respect to Theorems.7-9 are summarized by Algorithm.1 in Appendix. We apply Gemini 2.5 Pro as the proxy for their executions.

5.1 Bridging Theoretical-Empirical Gaps on Benchmark Data

To justify whether the theoretical results suit the practice, we conduct our compositional understanding experiment in ARO (Yuksekgonul et al. (2023)) that consists of four splits for evaluation: VG-Relation, VG-Attribution, COCO-Order, and Flickr30k-Order. We access their test splits then select the instances which belongs to the compositional reasoning cases described by Theorem 7-9. Besides, we also consider VALSE benchmark Parcalabescu et al. (2021) where the composition reasoning instances derived from five sources including MSCOCO, Visual7W, SWiG VisDial v1.0, SituNet are categorized into six cases, *i.e.*, *existence, plurality, counting, relations, actions, coreference*. Given this, we conduct the CLIP evaluation on the four test splits in ARO and six test splits in VALSE, where LLM-as-a-Judge strategy is employed to justify whether test instances can be categorized into the hard negative cases generated by our theorems, then report their percentages.

Fig.3,4 substantiate our core motivation: the proposed token-aware algorithms, instantiated from the SWAP/REPLACE/ADD theorems, can replicate a large fraction of the hard negative instances used by existing benchmarks. On ARO (Fig. 3) and VALSE (Fig. 4), the "overlap percentage" bars are high across splits, indicating that many benchmark negatives fall within the transformations our procedures generate. This alignment is not superficial: CLIP's accuracies on these subsets mirror the original benchmark trends, showing that our synthesized negatives preserve difficulty while being produced by a transparent, theoretically grounded process. Moreover, cases where accuracy on overlapped subsets matches the benchmark values reveal that pseudo-optimal text encoders remain insensitive to token permutations or rephrasings precisely as predicted. Together, these results demonstrate that our framework not only explains why CLIP fails on compositional variants, but also operationalizes this insight into practical data generation that faithfully reproduces real benchmark hard negatives—closing the theory-to-benchmark gap.

5.2 Evidences of g^{**}'s Existence

Theorems.7-9 demonstrate that we can not directly judge the existence of the pseudo-optimal text encoder g^{**}. Whereas some evidences are possibly observed if g^{**} is created. Specifically, we would like to observe the discrepancies between the features of $X^{(\text{text})}$ and its

Table 2: Results on CC3M and CC12M across Replace, Swap, and Add categories. Bold indicates the best in each column.

Methods	Replace			Swap		Add		Overall
	Object	Attribute	Relation	Object	Attribute	Object	Attribute	Avg.
CC3M								
NegCLIP	62.71	58.12	54.48	**56.33**	51.20	56.21	56.13	57.18
NegCLIP (+MC)	63.11	63.24	60.79	57.18	53.65	58.31	59.45	**59.02**
TripletCLIP	**69.92**	**69.03**	64.72	56.33	**57.96**	**62.61**	63.87	63.49
TripletCLIP (+MC)	**71.00**	**70.31**	63.22	55.93	58.67	**63.21**	64.90	**64.79**
CC12M								
NegCLIP	77.84	69.29	63.23	**66.53**	62.31	68.17	69.65	68.00
NegCLIP (+MC)	78.18	70.91	62.93	68.73	63.38	69.70	69.75	**68.87**
TripletCLIP	**83.66**	**81.22**	**79.02**	64.49	63.66	**73.67**	**75.43**	74.45
TripletCLIP (+MC)	84.86	80.02	79.82	67.52	64.55	72.67	76.43	**76.51**

hard negative counterparts as **SWAP**($X^{(\text{text})}$),**REPLACE**($X^{(\text{text})}$), or **ADD**($X^{(\text{text})}$), respectively. We employ $\mathcal{A}$-distances between the features of test instances drawn from SugarCREPE <$X^{(\text{text})}$,**SWAP**($X^{(\text{text})}$)>;<$X^{(\text{text})}$,**REPLACE**($X^{(\text{text})}$)>; <$X^{(\text{text})}$,**ADD**($X^{(\text{text})}$)>. We particularly consider the change before training with / without the hard negative generated by **SWAP**, **REPLACE**, and **ADD**. The results are presented as

- <$X^{(\text{text})}$,**SWAP**($X^{(\text{text})}$)>. with-1.91 , without-1.06.
- <$X^{(\text{text})}$,**REPLACE**($X^{(\text{text})}$)>. with-1.86 , without-0.98.
- <$X^{(\text{text})}$,**ADD**($X^{(\text{text})}$)>. with-1.84 , without-1.01.

With regards to the characteristic of $\mathcal{A}$ distance, we found that the generated hard negatives almost hold the same statistical evidences without post-training with hard negative, whereas hard negative can effectively isolate them. It implies the existence of g^{**}.

5.3 Multi-Calling of Composition Operators

In the last experiment, we are interested to observe whether iterative calling of composition operators **SWAP**($\cdot$),**REPLACE**($\cdot$), or **ADD**($\cdot$) to modify the text from the original description to hard negative, can lead to more challenging hard negative pairs. Specifically, we conduct the experiments on the benchmark with two train-test splits, *i.e.*, CC3M and CC12M. The evaluated baselines NegCLIP (Yuksekgonul et al. (2023)) and TripleCLIP (Patel et al. (2024)) both employed hard negative mining to augment their training paradigms. We accordingly use Algorithm.1 to generate hard negative to further augment the training instances, leading to our baselines NegCLIP (+MC) and TripleCLIP (+MC) to justify whether iterative-generated hard negative can further improve their performances.

Table 2 shows that iteratively applying SWAP/REPLACE/ADD during training yields consistent gains over their hard-negative baselines. On CC3M, NegCLIP(+MC) improves the Overall Avg. from 57.18 to 59.02 (+1.84), and TripletCLIP(+MC) from 63.49 to 64.79 (+1.30). The strongest per-type gains appear in Replace (e.g., CC3M Attribute: 69.03 $\rightarrow$ 70.31; CC12M Object: 83.66 $\rightarrow$ 84.86), aligning with our claim that stacking operators expands the difficult negative space beyond single edits. On CC12M, where base performance is higher, MC still adds +0.87 for NegCLIP and +2.06 for TripletCLIP, with notable boosts on Swap-Object (64.49 $\rightarrow$ 67.52) and Add-Attribute (75.43 $\rightarrow$ 76.43). Not all cells increase (e.g., CC3M Replace-Relation slightly drops for TripletCLIP), suggesting diminishing returns or coverage imbalance for certain relations. Overall, MC systematically enhances robustness across datasets and edit types, validating our hypothesis that compound compositional perturbations generate harder, complementary negatives that translate into better compositional generalization.

REFERENCES

Ziliang Chen, Xin Huang, Quanlong Guan, Liang Lin, and Weiqi Luo. A retrospect to multi-prompt learning across vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22190–22201, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

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

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

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

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

Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, et al. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing Systems*, 34:16158–16170, 2021.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

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

Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *The Eleventh International Conference on Learning Representations*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.