ON UNIDENTIFIED REPRESENTATION OPTIMALITY OF VISION-LANGUAGE COMPOSITIONAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Language-Image Pre-training (CLIP) is a cornerstone of multimodal representation learning, while it is born with a persistent inability to reason about compositional relations. What's worse, a deficiency can be inherited by the Multimodal Large Language Models (MLLMs) that adopt its image encoder as their visual receiver. Why such failures arise has remained without a rigorous account. We develop a token-aware causal representation learning framework built on sequential, language-token Structural Causal Models (SCMs), recasting block identifiability for tokenized text and proving that CLIP's contrastive objective recovers the modality-invariant latent variables at the token level. It yields a principled identifiability-based diagnosis, *i.e.*, compositionality is *nonidentifiable*: *pseudo-optimal* text encoders attain perfect modality-invariant alignment while remaining provably invariant to the SWAP, REPLACE and ADD operators over atomic concepts, so no contrastive optimum guarantees separation between a caption and its hard negatives. Through the modality gap, the problem propagates to the image encoder, which leads to an *algorithm-independent* impossibility result: no projector, LLM capacity, or instruction tuning can lift a downstream MLLM above chance on the affected queries. Guided by our theoretical recipe, we propose the remedy of hard negative strategies measurably repair both CLIP and MLLMs.

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 generalization and zero-shot inference capabilities using textual prompts. From the lens of causal representation learning 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 reasoning compositional relations across images and text. It often acts like a bag-of-words matcher Yuksekgonul et al. (2023); Ma et al. (2023); Hsieh et al. (2023), identifying atomic concepts individually but failing to bind them to their specified orders, attributes, or relationships. These failures reveal that its embedding space unreliably encodes the compositional structure required for precise, human-like understanding in vision-language tasks. This phenomenon is not confined to CLIP alone: Multimodal Large Language Models (MLLMs) such as LLaVA Liu et al. (2023; 2024), InstructBLIP Dai et al. (2023), and Qwen-VL Bai et al. (2023) that adopt CLIP's image encoder as their visual backbone inherit the same compositional blindness Tong et al. (2024), exhibiting systematic failures on elementary spatial, attribute-binding, and counting queries even when equipped with powerful language reasoning.

Although considerable benchmarks and empirical remedies have been proposed Hsieh et al. (2023); Patel et al. (2024), a rigorous theoretical explanation for ***why these failures arise—and why they propagate across the entire multimodal pipeline—remains elusive***. Much of the existing theoretical work on CLIP simplifies texts as monolithic, fixed-length vectors, thereby overlooking the compositional structure of atomic concepts that presents as tokens at the heart of the issue.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our research aims to provide the ***first principled, end-to-end theoretical recipe*** to elaborate failure roots of compositional reasoning spanning from CLIP's text encoder, through its image encoder, to downstream MLLMs. Our theoretical contributions are four-fold:

(I) Token-aware causal representation framework (Section 3). We propose a token-aware Structural Causal Model (SCM) of image–text generation that grounds atomic concepts in text-side tokens, and prove that CLIP's contrastive objective recovers the modality-invariant latent variables at the token level (Theorem 1, Corollary 2)—the mathematical foundation for all diagnoses below.

(II) Composition nonidentifiability of text encoders (Section 4.1). We certify "pseudo-optimal" text encoders that attain *perfect* modality-invariant alignment yet stay blind to the compositions used in prevalent benchmarks Ma et al. (2023); Hsieh et al. (2023) (Theorems 3–5): without extra data intervention, the objective itself cannot single out the truly optimal encoder.

(III) Propagation to visual encoders (Section 4.2). Via the constant modality gap Liang et al. (2022), text-side nonidentifiability provably forces the aligned image encoder to collapse images differing only in compositional structure (Proposition 8), with compound operations yielding progressively larger equivalence classes (Corollary 9).

(IV) From bottleneck to impossibility in MLLMs (Section 4.3). Since MLLMs consume exactly these collapsed visual tokens, we lift the failure from *one* encoder to *every* pipeline: a twin SCM that is observationally indistinguishable from the true one yet flips every compositional answer shows that *no* learning algorithm—any projector, any LLM capacity, any amount of tuning data—can beat chance on such queries (Theorem 10). To our knowledge this is the first *algorithm-independent* impossibility result for compositional reasoning; it is tight at chance level, explains the at-chance MLLM behavior on MMVP Tong et al. (2024), and pinpoints intervention beyond the observational distribution as the *only* remedy.

Guided by these, we propose algorithms that reconfigure hard negative mining strategies by iteratively stacking SWAP, REPLACE and ADD to reach compound perturbations unreachable by single edits. Experiments validate the methods: compound negatives yield consistent gains on SugarCREPE Hsieh et al. (2023), preserve zero-shot accuracy on VTAB and ImageNet-1k, and swapping in our encoder substantially recovers the compositional reasoning of CLIP-based MLLMs.

2 PRELIMINARIES

In this section, we briefly go through Contrastive Language-Image Pre-training (CLIP) and its theory derived from causal representation . More detailed background and notations refer to Appendix.A.

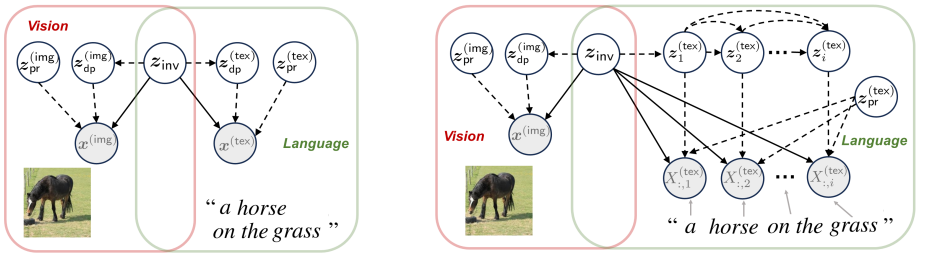
CLIP and Its Encoders. The CLIP family Radford et al. (2021); Jia et al. (2021); Cherti et al. (2023) receives data coupled by image and text in mutual semantic through contrastive pre-training Oord et al. (2018); He et al. (2020). Suppose $\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle \sim p_{\mathsf{mm}}\big(\boldsymbol{x}^{(\mathsf{img})}, \boldsymbol{x}^{(\mathsf{tex})}\big)$ 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

$$\min_{f,g} \mathbb{E}_{\mathcal{D}^{(K)} \sim p_{\mathsf{mm}}} \Big[\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) + \mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) \Big]$$

$$\text{s.t. } \mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}} = \sum_{i=1}^{K} -\log \frac{e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{j=1}^{K} e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_j^{(\mathsf{tex})})/\gamma\right)}},\ \mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}} = \sum_{i=1}^{K} -\log \frac{e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{j=1}^{K} e^{\left(f(x_j^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}} \tag{1}$$

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}.

Causal Representation Identified by CLIP. Under p_{mm} interpreted as the generative process defined 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 $\boldsymbol{z}^{(\mathsf{inv})}$ through a causal lens. More specifically, when image-text pairs are generated from some multimodal data distribution defined by the SCM Assumption.11 Daunhawer et al. (2022); Yao et al. (2023), the causal representation learned by CLIP yields the optimal representation alignment be-



(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 **(a)**.the sentence-level aspect (Assumption .11) and **(b)**.the token-level aspect (Assumption .12). CLIP seeks for the recovery of the modal-shared latent variable $\boldsymbol{z}_{\mathsf{inv}}$ via token-aware causal representation (Theorem.1), the tool to unveil the hardness of compositional reasoning in VLMs.

tween image and text encoders (f^*, g^*) under the principles of block identifiability (Von Kügelgen et al. (2021)) (see Them.9).While paradigm treats the entire text as a monolithic vector so that cannot reveal ***how the compositional arrangement of individual atomic concepts (represented by tokens) contribute to the learned alignment***. In the next section, we reframe the block identifiability via newly-developed token-aware causal representation. It becomes the main theoretical tool to justify the failures of reasoning compositionality by massive existing VLMs in the rest of our paper.

3 Token-Aware Causal Representation: A Theoretical Tool of Interpreting Compositional Reasoning Hardness

As discussed, Theorem.9 can be extended to interpret the hardness of vision-language compositionality: we reconsider the more practical setup with the nonparametric functions that indicates the text from a vector $x^{(\mathsf{tex})} \sim p_{\boldsymbol{x}^{(\mathsf{tex})}}$ to a k-column matrix $X^{(\mathsf{tex},\mathsf{k})} \sim p_{\mathbf{X}^{(\mathsf{tex},\mathsf{k})}}$, where $\forall \mathsf{k} \in \{1, \cdots, \mathsf{k}_{\max}\}$ indicates the sentence length and the ith column $X^{(\mathsf{tex},\mathsf{k})}_{:,i}$ indicates the ith token embedding. It promotes Assumption.12 extends the image-language SCM definition in Assumption.11 (Fig.1) by drawing the inspiration from the memory-argumented Bayesian LLM prior Wei et al. (2021). Derived from the token-level understanding to p_{mm}, we redevelop Them.9 from the sentence level to the token level:

Theorem 1. ***(Block-Identified Modal-invariant Alignment (Token-aware))*** *Consider the image-text pairs generated by Assumption.12 . If their densities and mappings meet:* ***1).*** $\mathbf{f}$ *is a diffeomorphism and* $\forall i \in \{1, \ldots, k_{\max}\}$, $\mathbf{g}_i : \mathcal{M}_i \to \mathbf{g}_i(\mathcal{M}_i)$ *is also a diffeomorphism.;* ***2).*** $\boldsymbol{z}^{(\mathsf{img})}$, $\boldsymbol{z}_i^{(\mathsf{tex})}$ $(\forall i \in \{1, \cdots, \mathsf{k}_{\mathsf{max}}\})$ *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* $g : \cup_i^{\mathsf{k}_{\mathsf{max}}} \mathcal{X}_{\mathsf{tex}}^{(i)} \to (0,1)^{n_{\mathsf{inv}}}$ *as smooth functions that are trained to jointly minimize the functionals,*

$$\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}} := \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}}} \left[||f(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})})|| \right] - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{X}^{(\mathsf{tex})})), \tag{2}$$

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:*

$$f^* = h_f \circ \mathbf{f}^{-1}_{1:n_{\mathsf{inv}}}, \; g^* = h_g \circ \mathbf{g}^{-1}_{1:n_{\mathsf{inv}}} \tag{3}$$

Corollary 2. *(Informal) The optimal encoders* f^*, g^* *in Theorem.1 are almost equivalent with* $(f^*, g^*) = \arg\min_{f,g} \mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}} + \mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}$ *with infinite training pairs.*

Theorem.1 and Corollary.2 mirror the insights of Theorem.9 and Corollary.10, which 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 rather than a generated vector. This granular view provides the causal-theoretic foundation to interpret VLMs' failures in reasoning compositionality, due to the pseudo-optimal text and image encoders that we certified next.

4 COMPOSITION NONIDENTIFIBILITY BEHIND VISION LANGUAGE MODELS

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

ADD form. The hard negative $\mathbf{ADD}(X^{(\mathsf{tex})})$ is created by inserting a new atomic concept into the text (*i.e.*, adding a new-concept column $\mathsf{ADD}(X^{(\mathsf{tex})}_{:,j})$ 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.

Having formalized the three canonical compositional operators that transform a caption into its hard negative, we are now equipped to ask the central theoretical question: ***does CLIP's contrastive objective guarantee that its text encoder can distinguish*** $X^{(\mathrm{tex})}$ ***from*** $\mathrm{SWAP}(X^{(\mathrm{tex})})$***,*** $\mathrm{REPLACE}(X^{(\mathrm{tex})})$***, or*** $\mathrm{ADD}(X^{(\mathrm{tex})})$***?*** The following theorems demonstrate the negative answer.

4.1 REASONING COMPOSITIONALITY UNIDENTIFIED BY TEXT ENCODERS OF CLIP

Derived from Theorem.1, we first establish the theoretical framework 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***. It consists of three theorems with regards to the well-trained image and text encoders f^*, g^* in CLIP (Eq.1). Specifically,

Theorem 3. ***(SWAP-form Language Composition Nonidentifibility)*** *Suppose image-text pairs generated by Assumption.12 with densities and mappings under the conditions in Theorem.1. If the optimal image encoder f^* and the optimal text encoder g^* satisfy Theorem.1 , $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*, g^*) \to 0$ 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 $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*, g^{**}) \to 0$, while if $g^{**}(X^{(\mathsf{tex})})$ equals to one of its column permutations,* i.e., *$\exists \pi(X^{(\mathsf{tex})}) \in \Pi_k(\{1, \cdots, k\})$:*

$$g^{**}([X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,2}, \cdots, X^{(\mathsf{tex})}_{:,k}]) = \quad g^{**}([X^{(\mathsf{tex})}_{:,\pi(1)}, X^{(\mathsf{tex})}_{:,\pi(2)}, \cdots, X^{(\mathsf{tex})}_{:,\pi(k)}]), \tag{4}$$

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

$$g^{**}([X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,2}, \cdots, X^{(\mathsf{tex})}_{:,k}]) = \quad g^{**}([X^{(\mathsf{tex})}_{:,\hat{\pi}(1)}, X^{(\mathsf{tex})}_{:,\hat{\pi}(2)}, \cdots, X^{(\mathsf{tex})}_{:,\hat{\pi}(k)}]), \tag{5}$$

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

Theorem 4. ***(REPLACE-form Language Composition Nonidentifibility)*** *Given g^{**} defined by Theorem.3, if there is a token embedding $X^{(\mathsf{tex})}_{:,j}$ with its rephrase embedding $\mathsf{RP}(X^{(\mathsf{tex})}_{:,j})$ that satisfies*

$$g^{**}([X^{(\mathsf{tex})}_{:,1}, \cdots, X^{(\mathsf{tex})}_{:,j}, \cdots, X^{(\mathsf{tex})}_{:,k}]) = \quad g^{**}([X^{(\mathsf{tex})}_{:,\pi(1)}, \cdots, \mathsf{RP}(X^{(\mathsf{tex})}_{:,j}), \cdots, X^{(\mathsf{tex})}_{:,\pi(k)}]), \tag{6}$$

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.3 (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.4 (REPLACE-form Composition Nonidentifibility)	OBJ,ATT,REL,QUA	"*a horse* **on** *the grass*"	"*the grass* ***under*** *a horse*"	"*the grass* ***on*** *a horse*"
Thm.5 (ADD-form Composition Nonidentifibility)	OBJ,ATT,NEG,QUA	"*flowers*"	$g^*(X^{(\text{tex})})=g^*(\text{ADD}(X^{(\text{tex})}))$	"*no flowers*"

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

$$g^{**}([X^{(\text{tex})}_{:,1},\cdots,\hat{X}^{(1)}_j,\cdots,X^{(\text{tex})}_{:,k}]) = g^{**}([X^{(\text{tex})}_{:,\hat{\pi}(1)},\cdots,\hat{X}^{(2)}_j,\cdots,X^{(\text{tex})}_{:,\hat{\pi}(k)}]). \tag{7}$$

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

Theorem 5. ***(ADD-form Language Composition Nonidentifibility)*** *Suppose image-text pairs generated by Assumption.12 with densities and mappings under the conditions in Theorem.1. If the optimal image encoder* f^* *and the optimal text encoder* g^* *satisfy Theorem.1 ,* $\mathcal{L}^{(\text{img,tex})}_{\text{MMAlign}}(f^*,g^*)\to 0$ *with invertible functions* h_{f^*} *and* h_{g^*} *that satisfy* $f^*=h_{f^*}\circ\mathbf{f}^{-1}_{1:n_{\text{inv}}}$ *and* $g^*=h_{g^*}\circ\mathbf{g}^{-1}_{1:n_{\text{inv}}}$*, there exists a pseudo-optimal text encoder* g^{**} *derived from* g^* *that satisfy*$\mathcal{L}^{(\text{img,tex})}_{\text{MMAlign}}(f^*,g^{**})\to 0$ *with the ADD-form hard negative* $\textbf{\textit{ADD}}(X^{(\text{tex})}) = \hat{\pi}(X^{(\text{tex})})$ *as the permutation where* $X^{(\text{tex})}\in\mathcal{X}_{\text{base}}$ *and* $\hat{\pi}(X^{(\text{tex})}) = ([X^{(\text{tex})}_{:,1},\cdots,X_j,\text{ADD}(X^{(\text{tex})}_{:,j}),\cdots,X^{(\text{tex})}_{:,k}])\in\mathcal{X}_{\text{ADD}}$*, such that* $\exists z^*_{\text{inv}}\in\mathcal{C}_{\text{inv}}$

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

then it holds

$$g^{**}([X^{(\text{tex})}_{:,1},\cdots,X^{(\text{tex})}_{:,k}]) = g^{**}([X^{(\text{tex})}_{:,1},\cdots,X^{(\text{tex})}_{:,j},\text{ADD}(X^{(\text{tex})}_{:,j}),\cdots,X^{(\text{tex})}_{:,k}]). \tag{8}$$

Interpretation. The statements in Theorems.3, 4, and 5 resemble the spirit of using Theorem.1 and Corollary.2 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.*, $\mathcal{L}^{(\text{img,tex})}_{\text{MMAlign}}(f^*,g^*)\simeq 0$ and $\mathcal{L}^{(\text{img,tex})}_{\text{MMAlign}}(f^*,g^{**})\simeq 0\Big)$ with the optimal image encoder f^* during pre-training. ***Nevertheless, distinct from*** g^* ***that could perfectly distinguish arbitrary permutations from a text*** $X^{(\text{tex})}$***,*** g^{**} ***fails to identify some token sequences re-permuted from the columns of*** $X^{(\text{tex})}$***, according to the compositional rules in Theorem.3-5.*** Since the encoders g^* and g^{**} share the same architecture and their parameters both achieve modal-invariant alignment during pre-training, ***there are no evidences and solutions to identify which one in*** g^****,*** g^{**} ***would be learned in practice.***

To clarify the meaning, we specify the statements in Theorem.3-5 with regards to **SWAP**, **REPLACE**, **ADD** that transfer the original text into the hard negative version in Appendix.A and Table.1.

Expressivity of cross-linguistic and compound-semantic cases. It is noteworthy that Theorems.3-5 are ***grammar-agnostic*** so can flexibly transfer across a broad range of language as long as they can convey the consistent semantic. Beyond this, the effects of Theorem.3-5 are also ***composable***: since each operator modifies a single atomic concept while preserving the alignment optimality of the resulting pseudo-optimal encoder, one may iteratively apply $\text{SWAP}(\cdot)$, $\text{REPLACE}(\cdot)$, or $\text{ADD}(\cdot)$ to distinct atomic concepts within the same caption, yielding compound hard negatives $F_2(F_1(\boldsymbol{X}^{(\text{tex})}))$ that involve multi-concept modifications. It almost encompasses all failure cases of isolating language composition in existing CLIP research.

4.2 Reasoning Compositionality Unidentified by Image Encoders of CLIP

Theorems 3–5 establish the composition nonidentifiability on the text-encoder side, certifying the existence of pseudo-optimal g^{**} provably blind to SWAP, REPLACE, and ADD operators. A mirror question arises: ***does this text-side failure propagate to the image encoder?*** Our answer is ***Yes***.

Specifically, the existing study Zhang et al. provide the evidences that in CLIP's learned embedding space, the difference between image and text embeddings for matched pairs approximates a constant vector that is orthogonal to the span of embeddings from either modality. It is validated across contrastive VLMs (CLIP, ConVIRT Zhang et al. (2022), VideoCLIP, etc), which we formalize as the ***Constant Modality Gap*** premise (Assumption.13). Given this characterization, the result is hold:

Proposition 6. ***(Visual Composition Nonidentifiability)*** *Suppose image-text pairs are generated by Assumption 12 with densities and mappings satisfying the conditions in Theorem 9. Let g^{**} be the pseudo-optimal text encoder from Theorem 3 (analogously Theorems 4, 5) satisfying $\mathcal{L}^{(\mathrm{img,tex})}_{\mathrm{MMAlign}}(f^*, g^{**}) \to 0$ with*

$$g^{**}(\boldsymbol{X}^{(\mathrm{tex})}) = g^{**}(\mathrm{OP}(\boldsymbol{X}^{(\mathrm{tex})})), \quad \mathrm{OP} \in \{\textsc{Swap}, \textsc{Replace}, \textsc{Add}\}. \tag{9}$$

With the constant modality gap (Assumption 13), for any two images $\boldsymbol{x}^{(\mathrm{img})}$ and $\boldsymbol{x}^{(\mathrm{img})}_{\mathrm{op}}$ whose ground-truth textual descriptions are $\boldsymbol{X}^{(\mathrm{tex})}$ and $\mathrm{OP}(\boldsymbol{X}^{(\mathrm{tex})})$ respectively, the image encoder f^* paired with g^{**} satisfies:

$$\|f^*(\boldsymbol{x}^{(\mathrm{img})}) - f^*(\boldsymbol{x}^{(\mathrm{img})}_{\mathrm{op}})\| = 0. \tag{10}$$

That is, the image encoder inherits the composition nonidentifiability from the text side: it provably collapses visually distinct images that differ only in their compositional structure.

Interpretation. Proposition 6 shows that text-side compositional nonidentifiability propagates to the visual modality: since the modality gap $\boldsymbol{\delta}$ acts as a semantically inert translation, whenever g^{**} conflates two compositionally distinct captions, the aligned image encoder f^* is forced to conflate the corresponding images regardless of their visual dissimilarity. By iteratively applying Proposition 6, diverse confused image pairs can be presented as a result of stacked multi-concept nonidentifiability:

Corollary 7. ***(Compound Visual Composition Nonidentifiability)*** *Let $F_2(F_1(\boldsymbol{X}^{(\mathrm{tex})}))$ be a compound hard negative obtained by sequentially applying two atomic-concept modifications as characterized by Theorems 3–5. If the pseudo-optimal text encoder g^{***} satisfies $g^{***}(F_2(F_1(\boldsymbol{X}^{(\mathrm{tex})}))) = g^{**}(F_1(\boldsymbol{X}^{(\mathrm{tex})})) = g^*(\boldsymbol{X}^{(\mathrm{tex})})$, then the aligned image encoder f^* conflates the images corresponding to all three captions, i.e., $f^*(x^{(\mathrm{img})}) = f^*(x^{(\mathrm{img})}_{F_1}) = f^*(x^{(\mathrm{img})}_{F_2 \circ F_1})$.*

4.3 Downstream Impossibility behind Compositional Reasoning in MLLMs

Since MLLMs consume exactly the visual tokens produced by the image encoders analyzed previously, the equivalence classes of collapsed images delimit what any downstream language model is able to resolve on compositional queries. We show that it can be propagated into a downstream large language model via a cross-modal projector, causing the risk of MLLM in compositional reasoning.

A formalized review of MLLM. We abstract the entire pipeline—encoder pre-training, projector, LLM decoding, and any fine-tuning—into a single measurable map $\mathcal{A} : (\mathcal{X} \times \mathcal{T})^n \times \mathcal{Q} \times [0,1] \to \{0,1\}$ of the training data D, the op-query q, and an internal randomness seed $R \sim \mathrm{Unif}[0,1]$, which imposes no assumption whatsoever on architectures or objectives. The asymmetry we exploit is that an op-query is a pair of *observed* inputs (as in the MMVP protocol), whereas its ground-truth label is issued by a semantic predicate $\varphi : \mathcal{Z} \to \{0,1\}$ acting on the *latent* variable; consequently, taking a permutation $\pi \in \Pi^*$ that is non-trivial on the concept slots bound by the operator (guaranteed by Theorems 3-5) and pushing it through the generative maps produces a twin SCM that is observationally *indistinguishable* from the true one (Lemma 18) yet assigns the *opposite* label to every op-query (Lemma 19).

Theorem 8 (**Generalized Downstream Impossibility**). *Let $\mathcal{M}_0 = (\mathcal{Z}, P_Z, g_v, g_l, \varphi)$ be the SCM of Assumption.12 with diffeomorphic generative maps $g_v : \mathcal{Z} \to \mathcal{X}$, $g_l : \mathcal{Z} \to \mathcal{T}$ and observational law $P_0 = \mathrm{Law}\big(g_v(Z), g_l(Z)\big)$, $Z \sim P_Z$, and let Q be the distribution of op-queries over observables. Suppose Π^* contains a permutation π that is non-trivial on the concept slots bound by the operator, and define*

$$\mathcal{M}_1 = \big(\mathcal{Z},\ \mathrm{Law}(\pi(Z)),\ g_v \circ \pi^{-1},\ g_l \circ \pi^{-1},\ \varphi\big). \tag{11}$$

Then for every (possibly randomized) learning pipeline $\mathcal{A}$ and every training sample size n,

$$\inf_{\mathcal{A}} \max_{b\in\{0,1\}} \Pr_{\mathcal{M}_b}\left[\mathcal{A} \textit{ errs on op-queries}\right] \;\geq\; \tfrac{1}{2}. \tag{12}$$

Interpretation. The bound follows from a two-point argument: the transcript (D, q, R) has the same law under both worlds, so the two error rates satisfy $e_0 + e_1 = 1$ and the worse one is at least chance (Appendix 6). It requires *no* modality-gap premise (Assumption 13), no regularity on the projector or LLM head, and no attained optimizer, since it quantifies over *algorithms* rather than over encoders—the strongest form that nonidentifiability can support, in the same sense as classical impossibility results for nonlinear ICA and disentanglement. Being exactly chance level for binary queries, the bound is tight, and as n never enters it, no amount of pre-training or instruction-tuning data escapes it. This explains the at- or below-chance behavior of MLLMs on the CLIP-blind pairs of MMVP Tong et al. (2024), and predicts a two-sided signature that we verify in the experiment: gains should appear only in categories covered by our operators and remain flat elsewhere.

4.4 Solution

CLIP and its variants. With regards to the discussion, it is obvious that hard nagative mining strategies play a key role to fix the problem. While existing strategies might ignore massive negative examples that contains multiple atomic-concept distortions defined by Theorem.3-5. So it would be possible to generate more high-quality hard-negative textual instances by stacking the compound nonidentified matching effects through **SWAP**$(\cdot)$,**REPLACE**$(\cdot)$, or **ADD**$(\cdot)$. This procedure was summarized into data curating algorithms (see **MC** in Algorithm.1 and **NT-MC** in Algorithm.2) that iteratively call the operators defined by Theorem 3-5, potentially improving the existing strategies to avoid pseudo-optimal text encoders. In the same spirit, we can also generate the hard negative images that holds consistent linguistic semantic with these hard-negative textual instances. In our implementation, we adopt SDXL-turbo Sauer et al. (2024) as our image-generation engine with regards to its balance between image quality and relatively high generation speed. It helps train the robust image encoder simultaneously for CLIP variants and MLLMs.

MLLMs. Since the problem in Theorem.8 arises from a pure CLIP-based encoders, a trivial remedy is to prevent their image encoders initiated from CLIP-style objective before training their MLLMs. The mainstream strategies align this conclusion with mid-training their visual encoders Wang et al. (2024), a fusion of multiple encodersWu et al. (2024), even a complete abandon of CLIP-based derivation Team et al. (2025); Bai et al. (2025); Luo et al. (2025), *etc*. Inspired from our theory, we show another way out: we initiate MLLMs' visual encoders from the CLIP model pre-trained with a portion of image-text pairs reconfigured by MC and NT-MC. Our experiments demonstrate the effectiveness of this strategy. It presents the potential of our theories in MLLM applications.

More implementation can be found in Appendix.C.

5 Experiments

In this section, we provide the experiments to verify our method (Section.4.4 and Algorithm.1,2). Since the method heavily depends on our theoretical results, our Appendix.D also provided the further empirical studies to verify the theoretical framework in this paper.

5.1 Experiments on Compositional Reasoning by CLIP

In this experiment, we focus on the evaluation of CLIP models trained and tested with regular hard negative pairs and hard negative pairs generated by iteratively calling to **SWAP**$(\cdot)$,**REPLACE**$(\cdot)$, or **ADD**$(\cdot)$, which reproduce more complex non-identified cases in the textual descriptions as well as the hard negative images close to these descriptions. The implementation of composition operators **SWAP**$(\cdot)$, **REPLACE**$(\cdot)$, and **ADD**$(\cdot)$ with respect to Theorems.3-5 are summarized by Algorithm.1. We apply Gemini 2.5 Pro as the proxy for their executions.

Setup. Similar with existing research, we focus on the evaluation of hard negative mining baselines, *e.g.*, NegCLIP (Yuksekgonul et al. (2023)) and TripleCLIP (Patel et al. (2024)) that both employed hard negative data to augment CLIP's training pool. Given this, we adopt Multi-Calling (MC) reconfiguration to these strategies, leading to our baselines NegCLIP (+MC) and TripleCLIP (+MC) to observe whether the iteratively-polished data generation can improve their hard negative mining

(a) Baselines provide hard negative data augmented to CC3M and CC12M respectively, then are evaluated across Replace, Swap, and Add composition categories in SugarCrepe. Bold indicates the best.

Methods	Replace			Swap		Add		Overall
	Object	Attribute	Relation	Object	Attribute	Object	Attribute	Avg.
CLIP (Vit-B/32)	56.90	56.85	51.99	50.81	50.00	53.93	60.55	54.43
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	59.39
NegCLIP (+NT-MC)	63.67	63.62	59.77	**58.92**	54.50	59.92	62.54	60.42
TripletCLIP	69.92	69.03	64.72	56.33	57.96	62.61	63.87	63.49
TripletCLIP (+MC)	71.42	70.88	64.42	57.30	59.17	62.55	64.90	64.38
TripletCLIP (+NT-MC)	**72.77**	**71.30**	**66.02**	58.87	**60.09**	**64.05**	**66.38**	**65.64**
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	69.08
NegCLIP (+MC-NT)	78.66	71.25	63.83	68.59	64.41	70.44	70.08	69.61
TripletCLIP	83.66	81.22	79.02	64.49	63.66	73.67	75.43	74.45
TripletCLIP (+MC)	**84.90**	80.85	80.25	67.97	**65.05**	73.44	**76.81**	75.61
TripletCLIP (+NT-MC)	84.61	**82.65**	**81.32**	**68.78**	64.79	**74.21**	76.64	**76.14**

(b) Zero-shot classification performance (Top-1 and Top-5 Accs) on VTAB and ImageNet1k.

	VTAB		ImageNet1k	
Methods	top-1	top-5	top-1	top-5
CC3M				
NegCLIP	12.25	36.38	4.67	12.69
NegCLIP (+MC)	12.95	35.91	6.04	16.62
NegCLIP (+NT-MC)	12.89	36.77	10.51	18.92
TripletCLIP	12.31	41.45	7.32	18.34
TripletCLIP (+MC)	12.91	42.67	9.42	23.11
TripletCLIP (+NT-MC)	**13.51**	**44.10**	**12.94**	**24.21**
CC12M				
NegCLIP	19.12	50.27	20.22	42.63
NegCLIP (+MC)	20.24	50.41	19.77	42.31
NegCLIP (+NT-MC)	21.11	49.77	21.17	42.94
TripletCLIP	20.81	53.40	23.31	47.33
TripletCLIP (+MC)	21.22	**54.71**	24.16	48.74
TripletCLIP (+NT-MC)	**22.36**	54.18	**24.94**	**48.89**

performances. To ensure the fairness, our text and image generation processes produce the consistent amount of hard negative samples used in TripleCLIP Patel et al. (2024). Our evaluation are conducted in two setups: (1) **Compositional Reasoning:** we demand all baselines to augment their hard negative training instances to the pre-training splits CC3M and CC12M in Changpinyo et al. (2021), then the evaluation is based on the challenging SugarCrepe benchmark Hsieh et al. (2023) with the pre-trained models. (2) **Zero-shot Classification:** the pre-trained models perform zero-shot classification on Visual Task Adaptation Benchmark (VTAB) and ImageNET1k[1] . We provide the ablation and more empirical verification in Appendix.D3.

Main Results. Tables (a),(b) jointly corroborate our theoretical claim that compound compositional perturbations expose pseudo-optimal text encoders far more effectively than single-edit hard negatives, while preserving—and often enhancing—general visual–semantic alignment:

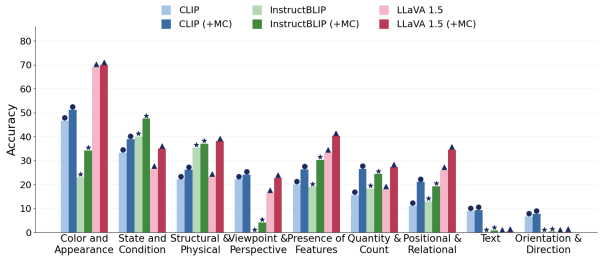
Compositional reasoning. In Table (a), every reconfigured variant improves upon its own baseline on the SugarCrepe overall average: MC lifts NegCLIP from 57.18 to 59.39 and TripletCLIP from 63.49 to 64.38 on CC3M, while NT-MC adds a further 1.03 and 1.26 points; the ordering NT-MC $>$ MC $>$ single-edit persists on CC12M, where TripletCLIP (+NT-MC) attains 76.14 against 54.43 for vanilla CLIP. Crucially, the gains concentrate exactly where Theorems 3–5 predict, i.e., Replace-Attribute ($58.12 \rightarrow 63.24$) and Replace-Relation ($54.48 \rightarrow 60.79$) for NegCLIP on CC3M, together with the Swap categories ($64.49 \rightarrow 68.78$ on CC12M) that single-concept negatives barely touch. Such selectivity corroborates our nonidentifiability diagnosis, i.e., only compound perturbations separate g^* from the pseudo-optimal g^{**} across multiple concept slots simultaneously, rather than a generic data-scaling effect.

Zero-shot classification. In Table (b), compound negatives never trade generality for compositionality. ImageNet1k top-1 nearly doubles on CC3M (NegCLIP $4.67 \rightarrow 10.51$; TripletCLIP $7.32 \rightarrow 12.94$) and VTAB top-5 rises from 41.45 to 44.10, whereas the CC12M gains stay positive but smaller (TripletCLIP $23.31 \rightarrow 24.94$ and $47.33 \rightarrow 48.89$), as a larger corpus already supplies broad concept coverage. The only regressions are marginal (e.g., NegCLIP VTAB top-5 $36.38 \rightarrow 35.91$ under MC), confirming that intervening beyond the observational distribution repairs composition while leaving modality-invariant alignment intact.

5.2 Experiments on Compositional Reasoning by MLLMs

Benchmarks and Baselines. We use MMVP Tong et al. (2024) as our primary benchmark because it is explicitly designed to surface CLIP-blind visual patterns implicitly aligned with the operators (SWAP, REPLACE, ADD) we analyzed (Table.4). To this, our evaluation enables the most direct

[1]https://github.com/LAION-AI/CLIP_benchmark



(a) Per-category Acc on MMVP. "+MC" denotes the models are pre-trained with the visual encoder derived from TripleCLIP (+MC).

(b) Downstream results on two VLM hosts with different visual encoders.

Host	Method	MMVP	CV-Bench	GQA
LLaVA-1.5	Base	24.7‡	56.6§	63.3§
	+NegCLIP†	25.9	56.4	62.5
	+TripletCLIP†	27.7	58.1	62.9
	— (+MC)	31.3	**59.7**	64.2
	— (**+MC-NT**)	**33.2**	59.3	**64.9**
InstructBLIP	Base	16.7‡	46.9§	49.5§
	+NegCLIP†	18.7	46.5	48.7
	+TripletCLIP†	21.8	48.4	51.1
	— (+MC)	26.6	50.0	51.8
	— (**+MC-NT**)	**28.3**	**51.7**	**52.9**

Figure 2: Compositional reasoning performances based on MLLMs.

test of whether our encoder-level improvements transfer downstream We adopt LLaVa1.5 Liu et al. (2024) and InstructBLIP Dai et al. (2023) as the MLLM base models due to their visual encoders pre-trained via CLIP-like strategies, then we replace their visual encoders by the image encoders optimized through NegCLIP, TripleCLIP, NegCLIP (+MC), TripleCLIP (+MC), respectively. The post-training phases are consistent with their releases. Beyond this, CV-Bench and GQA also serve as complementary probes for other-dimension ability, with a large amount of their test instances with regards to compositional reasoning ability.

Results. Fig. 2 validates Theorem.8 closing text-side composition nonidentifiability is what unlocks downstream MLLM gains. Three observations show from Table (b) in Fig. 2 [2]. **First**, our method improves every host $\times$ benchmark cell: on LLaVA-1.5, MMVP rises from 24.7 to 33.2, CV-Bench from 56.6 to 59.3 and GQA from 63.3 to 64.9; on InstructBLIP, from 16.7 to 28.3, 46.9 to 51.7 and 49.5 to 52.9, respectively. The improvement therefore holds across two structurally different hosts and is not tied to a particular vision tower or connector. **Second**, the advantage over existing hard-negative approaches persists downstream: NegCLIP *falls below* the Base host on both non-MMVP benchmarks and on both hosts (CV-Bench 56.4 vs. 56.6 and 46.5 vs. 46.9; GQA 62.5 vs. 63.3 and 48.7 vs. 49.5), indicating that its MMVP gain is bought at the cost of general VQA ability, whereas $+$MC-NT surpasses TripletCLIP on all six cells and never trades one axis for another. **Third**, the gains track the remaining headroom on the compositional axis: they are largest on MMVP, where both Bases leave the most room, and the weaker host benefits more on every benchmark, closing much of its gap to LLaVA-1.5. Finally, comparing $+$MC with $+$MC-NT isolates the contribution of NT: it lifts MMVP further ($31.3 \rightarrow 33.2$ and $26.6 \rightarrow 28.3$) and GQA ($64.2 \rightarrow 64.9$ and $51.8 \rightarrow 52.9$); the only exception is CV-Bench on LLaVA-1.5, where the two variants are within noise.

The per-category breakdown (Fig. 2a) in MMVP sharpens this picture and provides a two-sided falsification of our theory. The largest LLaVA-1.5 (+MC) gains concentrate on categories *covered* by our operators (cf. Tab. 4): *Positional & Relational* ($+\sim 8$, SWAP-OBJ), *Structural & Physical* ($+\sim 14$, REPLACE), and *Presence of Features* ($+\sim 5$, ADD-ATT); whereas *Orientation*, *Text (OCR)*, and *Viewpoint*—categories with *no* corresponding operator—remain flat across all variants ($\Delta \approx 0$). This absence of spurious gains rules out generic data-scaling effects and pinpoints the recovered information to exactly the operators licensed by Theorems 5-7. The same trend transfers to InstructBLIP (e.g., State & Condition $40 \rightarrow 48$), confirming that the bottleneck argument is backbone-agnostic.

In Appendix.D3, we provide a further MLLM architecture ablation to verify our method.

6 Conclusion

We developed a token-aware causal framework proving that CLIP's contrastive optimum leaves compositionality nonidentifiable, propagating through the modality gap to image encoders and yielding an algorithm-independent impossibility for MLLMs. Guided by this diagnosis, compound hard-negative curation (MC/NT-MC) measurably repairs compositional reasoning while preserving zero-shot generality, pointing to intervention beyond the observational distribution.

[2] **Bold** marks the best entry per column within each host block. † Existing hard-negative methods, compared downstream on the two hosts (LLaVA-1.5 and InstructBLIP) with different pretrained encoders.

AI USE STATEMENT

Large language model assistants (Claude, Anthropic) were used to help to implement the algorithm coding, evaluation and aggregation code, to diagnose failed runs, to draft and edit the text of this paper. All writing content, theory, and algorithms are derived from humans, and reference are checked to prevent hallucination. The theoretical statements are verified by authors and AI to promise the rigorousness, and all proofs are proposed and checked by authors. The authors reviewed every sentence and take responsibility for the content.

REFERENCES

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 5723–5738, 2022.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pp. 1298–1312. PMLR, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, pp. 231, 1951.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in neural information processing systems*, 35:21400–21413, 2022.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021.

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.

Jiahui Hu, Zongyen Liu, Kelvin Yang, Yu-Hsuan Shen, Sheng-Yu Chai, and Chen Sun. Cola: A compositional text-to-image benchmark. *arXiv preprint arXiv:2305.15472*, 2023.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–438, 1999.

Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9): 939–952, 1982.

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.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *International conference on machine learning*, pp. 2668–2677, 2018.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Disentanglement of grouped factors of variation by leveraging partial group supervision. *arXiv preprint arXiv:2010.08226*, 2021.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International conference on machine learning*, pp. 1078–1088. Pmlr, 2020.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

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

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

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24960–24971. IEEE, 2025.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pp. 186–195. PMLR, 2020.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4658–4664, 2019.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

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

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.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pp. 314–332. Springer, 2020.

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.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

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.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19412–19424, 2024.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

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.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

Ding Zhou and Xue-Xin Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. *Advances in neural information processing systems*, 33:7234–7247, 2020.

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.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

APPENDIX.A: BACKGROUND

A1.NOTATIONS

The notations frequently used are summarized in Table.3.

A2.RELATED WORK

Here we provide the technical background and relevant research as the foundation in this paper.

CLIP and its variants. CLIP (Contrastive Language-Image Pre-training) and its variants Radford et al. (2021); Cherti et al. (2023); Sun et al. (2023); Stevens et al. (2024) have emerged as a breakthrough in transferring visual representations through natural language supervision, enabling remarkable generalization across diverse visual recognition tasks Zareian et al. (2021); Ghiasi et al. (2021); Baevski et al. (2022). The core of CLIP lies in its contrastive pre-training on massive image-text datasets, which facilitates open-vocabulary prediction. This is achieved by using a prompt template, such as *i.e.*, "a photo of a [CLASS]," where any potential category name can be semantically encoded to serve as a category-specific classification weight.

Image-text compositional reasoning in CLIP. Recent research has focused on enhancing the compositional understanding of vision-language models through various training strategies, such as incorporating additional data, models, or specialized loss functions Hu et al. (2023). A prominent approach involves explicitly training models to differentiate correct captions from synthetically generated hard negatives. However, the effectiveness of these methods is often measured on benchmarks that may themselves be flawed. Several studies have highlighted that biased datasets can lead to an overestimation of a model's true capabilities Bender et al. (2021). To enable more faithful evaluations, dataset de-biasing methods have been proposed Le Bras et al. (2020); Ross et al. (2017); Pratt et al. (2020). Techniques like adversarial filtering, for instance, aim to remove "easy" or artifact-laden examples from datasets to ensure that models are evaluated on more challenging and representative data Le Bras et al. (2020); Ross et al. (2017). This focus on robust evaluation is critical to determine whether models are genuinely acquiring compositional reasoning or merely exploiting statistical biases within the evaluation benchmarks.

Causal representation learning (CRL) and concept discovery. In recent years, SCMs have found significant application in representation learning. In particular, causal representation learning (CRL) Scholkopf et al. (2021); Schölkopf (2019) aims to learn the latent generative factors behind high-dimensional data. This exciting field has seen significant progress in the last few years Khemakhem et al. (2020); Brehmer et al. (2022); Lachapelle et al. (2021); Monti et al. (2020); Yao et al. (2023); Zheng et al. (2024); Zhou & Wei (2020); Espinosa Zarlenga et al. (2022). A fundamental perspective in this field is to ensure that the model parameters we attempt to recover are identifiable Khemakhem et al. (2020); Hyvärinen & Pajunen (1999). Concept discovery is an important sub-field of machine learning which extracts human-intepretable concepts from pre-trained models. We do not attempt to list the numerous works in this direction, see e.g., Burns et al. (2022); Chen et al. (2020); Meng et al. (2022); Olah et al. (2020); Niven & Kao (2019); Kim et al. (2018); Ao et al. (2022); Zheng et al. (2024).

A3.CAUSAL REPRESENTATION AND OPTIMAL CROSS-MODAL ALIGNMENT: A REVIEW

Structural Causal Models (SCMs). The concept of SCM pioneered by Judea Pearl, have become a cornerstone of modern causal inference. They provide a mathematical framework for representing causal relationships within a system. An SCM consists of a set of variables and a set of equations that describe how each variable is determined by others in the model. This framework allows us to not only model statistical associations but also to predict the effects of interventions and to reason about counterfactuals. At its core, an SCM is defined by a collection of endogenous (or child) variables, whose values are determined by other variables within the model, and exogenous (or parent) variables, which are external to the model and treated as random noise or unobserved influences. The relationships between these variables are specified by structural equations, which are deterministic functions that define how each endogenous variable is generated from its direct causes and an associated exogenous noise term. The power of SCMs lies in their ability to make the causal assumptions explicit. By defining the causal graph—a directed acyclic graph (DAG) where nodes represent vari-

Category	Notation	Type/Space	Meaning
Data	$x^{(\mathsf{img})}$	$\mathcal{X}_{\mathsf{img}}$	Single image observation (image instance).
Data	$x^{(\mathsf{tex})}$	$\mathcal{X}_{\mathsf{tex}}$	Sentence-level text observation in the token-agnostic setting.
Data	$X^{(\mathsf{tex})}$	$\bigcup_{k \le k_{\mathsf{max}}} \mathbb{R}^{d \times k}$	Token-level text matrix; the i-th column represents the embedding of the i-th token.
Batch	$\mathcal{D}^{(K)}$	Set	A minibatch of paired samples $\{(x_i^{(\mathsf{img})}, x_i^{(\mathsf{tex})})\}_{i=1}^{K}$.
Distribution	p_{mm}	Joint distribution	Multimodal image–text joint distribution.
Latent (shared)	z_{inv}	$\mathbb{R}^{n_{\mathsf{inv}}}$	Modality-invariant latent variable capturing shared semantics.
Latent (img dep.)	$z_{\mathsf{dp}}^{(\mathsf{img})}$	$\mathbb{R}^{n_{\mathsf{dp}}^{(\mathsf{img})}}$	Image-dependent latent partition conditioned on z_{inv}.
Latent (img priv.)	$z_{\mathsf{pr}}^{(\mathsf{img})}$	$\mathbb{R}^{n_{\mathsf{pr}}^{(\mathsf{img})}}$	Image-private latent partition with an independent prior.
Latent (tex priv.)	$z_{\mathsf{pr}}^{(\mathsf{tex})}$	$\mathbb{R}^{n_{\mathsf{pr}}^{(\mathsf{tex})}}$	Text-private latent partition with an independent prior.
Latent (token)	$z_i^{(\mathsf{tex})}$	$\mathbb{R}^{n_i^{(\mathsf{tex})}}$	Token-dependent latent variable for the i-th token, generated autoregressively.
Mixing function	$\mathbf{f}$	$\mathcal{Z}_{\mathsf{img}} \to \mathcal{X}_{\mathsf{img}}$	Structural causal mixing function for image generation.
Mixing function	$\mathbf{g}$	$\mathcal{Z}_{\mathsf{tex}} \to \mathcal{X}_{\mathsf{tex}}$	Structural causal mixing function for sentence-level text generation.
Mixing function (token)	$\mathbf{g}_i$	$\mathcal{M}_i \to \mathbb{R}^d$	Token-level mixing function generating the i-th token embedding.
Encoder (CLIP)	f	$\mathcal{X}_{\mathsf{img}} \to (0,1)^{n_{\mathsf{inv}}}$	CLIP image encoder (distinct from the generative mixing function).
Encoder (CLIP)	g	$\mathcal{X}_{\mathsf{tex}} \to (0,1)^{n_{\mathsf{inv}}}$	CLIP text encoder; in the token-aware setting, the input is $X^{(\mathsf{tex})}$.
Representation	$\tilde{f}, \tilde{g}$	$\mathbb{S}^{m-1}$	Normalized representations used in the InfoNCE objective.
Objective	$\mathcal{L}_{\mathsf{InfoNCE}}$	Scalar	Contrastive learning objective for cross-modal retrieval.
Temperature	γ	> 0	Softmax temperature in the InfoNCE loss.
Alignment	$\mathcal{L}_{\mathsf{MMAlign}}$	Scalar	Multimodal alignment objective based on representation alignment and entropy regularization.
Entropy	$H(\cdot)$	Functional	Differential entropy of a continuous random variable.
Dimension	n_{inv}	$\mathbb{N}$	Dimensionality of the invariant semantic subspace.
Length	k, k_{max}	$\mathbb{N}$	Token sequence length and maximum allowed length.
Permutation	$\pi \in \Pi_k$	Permutation	Permutation over token positions.
Reordering	$\pi(X^{(\mathsf{tex})})$	$\mathbb{R}^{d \times k}$	Token matrix obtained by reordering columns according to π.
Pseudo-optimal	g^{**}	Function	Pseudo-optimal text encoder that achieves alignment but is insensitive to compositional perturbations.

Table 3: Notation used throughout the paper. 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.11 and 12 .

ables and directed edges represent causal relationships—we can analyze the flow of causal influence and determine which variables are causes and which are effects. This explicit representation is cru-

cial for tasks such as identifying causal effects from observational data, understanding confounding bias, and achieving robust predictions under distributional shifts.

A Structural Causal Model (SCM) is formally defined as a tuple $\mathcal{M} := \langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u}) \rangle$, where:

$\boldsymbol{V} = V_1, \ldots, V_n$ is a set of endogenous variables. These are the variables whose values are determined by other variables within the model. In the context of your assumption, the observed data, such as an image $x^{(\mathsf{img})}$ and a text description $x^{(\mathsf{tex})}$, are considered endogenous.

$\boldsymbol{U} = U_1, \ldots, U_n$ is a set of exogenous variables. These are mutually independent random variables that represent unobserved background conditions or noise. They are the ultimate sources of randomness in the model. In your assumption, the latent variables z_{inv}, $z^{(\mathsf{img})}\mathsf{dp}$, $z^{(\mathsf{img})}\mathsf{pr}$, $z^{(\mathsf{tex})}\mathsf{dp}$, and $z^{(\mathsf{tex})}\mathsf{pr}$ can be thought of as being determined by exogenous sources of variation.

$\mathcal{F} = f_1, \ldots, f_n$ is a set of structural equations, where each function f_i assigns a value to the corresponding endogenous variable V_i based on its direct causes $\mathsf{pa}(V_i) \subseteq \boldsymbol{V} \setminus V_i$ and its associated exogenous variable U_i:

$$V_i := f_i(\mathsf{pa}(V_i), U_i) \tag{13}$$

This equation states that the value of V_i is causally determined by the function f_i of its parents $\mathsf{pa}(V_i)$ and the exogenous noise U_i. $P(\boldsymbol{u})$ is a probability distribution over the exogenous variables $\boldsymbol{U}$.

Now, let's connect this general definition to the variables in your specific SCM assumption for image-text data generation. The assumption posits a hierarchical generation process that can be mapped onto the SCM framework. In particular, Exogenous Variables: The fundamental sources of variation are the latent variables drawn from their respective prior distributions: $z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}}$: The modal-invariant feature. $z_{\mathsf{pr}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{img})}}$: The image-private feature. $z_{\mathsf{pr}}^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{tex})}}$: The text-private feature. The dependent partitions, $z_{\mathsf{dp}}^{(\mathsf{img})}$ and $z_{\mathsf{dp}}^{(\mathsf{tex})}$, are also influenced by exogenous noise, but their generation is conditioned on z_{inv}. Endogenous Variables: These are the variables whose values are generated within the model. This includes the dependent latent variables and the final observed data: $z_{\mathsf{dp}}^{(\mathsf{img})}$: The image-dependent partition, generated based on $z\mathsf{inv}$. $z_{\mathsf{dp}}^{(\mathsf{tex})}$: The text-dependent partition, generated based on $z\mathsf{inv}$. $x^{(\mathsf{img})}$: The generated image. $x^{(\mathsf{tex})}$: The generated text. Structural Equations: The assumption provides the structural equations for the final observed variables, $x^{(\mathsf{img})}$ and $x^{(\mathsf{tex})}$:

$$x^{(\mathsf{img})} := \mathbf{f}\big(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{img})}, z^{(\mathsf{img})}\mathsf{pr}\big); \quad x^{(\mathsf{tex})} := \mathbf{g}\big(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{tex})}, z^{(\mathsf{tex})}\mathsf{pr}\big), \tag{14}$$

There are also implicit structural equations for the dependent partitions:

$$z_{\mathsf{dp}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{img})}}(\cdot|z_{\mathsf{inv}}) \quad z_{\mathsf{dp}}^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{tex})}}(\cdot|z_{\mathsf{inv}}) \tag{15}$$

These conditional distributions can be expressed as structural equations with their own exogenous noise terms. For example, $z_{\mathsf{dp}}^{(\mathsf{img})} := h_{\mathsf{img}}(z_{\mathsf{inv}}, U_{\mathsf{img_dp}})$, where $U_{\mathsf{img_dp}}$ is an exogenous noise variable.

By laying out the SCM in this manner, we can clearly see the causal dependencies. The modal-invariant feature z_{inv} is a common cause of both the image and the text, which is what creates the "mutual semantics" between them. The private features, $z^{(\mathsf{img})}\mathsf{pr}$ and $z^{(\mathsf{tex})}\mathsf{pr}$, account for the variability within each modality that is independent of the other. The dependent partitions, $z^{(\mathsf{img})}\mathsf{dp}$ and $z^{(\mathsf{tex})}\mathsf{dp}$, represent stylistic or content variations that are specific to a modality but are still influenced by the core shared semantics. This detailed causal structure is what allows for a rigorous analysis of how a model like CLIP might be able to disentangle and recover the causally meaningful feature z_{inv}.

Under p_{mm} interpreted as the generative process defined 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 $\boldsymbol{z}^{(\mathsf{inv})}$ through a causal lens. More specifically, when cross-modal pairs are generated from some multimodal data distribution defined by the SCM Assumption.11 Daunhawer et al. (2022); Yao et al. (2023), the causal representation learned by CLIP yields their optimal encoders (f^*, g^*) as follows:

Theorem 9. ***(Block-Identified Modal-invariant Alignment (Token-agnostic))*** *Consider the multi-modal (image-text) pair generated by Assumption.11. If their densities and mappings satisfy:* ***1).*** $\mathbf{f}$, $\mathbf{g}$ *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,*

$$\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}} := \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},x^{(\mathsf{tex})}\rangle \\ \sim p_{\mathsf{mm}}}} \left[||f(x^{(\mathsf{img})})-g(x^{(\mathsf{tex})})||\right] - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{x}^{(\mathsf{tex})})) \tag{16}$$

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 encoders* f^* *and* g^*, *there exist invertible functions* h_f *and* h_g *satisfying the following decompositions, respectively:*

$$f^*=h_f\circ\mathbf{f}^{-1}_{1:n_{\mathsf{inv}}},\ g^*=h_g\circ\mathbf{g}^{-1}_{1:n_{\mathsf{inv}}} \tag{17}$$

Note that Theorem. can be also observed as the repackaged results from theorems in ,

Corollary 10. *(Informal) The optimal encoders* f^*, g^* *in Theorem.9 are almost equivalent with* $(f^*, g^*)=\arg\min_{f,g}\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}+\mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}$ *with infinite training pairs.*

Insight Grounded in the principles of block identifiability (Von Kügelgen et al. (2021)), Theorem.9 implies how text and images encoders of CLIP can achieve modal invariance via learning causal representation. It justifies that the optimal encoders (f^*, g^*) learn features that isolate a shared latent variable, $\boldsymbol{z}_{\mathsf{inv}}$. This result provides a formal paradigm that CLIP's training objective leads to the feature matching for the image and language representation.

A4. Illustration of Theorem 3-5

We specify how the operational statements in Theorem.3-5 correspond to the operators **SWAP**, **REPLACE**, **ADD** that transfer the original text into the hard negative version. In the case of Theorem.3, when the sequence of token embedding $X^{(\mathsf{tex})}$ denotes a text *e.g.*,

"*a white cat and a black dog play together*"

and if there exists the permutation $\pi(X^{(\mathsf{tex})})$ such that the text that re-ordered $X^{(\mathsf{tex})}$ by $\pi(X^{(\mathsf{tex})})$ *i.e.*,

"*a black dog and a white cat play together*"

semantically aligned with $X^{(\mathsf{tex})}$ by g^{**}, the text

"*a white dog and a black cat play together*"

that re-ordered $X^{(\mathsf{tex})}$ by $\hat{\pi}(X^{(\mathsf{tex})})$ could be also provably aligned with $X^{(\mathsf{tex})}$ (Eq.10,11), despite its significant difference from the original text $X^{(\mathsf{tex})}$ in the semantic. In the case of Theorem.4, similarly, suppose that a sequence of token embedding $X^{(\mathsf{tex})}$ implying a textual description, *e.g.*,

"*a horse* ***on*** *the grass*",

which is correctly matched with its rephrased version, *e.g.*,

"*the grass* ***under*** *a horse*",

where "***under***" indicates the token rephrased from "***on***" (*i.e.*, $\mathsf{RP}(X^{(\mathsf{tex})}_{:,j})$ rephrased from $X^{(\mathsf{tex})}_{:,j}$ in Eq.12) after permuting the columns in $X^{(\mathsf{tex})}$ by $\pi(X^{(\mathsf{tex})})$. Given this, the pseudo-optimal text encoder g^{**} achieves the modal-invariant alignment with the incorrect matching between

"*the horse* ***on*** *the grass*"

and

"*the grass* ***on*** *the horse*"

according to Eq.12,13. Finally, in the case of Theorem.5, $\forall j$ and $X_{:,j}^{(\text{tex})}$, it is possible to identify which location j should be inserted with a new atomic concept (*e.g.*, *no*) conditioned on $X_{:,j}^{(\text{tex})}$ (*e.g.*, we have *no* as $\text{ADD}(X_{:,j}^{(\text{tex})})$ derived from *flowers* as $X_{:,j}^{(\text{tex})}$).

APPENDIX.B: ASSUMPTIONS AND PROOFS

Before we provide the proofs to our main theoretical results, we go through some assumptions used in our paper.

Assumption 11. (**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.*, $z_{\text{inv}} \sim p_{\boldsymbol{z}_{\text{inv}}}$; given z_{inv}, we obtain image-dependent partition $z_{\text{dp}}^{(\text{img})} \sim p_{\boldsymbol{z}_{\text{dp}}^{(\text{img})}}(\cdot|z_{\text{inv}})$ and text-dependent partition $z_{\text{dp}}^{(\text{tex})} \sim p_{\boldsymbol{z}_{\text{dp}}^{(\text{tex})}}(\cdot|z_{\text{inv}})$ specific to the image domain and text domain, respectively; and we also have the image-private partition $z_{\text{pr}}^{(\text{img})}$ and text-private partition $z_{\text{pr}}^{(\text{tex})}$ drawn from independent priors, *i.e.*, $z_{\text{pr}}^{(\text{img})} \sim p_{\boldsymbol{z}_{\text{pr}}^{(\text{img})}}$, $z_{\text{pr}}^{(\text{tex})} \sim p_{\boldsymbol{z}_{\text{pr}}^{(\text{tex})}}$; then each image-text pair $\langle x^{(\text{img})}, x^{(\text{tex})} \rangle$ is generated through the nonlinear mixing functions $\mathbf{f}$,$\mathbf{g}$ to specify p_{mm}:

$$x^{(\text{img})} := \mathbf{f}(z^{(\text{img})}) = \mathbf{f}\big(z_{\text{inv}}, z_{\text{dp}}^{(\text{img})}, z_{\text{pr}}^{(\text{img})}\big); \quad x^{(\text{tex})} := \mathbf{g}(z^{(\text{tex})}) = \mathbf{g}\big(z_{\text{inv}}, z_{\text{dp}}^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}\big), \tag{18}$$

where z_{inv}, $z_{\text{dp}}^{(\text{img})}$, $z_{\text{pr}}^{(\text{img})}$, $z_{\text{dp}}^{(\text{tex})}$, $z_{\text{pr}}^{(\text{tex})}$ denote real-value vectors drawn from the distributions with respect to $\boldsymbol{z}_{\text{inv}}$, $\boldsymbol{z}_{\text{dp}}^{(\text{img})}$, $\boldsymbol{z}_{\text{pr}}^{(\text{img})}$, $\boldsymbol{z}_{\text{dp}}^{(\text{tex})}$, $\boldsymbol{z}_{\text{pr}}^{(\text{tex})}$ 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 $\boldsymbol{z}_{\text{inv}}$ and $\boldsymbol{z}_{\text{dp}}^{(\text{tex})}$. Derived from the relaxed premise, CLIP holds the alignment to identify the modal-invariant part of each image-text pair.

Assumption 12. (**Token-aware continuous relaxation 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 $z_{\text{pr}}^{(\text{img})}$ and text-private partition $z_{\text{pr}}^{(\text{tex})}$ are drawn by $z_{\text{pr}}^{(\text{img})} \sim p_{\boldsymbol{z}_{\text{pr}}^{(\text{img})}}$, $z_{\text{pr}}^{(\text{tex})} \sim p_{\boldsymbol{z}_{\text{pr}}^{(\text{tex})}}$; and the image-dependent partition is obtained by $z_{\text{dp}}^{(\text{img})} \sim p_{\boldsymbol{z}_{\text{dp}}^{(\text{img})}}(\cdot|z_{\text{inv}})$. Suppose $z_i^{(\text{tex})}$ as the token-dependent partition of the ith token, and each of them is recursively sampled via $z_i^{(\text{tex})} \sim p_{\boldsymbol{z}_i^{(\text{tex})}}(\cdot|z_{\text{inv}}, \{z_j^{(\text{tex})}\}_{j=1}^{i-1})$; then each image-text pair $\langle x^{(\text{img})}, X^{(\text{tex})} \rangle$ is generated through the nonlinear mixing functions $\mathbf{f}$,$\{\mathbf{g}_i\}_{i=1}^{\text{k}_{\text{max}}}$ to specify p_{mm}

$$x^{(\text{img})} := \mathbf{f}\big(z_{\text{inv}}, z_{\text{dp}}^{(\text{img})}, z_{\text{pr}}^{(\text{img})}\big); \quad X_{:,i}^{(\text{tex})} := \mathbf{g}_i\big(z_{\text{inv}}, \{z_j^{(\text{tex})}\}_{j=1}^{i}, z_{\text{pr}}^{(\text{tex})}\big). \tag{19}$$

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

Assumption.12 extends the image-language SCM definition in Assumption.11 by drawing the inspiration from the recent memory-argumented Bayesian LLM prior Wei et al. (2021).

Assumption 13 (**Constant Modality Gap** (Zhang et al.))**.** Let $\langle \boldsymbol{x}^{(\text{img})}, \boldsymbol{X}^{(\text{tex})} \rangle \sim p_{\text{mm}}$ denote a matched image-text pair. There exists a constant vector $\boldsymbol{\delta} \in \mathbb{R}^{n_{\text{inv}}}$ such that for any image encoder $\hat{f}$ and text encoder $\hat{g}$ jointly trained under the contrastive objective (Eq. 1), the feature-level discrepancy satisfies:

$$\hat{f}(\boldsymbol{x}^{(\text{img})}) - \hat{g}(\boldsymbol{X}^{(\text{tex})}) = \boldsymbol{\delta}, \quad \text{a.s. } \langle \boldsymbol{x}^{(\text{img})}, \boldsymbol{X}^{(\text{tex})} \rangle \sim p_{\text{mm}}, \tag{20}$$

where $\boldsymbol{\delta}$ is orthogonal to the span of $\{\hat{g}(\boldsymbol{X}^{(\text{tex})})\}_{\boldsymbol{X}^{(\text{tex})} \sim p_{\boldsymbol{X}^{(\text{tex})}}}$ and to the span of $\{\hat{f}(\boldsymbol{x}^{(\text{img})}) - \boldsymbol{\delta}\}_{\boldsymbol{x}^{(\text{img})} \sim p_{\boldsymbol{x}^{(\text{img})}}}$.

B1. PROOF OF THEORM.9

The proof sketch of our Theorm.9 can be derived into three parts. In the first part, we show how to construct the optimal f^*, g^* to fullfil the objectives, further leading to h_f, h_g for their decomposition. In the second part, we prove h_f, h_g are modality-invariant with respect to any features in the image-specific partition $\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{img})}$, $\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{img})}$ and the text-specific partitions $\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{tex})}$, $\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{tex})}$, thus, they only recover the modal-invariant partitions of the inverses $\mathbf{f}^{-1}$, $\mathbf{g}^{-1}$. Finally we verify the invertability of h_f, h_g derived from Proposition.4.4 in Zimmermann et al. (2021) to fulfill the function decomposition.

Construction of h_f, h_g. The global minimum of $\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}$ is reached when their first term are minimized while the second and third terms are maximized, respectively. According to Jaynes (1982), the unique maximum entropy distribution on $(0,1)^{n_{\mathsf{inv}}}$ is uniform distribution without extra moment constraint. To this, we show how to construct a pair of f,g that map $\boldsymbol{x}^{(\mathsf{img})}$,$\boldsymbol{x}^{(\mathsf{tex})}$ into $(0,1)^{n_{\mathsf{inv}}}$, simultaneously attain the global minimization of $\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}$. They would further lead to the construction of h_f, h_g.

Let first consider f. To see this, we consider the smooth function $\mathbf{f}_{1:n_{\mathsf{inv}}}^{-1} : \mathcal{X}_{\mathsf{img}} \mapsto \mathcal{C}_{\mathsf{inv}}$, the inverse of $\mathbf{f}^{-1}$ restricted to its first n_{inv} dimension. This exists since $\mathbf{f}$ is invertible and smooth by the first primise. Based on Assumption.11, we obtain $\mathbf{f}_{1:n_{\mathsf{inv}}}^{-1}(\boldsymbol{x}^{(\mathsf{img})}) = \boldsymbol{z}^{(\mathsf{inv})}$. Here we further construct a function $\mathbf{d} : \mathcal{C}_{\mathsf{inv}} \to (0,1)^{n_{\mathsf{inv}}}$ to map $\boldsymbol{z}_{\mathsf{inv}}$ into a uniform random variable, which is achieved by a recursive building principles known as *Damois construction* Darmois (1951):

$$d_i(\boldsymbol{z}^{(\mathsf{inv})}) = F_i(z_i^{(\mathsf{inv})} | \boldsymbol{z}_{1:i-1}^{(\mathsf{inv})}), i = 1, \cdots, n_{\mathsf{inv}} \tag{21}$$

where $F_i(z_i^{(\mathsf{inv})} | \boldsymbol{z}_{1:i-1}^{(\mathsf{inv})})$ denotes the conditional cumulative distribution function (CDF) of $z_i^{(\mathsf{inv})}$ given $\boldsymbol{z}_{1:i-1}^{(\mathsf{inv})}$. Derived from such construction, $\mathbf{d}(\boldsymbol{z}^{(\mathsf{inv})})$ is uniformly distributed on $(0,1)^{n_{\mathsf{inv}}}$Darmois (1951), and is also smooth due to the third primise. To this, we define a composite smooth function $f^* := \mathbf{d} \circ \mathbf{f}_{1:n_{\mathsf{inv}}}^{-1}$.

Then we turn to consider g. Similarly, we also have the inverse smooth function $\mathbf{g}_{1:n_{\mathsf{inv}}}^{-1} : \mathcal{X}_{\mathsf{tex}} \mapsto \mathcal{C}_{\mathsf{inv}}$. Based on Assumption.11, it also holds a smooth restricted function $\mathbf{g}_{1:n_{\mathsf{inv}}}^{-1}(\boldsymbol{x}^{(\mathsf{tex})}) = \boldsymbol{z}^{(\mathsf{inv})}$. Using the Damois construction $\mathbf{d} : \mathcal{C}_{\mathsf{inv}} \to (0,1)^{n_{\mathsf{inv}}}$ above, we also define the other composite smooth function $g^* := \mathbf{d} \circ \mathbf{g}_{1:n_{\mathsf{inv}}}^{-1}$.

Given this, we consider the following derivation:

$$\begin{aligned}
\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f^*, g^*) &= \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}}} \left[\left|\left| f^*(x^{(\mathsf{img})}) - g^*(x^{(\mathsf{tex})}) \right|\right|_2^2 \right] - H\Big(f^*(x^{(\mathsf{img})})\Big) - H\Big(g^*(x^{(\mathsf{tex})})\Big) \\
&= \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}}} \left[\left|\left| \mathbf{d}(z_{(\mathsf{inv})}) - \mathbf{d}(z_{(\mathsf{inv})}) \right|\right|_2^2 \right] - H\Big(\mathbf{d}(z_{(\mathsf{inv})})\Big) - H\Big(\mathbf{d}(z_{(\mathsf{inv})})\Big) \\
&= 0.
\end{aligned} \tag{22}$$

Given $f^* : \mathcal{X}_{\mathsf{img}} \mapsto (0,1)^{n_{\mathsf{inv}}}$ and $g^* : \mathcal{X}_{\mathsf{tex}} \mapsto (0,1)^{n_{\mathsf{inv}}}$ as the functions that obtain the global minimum of $\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f^*, g^*)$, *i.e.*

$$\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f^*, g^*) = \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}}} \left[\left|\left| f^*(x^{(\mathsf{img})}) - g^*(x^{(\mathsf{tex})}) \right|\right|_2^2 \right] - H\Big(f(x^{(\mathsf{img})})\Big) - H\Big(g(x^{(\mathsf{tex})})\Big) \tag{23}$$

Let define $h_f{=}f^* \circ \mathbf{f}$ and $h_g{=}g^* \circ \mathbf{g}$. In terms of Eq.6, the formulation above implies h_f, h_g with

$$\begin{aligned} &\mathop{\mathbb{E}}_{p_{\mathsf{mm}}}\left[\left|\left|h_f(z^{(\mathsf{img})}){-}h_g(z^{(\mathsf{tex})})\right|\right|_2^2\right]=0 \\ \iff &\mathop{\mathbb{E}}_{p_{\mathsf{mm}}}\left[\left|\left|h_f\Big((z_{\mathsf{inv}},z_{\mathsf{dp}}^{(\mathsf{img})},z_{\mathsf{pr}}^{(\mathsf{img})})\Big){-}h_g\Big((z_{\mathsf{inv}},z_{\mathsf{dp}}^{(\mathsf{tex})},z_{\mathsf{pr}}^{(\mathsf{tex})})\Big)\right|\right|_2^2\right]=0, \\ &H\big(h_f(\boldsymbol{z}^{(\mathsf{img})})\big)=0, \\ &H\big(h_g(\boldsymbol{z}^{(\mathsf{tex})})\big)=0. \end{aligned} \tag{24}$$

The second and third terms are typically satisfied due to the uniformity to their distributions. The first term implies the modal-invariance condition by Assumption.11.

Modal Invariance of h_f, h_g. Here we prove that $h_f(\cdot)$ and $h_g(\cdot)$ are modal-invariant. Thus, given $z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}}$, for all $i \in\{1,\cdots,n_{\mathsf{img}^{(\mathsf{dp})}}\}$ and $j \in\{1,\cdots,n_{\mathsf{tex}^{(\mathsf{dp})}}\}$ resulting $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},i}^{(\mathsf{img})}}{=}0$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},i}^{(\mathsf{img})}}{=}0$, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},j}^{(\mathsf{tex})}}{=}0$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},j}^{(\mathsf{tex})}}{=}0$; and for all $i \in\{1,\cdots,n_{\mathsf{img}^{(\mathsf{pr})}}\}$ and $j \in\{1,\cdots,n_{\mathsf{tex}^{(\mathsf{pr})}}\}$, resulting $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},i}^{(\mathsf{img})}}{=}0$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},i}^{(\mathsf{img})}}{=}0$, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},j}^{(\mathsf{tex})}}{=}0$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},j}^{(\mathsf{tex})}}{=}0$. It is obvious that given $z^{(\mathsf{inv})}$ fixed, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},i}^{(\mathsf{img})}}{=}0$, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},j}^{(\mathsf{tex})}}{=}0$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},i}^{(\mathsf{img})}}{=}0$, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},j}^{(\mathsf{tex})}}{=}0$.

So we only need to prove $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},i}^{(\mathsf{img})}}{=}0$, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},j}^{(\mathsf{img})}}{=}0\ \forall\, i \in\{1,\cdots,n_{\mathsf{img}^{(\mathsf{dp})}}\}, \forall j \in\{1,\cdots,n_{\mathsf{img}^{(\mathsf{pr})}}\}$; and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},j}^{(\mathsf{tex})}}{=}0$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},i}^{(\mathsf{tex})}}{=}0\ \forall\, i \in\{1,\cdots,n_{\mathsf{img}^{(\mathsf{dp})}}\}, \forall j \in\{1,\cdots,n_{\mathsf{img}^{(\mathsf{pr})}}\}$. To simplify the proof, we consider the surrogate image variable $\boldsymbol{z}_{\mathsf{sp}}^{(\mathsf{img})} = [\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{img})};\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{img})}]$ and the surrogate text variable $\boldsymbol{z}_{\mathsf{sp}}^{(\mathsf{tex})} = [\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{tex})};\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{tex})}]$, according to the concatenation, we rewrite h_f, h_g into the equivalent forms:

$$h_f(z_{\mathsf{inv}},z_{\mathsf{sp}}^{(\mathsf{img})}) = h_f(z_{\mathsf{inv}},z_{\mathsf{dp}}^{(\mathsf{img})},z_{\mathsf{pr}}^{(\mathsf{img})}); h_g(z_{\mathsf{inv}},z_{\mathsf{sp}}^{(\mathsf{tex})}) = h_g(z_{\mathsf{inv}},z_{\mathsf{dp}}^{(\mathsf{tex})},z_{\mathsf{pr}}^{(\mathsf{tex})}). \tag{25}$$

In this way, we only need to prove

$$\begin{aligned} &\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{sp},i}^{(\mathsf{img})}}=0,\ \ \mathsf{s.t.}\ \forall i \in \{1,\cdots,n_{\mathsf{img}^{(\mathsf{dp})}}+n_{\mathsf{img}^{(\mathsf{pr})}}\}; \\ &\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{sp},j}^{(\mathsf{tex})}}=0,\ \ \mathsf{s.t.}\ \forall j \in \{1,\cdots,n_{\mathsf{tex}^{(\mathsf{dp})}}+n_{\mathsf{tex}^{(\mathsf{pr})}}\}, \end{aligned} \tag{26}$$

then the statements would be satisfied.

We first go for $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{sp},i}^{(\mathsf{img})}}{=}0$. Let seek for a contradiction that satisfies

$$\exists l \in \{1,\cdots,n_{\mathsf{img}^{(\mathsf{dp})}}+n_{\mathsf{img}^{(\mathsf{pr})}}\},\ \ (\overline{z}_{\mathsf{inv}},\overline{z}_{\mathsf{sp}}^{\mathsf{img}}) \sim p_{\boldsymbol{z}_{\mathsf{sp}}^{(\mathsf{img})}}\ \ \mathsf{s.t.}\ \ \frac{\partial h_f\Big((\overline{z}_{\mathsf{inv}},\overline{z}_{\mathsf{sp}}^{\mathsf{img}})\Big)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}} \neq 0, \tag{27}$$

thus, we assume that the partial derivative of h_f with respect to the lth image-private latent variable is non-zero at some point in the support of $p_{\boldsymbol{z}_{\mathsf{sp}}^{(\mathsf{img})}}$, *i.e.*, $\mathcal{Z}_{\mathsf{img}} = \mathcal{C} \times \mathcal{Z}_{\mathsf{img}^{\mathsf{sp}}}$ ($\mathcal{C}$ and $\mathcal{Z}_{\mathsf{img}^{\mathsf{sp}}}$ are the subspaces that represent the supports of $\boldsymbol{z}_{\mathsf{inv}}$ and $\boldsymbol{z}_{\mathsf{sp}}^{(\mathsf{img})}$). Since f and $\mathbf{f}$ are smooth, so is $h_f{=}f^* \circ \mathbf{f}$. Hence h_f has continuous (first) partial derivatives, so is h_g. To satisfy this for h_f, it must be $\frac{\partial h_f}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}} \neq 0$ in a neighborhood of $(\overline{z}_{\mathsf{inv}},\overline{z}_{\mathsf{sp}}^{\mathsf{img}})$:

$$\exists \gamma>0,\ \ \mathsf{s.t.}\ \ z_l' \mapsto h_f\Big(\overline{z}_{\mathsf{inv}},(\overline{z}_{\mathsf{sp},-\mathsf{l}}^{\mathsf{img}},z_l')\Big)\ \ \mathsf{is\ strictly\ monotonic\ on}\ \ (z_l'-\gamma,z_l'+\gamma) \tag{28}$$

where $\overline{z}_{\mathsf{sp},-l}^{\mathsf{img}} \in \mathcal{S}_{-l}$ denotes the vector of remaining the variables in $\overline{z}_{\mathsf{sp}}^{\mathsf{img}}$ except the lth variable.

From now on, we consider the $z_{\mathsf{sp}}^{(\mathsf{img})} \times z_{\mathsf{sp}}^{(\mathsf{tex})}$ defined in a sufficiently small neighborhood $\mathcal{Z}_{\mathsf{sp}}^{(\mathsf{img})} \times \mathcal{Z}_{\mathsf{sp}}^{(\mathsf{tex})}$ such that

Under the condition in Eq.28, we separately consider two cases.

Case.1: $\forall \hat{l} \in \{1, \cdots, n_{\mathsf{tex}^{(\mathsf{dp})}} + n_{\mathsf{tex}^{(\mathsf{pr})}}\}$, $\frac{\partial h_g\left(\overline{z}_{\mathsf{inv}}, \overline{z}_{\mathsf{sp}}^{\mathsf{tex}}\right)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}} = 0$. In this case, given $\overline{z}_{\mathsf{inv}}$ it holds a constant n_{inv}-dim vector $\mathbf{v}_{z_{\mathsf{inv}}} = h_g\left(z_{\mathsf{inv}}, z_{\mathsf{sp}}^{(\mathsf{tex})}\right)$ and we have

$$\mathop{\mathbb{E}}_{p_{\mathsf{mm}}}\left[\left|\left|h_f\left(\overline{z}_{\mathsf{inv}}, \overline{z}_{\mathsf{sp}}^{(\mathsf{img})}\right) - h_g\left(\overline{z}_{\mathsf{inv}}, \overline{z}_{\mathsf{sp}}^{(\mathsf{tex})}\right)\right|\right|_2^2\right] = 0 \iff \mathop{\mathbb{E}}_{p_{\mathsf{mm}}}\left[\left|\left|h_f\left(\overline{z}_{\mathsf{inv}}, \overline{z}_{\mathsf{sp}}^{(\mathsf{img})}\right) - \mathbf{v}_{\overline{z}_{\mathsf{inv}}}\right|\right|_2^2\right] = 0 \tag{29}$$

Given this, the optima of h_f with respect to $\boldsymbol{z}_{\mathsf{sp}}^{(\mathsf{img})}$ in the range of Eq.28 refers to a constant function, therefore

$$\forall \hat{z}_l \in (z'_l - \gamma, z'_l + \gamma), \quad h_f\left(\overline{z}_{\mathsf{inv}}, (\overline{z}_{\mathsf{sp},-l}^{(\mathsf{img})}, \hat{z}_l)\right) = \mathbf{v}_{\overline{z}_{\mathsf{inv}}}, \tag{30}$$

so that $\frac{\partial h_f\left(\overline{z}_{\mathsf{inv}}, (\overline{z}_{\mathsf{sp},-l}^{(\mathsf{img})}, \hat{z}_l)\right)}{\partial z_{\mathsf{dp},l}^{\mathsf{img}}} = 0$ on $(z'_l - \gamma, z'_l + \gamma)$. It violates the condition in Eq.28.

Case.2: $\exists \hat{l} \in \{1, \cdots, n_{\mathsf{tex}^{(\mathsf{dp})}} + n_{\mathsf{tex}^{(\mathsf{pr})}}\}$, $\frac{\partial h_g\left(\overline{z}_{\mathsf{inv}}, \overline{z}_{\mathsf{sp}}^{\mathsf{tex}}\right)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}} \neq 0$. In such case, we consider the auxiliary function Ω: $\mathcal{C} \times \mathcal{Z}_{\mathsf{img}^{\mathsf{sp}}} \times \mathcal{Z}_{\mathsf{tex}^{\mathsf{sp}}} \mapsto \mathbb{R}_{\geq 0}$ as follows:

$$\Omega(z_{\mathsf{inv}}, z_1, z_2) = \left| h_f\left(z_{\mathsf{inv}}, z_1\right) - h_g\left(z_{\mathsf{inv}}, z_2\right) \right| \geq 0. \tag{31}$$

Our goal is to show that Ω is strictly positive with probability greater than zero with respect to p_{mm}. Specifically, given γ defined from Eq.28, we may define $\eta(\gamma) > 0$, such that given $z''_l \in (z'_l - \gamma, z'_l)$, it holds

$$z'_{\hat{l}} \mapsto h_g\left(\overline{z}_{\mathsf{inv}}, (\overline{z}_{\mathsf{sp},-\hat{l}}^{(\mathsf{tex})}, z'_{\hat{l}})\right) \text{ is strictly monotonic on } (z'_{\hat{l}} - \eta(\gamma), z'_{\hat{l}} + \eta(\gamma)), \tag{32}$$

which is achieved due to the continunity of the first partial derivative of h_g *w.r.t.* $z_{\mathsf{sp},\hat{l}}^{(\mathsf{tex})}$. To achieve the strict positivity of Eq.31, we are going to prove $\frac{\partial h_f\left(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{img}}\right)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}} - \frac{\partial h_g\left(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{tex}}\right)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}} \neq 0$ in an open subset $\mathcal{Z}' \subset \overline{z}_{\mathsf{inv}} \times \left(\mathcal{Z}_{-l}^{\mathsf{img}^{\mathsf{sp}}} \times (z'_l - \gamma, z'_l)\right) \times \left(\mathcal{Z}_{-\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}} \times (z'_{\hat{l}} - \eta(\gamma), z'_{\hat{l}} + \eta(\gamma))\right)$ where $\mathcal{Z}_{-l}^{\mathsf{img}^{\mathsf{sp}}}$ and $\mathcal{Z}_{-\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}}$ denote the subspaces of $\mathcal{Z}_{\mathsf{img}^{\mathsf{sp}}}$ and $\mathcal{Z}_{\mathsf{tex}^{\mathsf{sp}}}$ except for the lth dimension and the $\hat{l}^{\mathsf{th}}$ dimension, respectively.

In particular, if no solution of $\frac{\partial h_f\left(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{img}}\right)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}} - \frac{\partial h_g\left(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{tex}}\right)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}} = 0$ in the range of $\overline{z}_{\mathsf{inv}} \times \left(\mathcal{Z}_{-l}^{\mathsf{img}^{\mathsf{sp}}} \times (z'_l - \gamma, z'_l)\right) \times \left(\mathcal{Z}_{-\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}} \times (z'_{\hat{l}} - \eta(\gamma), z'_{\hat{l}} + \eta(\gamma))\right)$, we know that $h_f(z_{\mathsf{inv}}, z_1) - h_g(z_{\mathsf{inv}}, z_2)$ is monotonic in the range of $\overline{z}_{\mathsf{inv}} \times \left(\mathcal{Z}_{-l}^{\mathsf{img}^{\mathsf{sp}}} \times (z'_l - \gamma, z'_l)\right) \times \left(\mathcal{Z}_{-\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}} \times (z'_{\hat{l}} - \eta(\gamma), z'_{\hat{l}} + \eta(\gamma))\right)$ due to the continunity of $\frac{\partial h_f(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{img}})}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}} - \frac{\partial h_g(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{tex}})}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}$. So we can set $\mathcal{Z}' = \overline{z}_{\mathsf{inv}} \times \left(\mathcal{Z}_{-l}^{\mathsf{img}^{\mathsf{sp}}} \times (z'_l - \gamma, z'_l)\right) \times \left(\mathcal{Z}_{-\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}} \times (z'_{\hat{l}} - \eta(\gamma), z'_{\hat{l}} + \eta(\gamma))\right)$.

On the other hand, suppose that $\overline{z}_{\mathsf{inv}} \times \left(\hat{\mathbf{z}}_{-l} \times \hat{z}_l\right) \times \left(\hat{\mathbf{z}}_{-\hat{l}} \times \hat{z}_{\hat{l}}\right) \in \overline{z}_{\mathsf{inv}} \times \left(\mathcal{Z}_{-l}^{\mathsf{img}^{\mathsf{sp}}} \times (z'_l - \gamma, z'_l)\right) \times \left(\mathcal{Z}_{-\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}} \times (z'_{\hat{l}} - \eta(\gamma), z'_{\hat{l}} + \eta(\gamma))\right)$ is a solution of $\frac{\partial h_f\left(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{img}}\right)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}} - \frac{\partial h_g\left(\overline{z}_{\mathsf{inv}}, z_{\mathsf{sp}}^{\mathsf{tex}}\right)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}} = 0$. Given this, let consider the ranges $(z'_l - \gamma, \hat{z}_l)$ and $(\hat{z}_l - \gamma, z'_l)$ to the lth dimension of $\mathcal{Z}_{\mathsf{img}^{\mathsf{dp}}}$. According to the monotonicity of h_f and the continunity of $\frac{\partial h_f}{\partial \boldsymbol{z}_{\mathsf{sp},l}^{\mathsf{img}}}$ with respect to $\boldsymbol{z}_{\mathsf{sp},l}^{\mathsf{img}}$, there is

$\mathcal{Z}_l^{(1)}(\gamma)\in\Big\{(z_l'-\gamma,\hat{z}_l),(\hat{z}_l-\gamma,z_l')\Big\}$ so that given $z_l^{(1)}\in\mathcal{Z}_l^{(1)}(\gamma)$, it holds:

$$\frac{\partial h_f\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times z_l^{(1)}\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\Big)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}}-\frac{\partial h_f\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\Big)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}}>0; \quad (33)$$

Similarly, according to the monotonicity of h_g and the continunity of $\frac{\partial h_g}{\partial \boldsymbol{z}_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}$ with respect to $\boldsymbol{z}_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}$, there is $\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma))\in\Big\{(z_{\hat{l}}'-\eta(\gamma),\hat{z}_{\hat{l}}),(\hat{z}_{\hat{l}},z_{\hat{l}}'+\eta(\gamma))\Big\}$ so that given $z_{\hat{l}}^{(2)}\in\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma))$, it holds:

$$\frac{\partial h_g\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\Big)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}-\frac{\partial h_g\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times z_{\hat{l}}^{(2)}\big)\Big)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}>0. \quad (34)$$

Combine Eq.33 and Eq.34, then we obtain

$$\begin{aligned}
&\frac{\partial h_f\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times z_l^{(1)}\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\Big)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}}-\frac{\partial h_f\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\Big)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}}\\
&+\frac{\partial h_g\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\Big)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}-\frac{\partial h_g\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times z_{\hat{l}}^{(2)}\big)\Big)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}>0\\
\Longleftrightarrow\ &\frac{\partial h_f\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times z_l^{(1)}\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\Big)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}}-\frac{\partial h_g\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times z_{\hat{l}}^{(2)}\big)\Big)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}>0\\
&\Big(\overline{z}_{\mathsf{inv}}\times\big(\hat{\mathbf{z}}_{-l}\times\hat{z}_l\big)\times\big(\hat{\mathbf{z}}_{-\hat{l}}\times\hat{z}_{\hat{l}}\big)\mathsf{\ is\ a\ solution\ of}\frac{\partial h_f\Big(\overline{z}_{\mathsf{inv}},\overline{z}_{\mathsf{dp}}^{\mathsf{img}}\Big)}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}}-\frac{\partial h_g\Big(\overline{z}_{\mathsf{inv}},z_{\mathsf{sp}}^{\mathsf{tex}}\Big)}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}=0\Big).
\end{aligned} \quad (35)$$

To this, in the range of $\mathcal{Z}'=\overline{z}_{\mathsf{inv}}\times\Big(\hat{\mathbf{z}}_{-l}\times\mathcal{Z}_l^{(1)}(\gamma)\Big)\times\Big(\hat{\mathbf{z}}_{-\hat{l}}\times\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma))\Big)\subset\overline{z}_{\mathsf{inv}}\times\Big(\mathcal{Z}_{-l}^{\mathsf{img}^{\mathsf{sp}}}\times(z_l'-\gamma,z_l')\Big)\times\Big(\mathcal{Z}_{-\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}}\times(z_{\hat{l}}'-\eta(\gamma),z_{\hat{l}}'+\eta(\gamma))\Big)$, it holds $\frac{\partial h_f(\overline{z}_{\mathsf{inv}},z_{\mathsf{sp}}^{\mathsf{img}})}{\partial z_{\mathsf{sp},l}^{\mathsf{img}}}-\frac{\partial h_g(\overline{z}_{\mathsf{inv}},z_{\mathsf{dp}}^{\mathsf{tex}}))}{\partial z_{\mathsf{sp},\hat{l}}^{\mathsf{tex}}}>0$.

Given the strict positive monotonicity of $h_f(z_{\mathsf{inv}},z_1)$–$h_g(z_{\mathsf{inv}},z_2)$ in $\mathcal{Z}'$, we consider the solution of $h_f(z_{\mathsf{inv}},z_1)$–$h_g(z_{\mathsf{inv}},z_2)=0$. If no solution, we set $\mathcal{Z}''=\mathcal{Z}'$. If there is a solution $\overline{z}_{\mathsf{inv}}\times\Big(\hat{\mathbf{z}}_{-l}\times z_l^{(3)}\Big)\times\Big(\hat{\mathbf{z}}_{-\hat{l}}\times z_{\hat{l}}^{(4)}\Big)\in\overline{z}_{\mathsf{inv}}\times\Big(\hat{\mathbf{z}}_{-l}\times\mathcal{Z}_l^{(1)}(\gamma)\Big)\times\Big(\hat{\mathbf{z}}_{-\hat{l}}\times\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma))\Big)$ with $z_l^{(3)}\in\mathcal{Z}_l^{(1)}(\gamma)$ and $z_{\hat{l}}^{(4)}\in\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma))$, we turn to consider

$$\begin{aligned}
\mathcal{Z}^{(1)}(\gamma,\eta)&:=\overline{z}_{\mathsf{inv}}\times\Big(\hat{\mathbf{z}}_{-l}\times(\inf_{\mathcal{Z}_l^{\mathsf{img}^{\mathsf{sp}}}}\mathcal{Z}_l^{(1)}(\gamma),z_l^{(3)})\Big)\times\Big(\hat{\mathbf{z}}_{-\hat{l}}\times(\inf_{\mathcal{Z}_{\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}}}\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma))\times z_{\hat{l}}^{(4)})\Big);\\
\mathcal{Z}^{(2)}(\gamma,\eta)&:=\overline{z}_{\mathsf{inv}}\times\Big(\hat{\mathbf{z}}_{-l}\times(z_l^{(3)},\sup_{\mathcal{Z}_l^{\mathsf{img}^{\mathsf{sp}}}}\mathcal{Z}_l^{(1)}(\gamma))\Big)\times\Big(\hat{\mathbf{z}}_{-\hat{l}}\times(\inf_{\mathcal{Z}_{\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}}}\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma)),z_{\hat{l}}^{(4)})\Big);\\
\mathcal{Z}^{(3)}(\gamma,\eta)&:=\overline{z}_{\mathsf{inv}}\times\Big(\hat{\mathbf{z}}_{-l}\times(z_l^{(3)},\sup_{\mathcal{Z}_l^{\mathsf{img}^{\mathsf{sp}}}}\mathcal{Z}_l^{(1)}(\gamma))\Big)\times\Big(\hat{\mathbf{z}}_{-\hat{l}}\times(z_{\hat{l}}^{(4)},\sup_{\mathcal{Z}_{\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}}}\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma)))\Big);\\
\mathcal{Z}^{(4)}(\gamma,\eta)&:=\overline{z}_{\mathsf{inv}}\times\Big(\hat{\mathbf{z}}_{-l}\times(\inf_{\mathcal{Z}_l^{\mathsf{img}^{\mathsf{sp}}}}\mathcal{Z}_l^{(1)}(\gamma),z_l^{(3)})\Big)\times\Big(\hat{\mathbf{z}}_{-\hat{l}}\times(z_{\hat{l}}^{(4)},\sup_{\mathcal{Z}_{\hat{l}}^{\mathsf{tex}^{\mathsf{sp}}}}\mathcal{Z}_{\hat{l}}^{(2)}(\eta(\gamma)))\Big).
\end{aligned} \quad (36)$$

With regards to the strict monotonicity and continunity of $h_f(z_{\mathsf{inv}},z_1)-h_g(z_{\mathsf{inv}},z_2)$, the region $\mathcal{Z}''(\gamma,\eta)\in\Big\{\mathcal{Z}^{(1)}(\gamma,\eta),\mathcal{Z}^{(2)}(\gamma,\eta),\mathcal{Z}^{(3)}(\gamma,\eta),\mathcal{Z}^{(4)}(\gamma,\eta)\Big\}$ satisfies that $\forall(z_{\mathsf{inv}},z_1,z_2)\in\mathcal{Z}''(\gamma,\eta),h_f\Big(z_{\mathsf{inv}},z_1\Big)-h_g\Big(z_{\mathsf{inv}},z_2\Big)>0$.

Therefore $\mathcal{Z}''(\gamma,\eta)$ satisfy some conditions: 1). non-empty; 2). it is an open subset in the topological subspace of $\mathcal{C}\times\mathcal{Z}_{\mathsf{img}^{\mathsf{sp}}}\times\mathcal{Z}_{\mathsf{tex}^{\mathsf{sp}}}$; 3).$\forall(z_{\mathsf{inv}},z_1,z_2)\in\mathcal{Z}''(\gamma,\eta)$, it holds $\Omega(z_{\mathsf{inv}},z_1,z_2)=$

$\left|h_f((z_{\mathsf{inv}}, z_1) - h_g(z_{\mathsf{inv}}, z_2)\right| > 0$; 4). it is fully supported with respect to p_{mm} generated by Assumption.11. As a consequence,

$$p_{\mathsf{mm}}\Big(\Omega(z_{\mathsf{inv}}, z_1, z_2) > 0\Big) > p_{\mathsf{mm}}\Big(\mathcal{Z}''(\gamma, \eta)\Big) > 0. \tag{37}$$

So it leads to

$$\begin{aligned} &\mathop{\mathbb{E}}_{p_{\mathsf{mm}}}\left[\left|\left|h_f\Big((z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{img})}, z_{\mathsf{pr}}^{(\mathsf{img})})\Big) - h_g\Big((z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{tex})}, z_{\mathsf{pr}}^{(\mathsf{tex})})\Big)\right|\right|_2^2\right] \\ \geq &\mathop{\mathbb{E}}_{p_{\mathsf{mm}}\left(\mathcal{Z}''(\gamma,\eta)\right)}\left[\left|\left|\Omega(z_{\mathsf{inv}}, z_{\mathsf{sp}}^{(\mathsf{img})}, z_{\mathsf{sp}}^{(\mathsf{tex})})\right|\right|_2^2\right] \\ >&0. \end{aligned} \tag{38}$$

It results in the contradiction to Eq.24 .

Concluding Case.1 and Case.2, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{sp},i}^{(\mathsf{img})}}=0$ is proved, thus, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},i}^{(\mathsf{img})}}=0$ and $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},i}^{(\mathsf{img})}}=0$ have been proven. In terms of the symmetry between $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{sp},i}^{(\mathsf{img})}}=0$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{sp},j}^{(\mathsf{tex})}}=0$ as well as the generative processes between $\boldsymbol{x}^{(\mathsf{img})}$ and $\boldsymbol{x}^{(\mathsf{tex})}$, we may follow the same rountine to prove $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{sp},j}^{(\mathsf{tex})}}=0$, thus, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{pr},j}^{(\mathsf{tex})}}=0$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z_{\mathsf{dp},j}^{(\mathsf{tex})}}=0$ have also been proven (skipped for simplicity).

To this end, we have restricted h_f and h_g taking value in $\mathcal{C}$, thus, $h_f = f^* \circ \mathbf{f}_{1:n_{\mathsf{inv}}}$ and $h_g = g^* \circ \mathbf{g}_{1:n_{\mathsf{inv}}}$.

Invertability of h_f, h_g. We derive the proof of step.3 from proving the theorem 4.4 in Von Kügelgen et al. (2021), in order to justify the invertability of h_f, h_g. Specifically, we introduce a lemma from Zimmermann et al. (2021)

Lemma 14. *(Proposition 5 of Zimmermann et al. (2021)) Let $\mathcal{M}$, $\mathcal{N}$ be simply connected and oriented C^1 manifolds without boundaries and $h : \mathcal{M} \mapsto \mathcal{N}$ be a differentiable map. Further, let the random variable $z \in \mathcal{M}$ be distributed according to $z \sim p(z)$ for a regular density function p,* i.e.*, $0 < p < +\infty$. If the pushforward $p\sharp h(z)$ of p through h is also a regular density, i.e., $0 < p\sharp h(z) < \infty$, then h is a bijection.*

We apply this result to "the simply connected and oriented C^1 manifolds without boundaries" by setting $\mathcal{M}=\mathcal{C}$ and $\mathcal{N}=(0,1)^{n_{\mathsf{inv}}}$. In terms of the smoothness of h_f and h_g, they are differentiable maps so that both satisfy h in the lemma by mapping the random variable $\boldsymbol{z}_{\mathsf{inv}}$ into a uniform random variable. Notice that $p_{\boldsymbol{z}_{\mathsf{inv}}}$ (Assumption.11) and the uniform distribution (the pushforward of $p_{\boldsymbol{z}_{\mathsf{inv}}}$) are regular densities in the sense of Lemma.14, therefore h_f and h_g are bijective maps, *i.e.*, invertable.

B2. Proof of Corollary.10

Here we derive the formal version of Corollary.10 with its proof:

Proposition 15. *For fixed $\gamma > 0$, as the number of negative samples $K - 1 \to \infty$, the (normalized) $\mathcal{L}_{\mathsf{InfoNCE}}^{(\mathsf{img}\to\mathsf{tex})} - \log(K-1)$ and $\mathcal{L}_{\mathsf{InfoNCE}}^{(\mathsf{tex}\to\mathsf{img})} - \log(K-1)$ converge to*

$$\begin{aligned} &-\frac{1}{\gamma}\mathbb{E}_{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})}\rangle \sim p_{\mathsf{mm}}}\Big(f(x^{(\mathsf{img})})^\top g(x^{(\mathsf{tex})})\Big) \\ &\qquad + \mathbb{E}_{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})}\rangle \sim p_{\mathsf{mm}}}\left(\log \mathbb{E}_{\hat{x}^{(\mathsf{tex})} \sim p(\boldsymbol{x}^{(\mathsf{tex})})}\left[e^{f(x^{(\mathsf{img})})^\top g(\hat{x}^{(\mathsf{tex})})/\gamma}\right]\right); \end{aligned} \tag{39}$$

$$\begin{aligned} &-\frac{1}{\gamma}\mathbb{E}_{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})}\rangle \sim p_{\mathsf{mm}}}\Big(f(x^{(\mathsf{img})})^\top g(x^{(\mathsf{tex})})\Big) \\ &\qquad + \mathbb{E}_{\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})}\rangle \sim p_{\mathsf{mm}}}\left(\log \mathbb{E}_{\hat{x}^{(\mathsf{img})} \sim p(\boldsymbol{x}^{(\mathsf{img})})}\left[e^{f(\hat{x}^{(\mathsf{img})})^\top g(x^{(\mathsf{tex})})/\gamma}\right]\right), \end{aligned} \tag{40}$$

respectively, with the following results:

1. *The first terms of Eq.39 and 40, are minimized iff f, g are perfectly aligned,* i.e., $||f(x^{(\text{img})})$-$g(x^{(\text{tex})})||_2^2 \to 0$ *across image-text pairs* $\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}}$.
2. *With perfectly aligned image-text feature pairs extracted from f,g, the second terms of Eq.39 and 40 refer to the resubstitution entropy estimators with respect to von Mises-Fisher (vMF) kernel density estimation.*

Proof. Consider each image-text pair $\langle x^{(\text{img})}, x^{(\text{tex})} \rangle$ with $K-1$ images $\{\hat{x}_k^{(\text{img})}\}_{k=1}^{K-1}$ to construct $K-1$ negative pairs with $x^{(\text{tex})}$ in $\mathcal{L}_{\text{InfoNCE}}^{(\text{img}\to\text{tex})}$, and with $K-1$ texts $\{\hat{x}_k^{(\text{tex})}\}_{k=1}^{K-1}$ to construct $K-1$ negative pairs with $x^{(\text{img})}$ in $\mathcal{L}_{\text{InfoNCE}}^{(\text{tex}\to\text{img})}$. Note that $x^{(\text{img})} \overset{\text{i.i.d.}}{\sim} p_{\text{mm}}(\boldsymbol{x}^{(\text{img})})$; $\{\hat{x}_k^{(\text{img})}\}_{k=1}^{K-1} \overset{\text{i.i.d.}}{\sim} p_{\text{mm}}(\boldsymbol{x}^{(\text{img})})$, therefore we have

$$\lim_{K-1\to\infty} \log\left(\frac{e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma}}{K-1} + \sum_{k=1}^{K-1}\frac{e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma}}{K-1}\right) = \log \mathbb{E}_{\hat{x}^{(\text{tex})}\sim p(\boldsymbol{x}^{(\text{tex})})}\left[e^{f(x^{(\text{img})})^\top g(\hat{x}^{(\text{tex})})/\gamma}\right] \tag{41}$$

with the strong law of large numbers and the Continuous Mapping Theorem. Eq.41 results in

$$\begin{aligned}
&\mathcal{L}_{\text{InfoNCE}}^{(\text{img}\to\text{tex})} - \log(K-1) \\
=&\mathbb{E}_{\substack{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}} \\ \langle x^{(\text{img})}, \hat{x}_k^{(\text{tex})}\rangle \nsim p_{\text{mm}}}} - \log \frac{(K-1)e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma}}{e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma} + \sum_{k=1}^{K-1} e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma}} \\
=&\mathbb{E}_{\substack{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}} \\ \langle x^{(\text{img})}, \hat{x}_k^{(\text{tex})}\rangle \nsim p_{\text{mm}}}} \left(- \frac{f(x^{(\text{img})})^\top g(x^{(\text{tex})})}{\gamma} + \log \frac{\left(e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma} + \sum_{k=1}^{K-1} e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma}\right)}{K-1}\right) \\
=& -\frac{1}{\gamma}\mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}}}\left(f(x^{(\text{img})})^\top g(x^{(\text{tex})})\right) \\
&\qquad + \mathbb{E}_{\substack{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}} \\ \langle x^{(\text{img})}, \hat{x}_k^{(\text{tex})}\rangle \nsim p_{\text{mm}}}} \left(\log \frac{\left(e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma} + \sum_{k=1}^{K-1} e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma}\right)}{K-1}\right)
\end{aligned} \tag{42}$$

So

$$\begin{aligned}
&\lim_{K-1\to\infty} \mathcal{L}_{\text{InfoNCE}}^{(\text{img}\to\text{tex})} - \log(K-1) \\
=& -\frac{1}{\gamma}\mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}}}\left(f(x^{(\text{img})})^\top g(x^{(\text{tex})})\right) \\
&\qquad + \mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}}}\left(\log \mathbb{E}_{\hat{x}^{(\text{tex})}\sim p(\boldsymbol{x}^{(\text{tex})})}\left[e^{f(x^{(\text{img})})^\top g(\hat{x}^{(\text{tex})})/\gamma}\right]\right)
\end{aligned} \tag{43}$$

Similarly we obtain

$$\begin{aligned}
&\lim_{K-1\to\infty} \mathcal{L}_{\text{InfoNCE}}^{(\text{tex}\to\text{img})} - \log(K-1) \\
=& -\frac{1}{\gamma}\mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}}}\left(f(x^{(\text{img})})^\top g(x^{(\text{tex})})\right) \\
&\qquad + \mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})}\rangle \sim p_{\text{mm}}}\left(\log \mathbb{E}_{\hat{x}^{(\text{img})}\sim p(\boldsymbol{x}^{(\text{img})})}\left[e^{f(\hat{x}^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma}\right]\right)
\end{aligned} \tag{44}$$

So the main result has been proven.

Here we turn to prove the two statements based on the main result:

1. Note that $||f(x^{(\text{img})})$-$g(x^{(\text{tex})})||_2^2 = 1 - f(x^{(\text{img})})^\top g(x^{(\text{tex})})$. The minimization in the first term in Eq.39 and Eq.40 is equivalent with $||f(x^{(\text{img})})$-$g(x^{(\text{tex})})||_2^2 = 0$.
2. With perfectly aligned image-text feature pairs extracted from f,g, as we known the pair drawn from p_{mm}, it holds $p_{\text{mm}} = p_{\boldsymbol{x}^{(\text{img})}}\delta\Big(f(\boldsymbol{x}^{(\text{img})})$-$g(\boldsymbol{x}^{(\text{tex})})\Big) = p_{\boldsymbol{x}^{(\text{tex})}}\delta\Big(f(\boldsymbol{x}^{(\text{img})})$-

$g(\boldsymbol{x}^{(\mathsf{tex})})\Big)$. Therefore we have

$$\begin{aligned}
&\mathbb{E}_{\langle x^{(\mathsf{img})},x^{(\mathsf{tex})}\rangle\sim p_{\mathsf{mm}}}\left(\log \mathbb{E}_{\hat{x}^{(\mathsf{tex})}\sim p(\boldsymbol{x}^{(\mathsf{tex})})}\left[e^{f(x^{(\mathsf{img})})^\top g(\hat{x}^{(\mathsf{tex})})/\gamma}\right]\right)\\
=&\mathbb{E}_{\substack{x^{(\mathsf{tex})}\sim p(\boldsymbol{x}^{(\mathsf{tex})})\\ x^{(\mathsf{img})}\sim\delta\left(f(\boldsymbol{x}^{(\mathsf{img})})-g(x^{(\mathsf{tex})})\right)}}\left(\log \mathbb{E}_{\hat{x}^{(\mathsf{tex})}\sim p(\boldsymbol{x}^{(\mathsf{tex})})}\left[e^{f(x^{(\mathsf{img})})^\top g(\hat{x}^{(\mathsf{tex})})/\gamma}\right]\right)\\
=&\mathbb{E}_{x^{(\mathsf{tex})}\sim p(\boldsymbol{x}^{(\mathsf{tex})})}\left(\log \mathbb{E}_{\hat{x}^{(\mathsf{tex})}\sim p(\boldsymbol{x}^{(\mathsf{tex})})}\left[e^{g(x^{(\mathsf{tex})})^\top g(\hat{x}^{(\mathsf{tex})})/\gamma}\right]\right)\\
\triangleq&\frac{1}{N}\sum_{i=1}^{N}\log\Big(\frac{1}{N}\sum_{i=1}^{N}\left[e^{g(x_i^{(\mathsf{tex})})^\top g(x_j^{(\mathsf{tex})})/\gamma}\right]\Big)\\
=&\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}_{\mathsf{vMF-KDE}}\Big(g(x_i^{(\mathsf{tex})})\Big)+\log Z_{\mathsf{vMF}}\\
\triangleq&-H\Big(g(\boldsymbol{x}^{(\mathsf{tex})})\Big)+\log Z_{\mathsf{vMF}}
\end{aligned}\tag{45}$$

where $H\Big(f(\boldsymbol{x}^{(\mathsf{tex})})\Big)$ implies the the resubstitution entropy estimator with respect to von Mises-Fisher (vMF) kernel density estimation (KDE) based on N samples that constructs a vMF kernel with $\kappa=\gamma^{-1}$; Z_{vMF} denotes the normalization constant for vMF distribution with $\kappa=\gamma^{-1}$. Using the same proof technique, we also obtain

$$\mathbb{E}_{\langle x^{(\mathsf{img})},x^{(\mathsf{tex})}\rangle\sim p_{\mathsf{mm}}}\left(\log \mathbb{E}_{\hat{x}^{(\mathsf{img})}\sim p(\boldsymbol{x}^{(\mathsf{img})})}\left[e^{f(\hat{x}^{(\mathsf{img})})^\top g(x^{(\mathsf{tex})})/\gamma}\right]\right)\triangleq -H\big(f(\boldsymbol{x}^{(\mathsf{img})})\big)+\log Z_{\mathsf{vMF}}\tag{46}$$

The proposition has been proven. □

B3. Proof of Theorem.1

Similar with the proof of Theorem.9, our proof of Theorem.1 can be also divided into three steps: 1). construct the optimal f^*, g^* to fulfill the objectives, further leading to h_f, h_g for their decomposition; 2), h_f, h_g are modality-invariant with respect to any modality-specific features (only recover the modal-invariant partitions of the inverses $\mathbf{f}^{-1}$, $\mathbf{g}_i^{-1}$, $\forall i\in\{1,\cdots,\mathsf{k}_{\mathsf{max}}\}$); 3). Verify the invertability of h_f, h_g to fulfill the function decomposition.

Construction of h_f, h_g. Let first consider f. It is easy to observe that the image generation in Assumption.12 is consistent with Assumption.11, it leads to the same construction process of f^* and h_f in the proof of Theorem.9 .

Here we turn to g^* and h_g. Observe that g is defined on the union of $\mathsf{k}_{\mathsf{max}}$ real-value matrix spaces $\{\mathcal{X}_{\mathsf{tex}}^{(k)}\}_{k=1}^{\mathsf{k}_{\mathsf{max}}}$ where the kth space $\mathcal{X}_{\mathsf{tex}}^{(k)}$ indicates the sentence matrix with k token columns ($k\leq\mathsf{k}_{\mathsf{max}}$) and can be decomposed by token spaces, *i.e.*, $\mathcal{X}_{\mathsf{tex}}^{(k)}=\mathcal{T}_{\mathsf{tex}}^{(1)}\times\cdots\times\mathcal{T}_{\mathsf{tex}}^{(k)}$. Since $\mathbf{g}$ is a diffeomorphism on generated sentence matrices, therefore $\forall k\in\mathsf{k}_{\mathsf{max}}$, there must exist a manifold $\mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)})$ and a function $\mathbf{g}^{(k)}$ derived from $\mathbf{g}$, which satisfies $\mathbf{g}^{(k)}$: $\mathcal{C}_{\mathsf{inv}}\times\Big(\mathcal{S}_1\times\cdots\mathcal{S}_k\Big)\times\mathcal{S}_{\mathsf{pr}}^{\mathsf{tex}}\to\mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)})$ is smooth and invertable with respect to the generation of k-length sentence matrices, where $\mathcal{S}_k$ ($\forall k\in\{1,\cdots,\mathsf{k}_{\mathsf{max}}\}$) indicates latent feature spaces with respect to the text-dependent variable $\boldsymbol{z}_k^{(\mathsf{tex})}$, and $\mathcal{S}_{\mathsf{pr}}^{\mathsf{tex}}$ indicates the text-private feature space with dimension $n_{\mathsf{pr}}^{(\mathsf{tex})}$. Note that, $\mathbf{g}^{(k)}$ is represented by $\{\mathbf{g}_i\}_{i=1}^{k}$:

$$\begin{aligned}
&\forall\big(z_{\mathsf{inv}},z_1^{(\mathsf{tex})},\cdots,z_k^{(\mathsf{tex})},z_{\mathsf{pr}}^{(\mathsf{tex})}\big)\in\mathcal{C}_{\mathsf{inv}}\times\Big(\mathcal{S}_1\times\cdots\mathcal{S}_k\Big)\times\mathcal{S}_{\mathsf{pr}}^{\mathsf{tex}}\\
&\qquad\mathbf{g}^{(k)}\Big(z_{\mathsf{inv}},z_1^{(\mathsf{tex})},\cdots,z_k^{(\mathsf{tex})},z_{\mathsf{pr}}^{(\mathsf{tex})}\Big)=\Big[\mathbf{g}_1\big(z_{\mathsf{inv}},z_1^{(\mathsf{tex})},z_{\mathsf{pr}}^{(\mathsf{tex})}\big),\cdots,\mathbf{g}_k\big(z_{\mathsf{inv}},\{z_j^{(\mathsf{tex})}\}_{j=1}^{k},z_{\mathsf{pr}}^{(\mathsf{tex})}\big)\Big].
\end{aligned}\tag{47}$$

Based on the condition, $\mathbf{g}^{(k)}$ holds its smooth inverse $(\mathbf{g}^{(k)})^{-1}$ such that

$$(\mathbf{g}^{(k)})^{-1}\Big(\big[\mathbf{g}_1\big(z_{\mathsf{inv}}, z_1^{(\mathsf{tex})}, z_{\mathsf{pr}}^{(\mathsf{tex})}\big), \cdots, \mathbf{g}_k\big(z_{\mathsf{inv}}, \{z_j^{(\mathsf{tex})}\}_{j=1}^k, z_{\mathsf{pr}}^{(\mathsf{tex})}\big)\big]\Big) = \big(z_{\mathsf{inv}}, z_1^{(\mathsf{tex})}, \cdots, z_k^{(\mathsf{tex})}, z_{\mathsf{pr}}^{(\mathsf{tex})}\big). \tag{48}$$

Hence for each X^{tex} generated by Assumption.12, *i.e.*, $X^{\mathsf{tex}} = \big[\mathbf{g}_1\big(z_{\mathsf{inv}}, z_1^{(\mathsf{tex})}, z_{\mathsf{pr}}^{(\mathsf{tex})}\big), \cdots, \mathbf{g}_k\big(z_{\mathsf{inv}}, \{z_j^{(\mathsf{tex})}\}_{j=1}^k, z_{\mathsf{pr}}^{(\mathsf{tex})}\big)\big]$ ($\forall k \in \{1, \cdots, \mathsf{k}_{\max}\}$), we can restrict their outputs in the first n_{inv} dimensions such that $(\mathbf{g}^{(k)})_{1:n_{\mathsf{inv}}}^{-1}(X^{(\mathsf{tex})}) = z^{(\mathsf{inv})}$. Then we employ the same Damois construction technique used in first step of the proof in Theorem.9 to define the function $\mathbf{d} : \mathcal{C}_{\mathsf{inv}} \to (0,1)^{n_{\mathsf{inv}}}$ that map $\boldsymbol{z}_{\mathsf{inv}}$ into a uniform random variable. Derived from such construction, $\mathbf{d}(\boldsymbol{z}^{(\mathsf{inv})})$ is uniformly distributed on $(0,1)^{n_{\mathsf{inv}}}$Darmois (1951), and is also smooth due to the third primise.

Note that given $\forall k \in \{1, \cdots, \mathsf{k}_{\max}\}$, it exists a manifold support $\mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)})$ derived to construct $(\mathbf{g}^{(k)})_{1:n_{\mathsf{inv}}}^{-1}$ and $\mathbf{d}$, where $\mathbf{d}$ are regardless of i. Since $\forall i_1, i_2$ $(i_1 \neq i_2)$, $\mathcal{M}^{(i_2)}(\mathcal{X}_{\mathsf{tex}}^{(i_2)}) \cap \mathcal{M}^{(i_1)}(\mathcal{X}_{\mathsf{tex}}^{(i_1)}) = \emptyset$, we can define a piecewise composite function g^* on $\cup_i^{\mathsf{k}_{\max}} \mathcal{M}^{(i)}(\mathcal{X}_{\mathsf{tex}}^{(i)})$:

$$g^*(X^{(\mathsf{tex})}) := \mathbf{d} \circ (\mathbf{g}^{(k)})_{1:n_{\mathsf{inv}}}^{-1}(X^{(\mathsf{tex})}) \text{ if. } X^{(\mathsf{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)}), \forall i \in \{1, \cdots, \mathsf{k}_{\max}\}, \tag{49}$$

which is smooth on each sub-manifold. Given this, we consider the following derivation:

$$\begin{aligned}
\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f^*, g^*) &= \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}}} \left[\left|\left| f^*(x^{(\mathsf{img})}) - g^*(X^{(\mathsf{tex})}) \right|\right|_2^2\right] - H(f^*(\boldsymbol{x}^{(\mathsf{img})})) - H\Big(g^*(\mathbf{X}^{(\mathsf{tex})})\Big) \\
&= \sum_{k=1}^{\mathsf{k}_{\max}} p(\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k) \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}(|\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k)}} \left[\left|\left| f^*(x^{(\mathsf{img})}) - g^*(X^{(\mathsf{tex})}) \right|\right|_2^2\right] \\
&\qquad - H(f^*(\boldsymbol{x}^{(\mathsf{img})})) - H\Big(g^*(\mathbf{X}^{(\mathsf{tex})})\Big) \\
&= \sum_{k=1}^{\mathsf{k}_{\max}} p(\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k) \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}(|\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k)}} \left[\left|\left| f^*(x^{(\mathsf{img})}) - \mathbf{d} \circ \big(\mathbf{g}^{(k)}\big)_{1:n_{\mathsf{inv}}}^{-1}(X^{(\mathsf{tex})}) \right|\right|_2^2\right] \\
&\qquad - H(f^*(\boldsymbol{x}^{(\mathsf{img})})) - H\Big(g^*(\mathbf{X}^{(\mathsf{tex})})\Big) \\
&= \sum_{k=1}^{\mathsf{k}_{\max}} p(\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k) \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}(|\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k)}} \left[\left|\left| \mathbf{d}(z_{(\mathsf{inv})}) - \mathbf{d}(z_{(\mathsf{inv})}) \right|\right|_2^2\right] \\
&\qquad - H\Big(\mathbf{d}(z_{(\mathsf{inv})})\Big) - H\Big(\mathbf{d}(z_{(\mathsf{inv})})\Big) \\
&= 0,
\end{aligned} \tag{50}$$

where $p\big(\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k\big)$ indicates the proportion that the number of column $X^{(\mathsf{tex})}$ equals to k. Consider f^*, g^* that satisfy

$$\mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f^*, g^*) = \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}}} \left[\left|\left| f^*(x^{(\mathsf{img})}) - g^*(X^{(\mathsf{tex})}) \right|\right|_2^2\right] - H(f^*(\boldsymbol{x}^{(\mathsf{img})})) - H\Big(g^*(\mathbf{X}^{(\mathsf{tex})})\Big), \tag{51}$$

which we take to define $h_f = f^* \circ \mathbf{f}$ and the piecewise function

$$h_g(z) = g^* \circ \mathbf{g}^{(k)}(z), \text{ if } z \in \mathcal{C}_{\mathsf{inv}} \times \Big(\mathcal{S}_1 \times \cdots \mathcal{S}_k\Big) \times \mathcal{S}_{\mathsf{pr}}, \ \forall k \in \{1, \cdots, \mathsf{k}_{\max}\}. \tag{52}$$

In terms of Eq.6, the formulation above implies h_f, h_g with

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p_{\mathsf{mm}}} \left[\left|\left| h_f(z^{(\mathsf{img})}) - h_g(z^{(\mathsf{tex})}) \right|\right|_2^2\right] = 0 \\
&= \sum_{k=1}^{\mathsf{k}_{\max}} p(\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k) \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})} \rangle \\ \sim p_{\mathsf{mm}}(|\dim_{\mathsf{col}}(X^{(\mathsf{tex})}) = k)}} \left[\left|\left| h_f(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{img})}, z_{\mathsf{pr}}^{(\mathsf{img})}) - h_g(z_{\mathsf{inv}}, \{z_j^{(\mathsf{tex})}\}_{j=1}^i, z_{\mathsf{pr}}^{(\mathsf{tex})}) \right|\right|_2^2\right] = 0, \\
&H\big(h_f(\boldsymbol{z}^{(\mathsf{img})})\big) = 0, \quad H\big(h_g(\mathbf{X}^{(\mathsf{tex})})\big) = 0.
\end{aligned} \tag{53}$$

The second and third terms are typically satisfied due to the uniformity to their distributions. The first term implies the modal-invariance condition by Assumption.12 .

Modal Invariance of h_f, h_g. Here we prove that $h_f(\cdot)$ and $h_g(\cdot)$ are modal-invariant. Since h_f is consistent with Theorem.9, it satisfies that given $z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}}$, for all $i \in\{1, \cdots, n_{\mathsf{img}^{(\mathsf{dp})}}\}$ and $j \in\{1, \cdots, n_{\mathsf{img}^{(\mathsf{pr})}}\}$, it results in $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{img})}_{\mathsf{dp},i}}$=0, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{img})}_{\mathsf{pr},j}}$=0 ; for all $i \in\{1, \cdots, n_{\mathsf{tex}^{(\mathsf{dp})}}\}$ and $j \in\{1, \cdots, n_{\mathsf{tex}^{(\mathsf{pr})}}\}$, it results in $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},j}}$=0, $\frac{\partial h_f(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{dp},i}}$=0. They are consistent with the proof of Theorem.9 .

Here we consider the modal invariant property of h_g. Note that $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{img})}_{\mathsf{pr},i}}$=0 and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{img})}_{\mathsf{dp},i}}$=0 are also obviously satisfied given z_{inv} fixed. To this end, we only need to prove that $\forall k \in \{1, \cdots, \mathsf{k}_{\max}\}$, in terms of $h_g(\cdot) = g^* \circ \mathbf{g}^{(k)}(\cdot)$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k,i}}$=0 for all $i \in \{1, \cdots, n^{(\mathsf{tex})}_k\}$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},j}}$=0 ($\forall j\in\{1, \cdots, n^{(\mathsf{tex})}_{\mathsf{pr}}\}$).

When $k = 1$, it can be reduced to prove $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{dp},j}}$=0 and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},j}}$=0 in the proof of Theorem.9, so it is satisfied. Regarding this as the first step, we construct a mathematical induction procedure to prove the rest.

Specifically, suppose that $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k,l}}$=0 ($\forall l \in \{1, \cdots, n^{(\mathsf{tex})}_k\}$) and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}}$=0 ($\forall \hat{j}\in \{1, \cdots, n^{(\mathsf{tex})}_{\mathsf{pr}}\}$) with

$$h_g(\cdot) = g^* \circ \mathbf{g}^{(k)}(\cdot) = g^* \circ \big([\mathbf{g}_1(\cdot), \mathbf{g}_2(\cdot), \cdots, \mathbf{g}_k(\cdot)]\big), \tag{54}$$

thus,

$$\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = \frac{\partial\Big(g^* \circ \big([\mathbf{g}_1(z_{\mathsf{inv}}, \cdot), \cdots, \mathbf{g}_k(z_{\mathsf{inv}}, \cdot)]\big)\Big)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = \sum_{i=1}^{k}\sum_{i'=1}^{m} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}}, \cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}}, \cdot)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = 0 \tag{55}$$

and

$$\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k,l}} = \frac{\partial\Big(g^* \circ \big([\mathbf{g}_1(z_{\mathsf{inv}}, \cdot), \cdots, \mathbf{g}_k(z_{\mathsf{inv}}, \cdot)]\big)\Big)}{\partial z^{(\mathsf{tex})}_{k,l}} = \sum_{i=1}^{k}\sum_{i'=1}^{m} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}}, \cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}}, \cdot)}{\partial z^{(\mathsf{tex})}_{k,l}} = 0. \tag{56}$$

where $\mathbf{g}_{i,i'}(\cdot)$ indicates the function output of i'-th element with respect to the i-th token embedding. Given this, we first prove $\forall k' < k$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l'}}$=0 ($\forall l' \in \{1, \cdots, n^{(\mathsf{tex})}_{k'}\}$).

Let's begin by $k' = k-1$. It is obvious that $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l'}}$=$\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k-1,l'}}$=$\sum_{l=1}^{n^{(\mathsf{tex})}_k}$ $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k,l}} \frac{\partial z^{(\mathsf{tex})}_{k,l}}{\partial z^{(\mathsf{tex})}_{k-1,l'}}$=0 (since $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k,l}}$=0 for all $l \in \{1, \cdots, n^{(\mathsf{tex})}_k\}$). Similarly, for $k' = k-2$, it also holds

$$\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l'}} = \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k-2,l'}} = \sum_{l=1}^{n^{(\mathsf{tex})}_{k-1}} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k-1,l}} \frac{\partial z^{(\mathsf{tex})}_{k-1,l}}{\partial z^{(\mathsf{tex})}_{k-2,l'}} + \sum_{l=1}^{n^{(\mathsf{tex})}_{k}} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k,l}} \frac{\partial z^{(\mathsf{tex})}_{k,l}}{\partial z^{(\mathsf{tex})}_{k-2,l'}} = 0. \tag{57}$$

which is fulfilled because $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k'',l}}$=0, $\forall\, l \in \{1, \cdots, n^{(\mathsf{tex})}_{k''}\}$, $\forall\, k'' \in \{k-1, k-2\}$. Follow this induction chain, $\forall k' < k$, it holds the decomposition as

$$\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l'}} = \sum_{t=1}^{k-k'} \sum_{l=1}^{n^{(\mathsf{tex})}_{k-t+1}} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k-t+1,l}} \frac{\partial z^{(\mathsf{tex})}_{k-t+1,l}}{\partial z^{(\mathsf{tex})}_{k',l'}} \tag{58}$$

with $\forall k'' \in \{k', \cdots, k\}$, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k'',l}} = 0$. So $\sum_{t=1}^{k-k'} \sum_{l=1}^{n^{(\mathsf{tex})}_{k-t+1}} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k-t+1,l}} \frac{\partial z^{(\mathsf{tex})}_{k-t+1,l}}{\partial z^{(\mathsf{tex})}_{k',l}} = 0$ and we have $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l}}$=0 ($\forall l \in \{1, \cdots, n^{(\mathsf{tex})}_{k'}\}$, $\forall k' \in \{1, \cdots, k\}$) and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}}$=0 ($\forall \hat{j}\in \{1, \cdots, n^{(\mathsf{tex})}_{\mathsf{pr}}\}$.

Following the mathematical induction rule, we turn to the case with $k+1$ in $h_g(\cdot) = g^* \circ \mathbf{g}^{(k+1)}(\cdot) = g^* \circ \big([\mathbf{g}^{(k)}(\cdot), \mathbf{g}_{k+1}(\cdot)]\big)$, then attempt to prove $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l}}=0$ $(\forall l \in \{1,\cdots,n^{(\mathsf{tex})}_{k'}\}, \forall k' \in \{1,\cdots,k+1\})$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}}=0$ $(\forall \hat{j} \in \{1,\cdots,n^{(\mathsf{tex})}_{\mathsf{pr}}\}$. Ought to be noted that, if $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k+1,l}}=0$ $(\forall l \in \{1,\cdots, n^{(\mathsf{tex})}_{k+1}\})$ is satisfied, we can take the similar induction above to verify $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l}}=0$ $(\forall l \in \{1,\cdots, n^{(\mathsf{tex})}_{k'}\}, \forall k' \in \{1,\cdots,k+1\})$. So we only need to prove $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k+1,l}}=0$ $(\forall l \in \{1,\cdots, n^{(\mathsf{tex})}_{k+1}\})$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}}=0$ $(\forall \hat{j} \in \{1,\cdots,n^{(\mathsf{tex})}_{\mathsf{pr}}\}$. Observe that

$$\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = \sum_{i=1}^{k+1}\sum_{i'=1}^{m} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}},\cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}},\cdot)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = \sum_{i'=1}^{m} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial \mathbf{g}_{k+1,i'}(z_{\mathsf{inv}},\cdot)} \frac{\partial \mathbf{g}_{k+1,i'}(z_{\mathsf{inv}},\cdot)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} \tag{59}$$

and

$$\begin{aligned}\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k+1,l}} &= \sum_{i=1}^{k}\sum_{i'=1}^{m}\sum_{k'=1}^{k} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}},\cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\mathsf{inv}},\cdot)}{\partial z^{(\mathsf{tex})}_{k',l}} \frac{\partial z^{(\mathsf{tex})}_{k',l}}{\partial z^{(\mathsf{tex})}_{k+1,l}} + \sum_{i'=1}^{m} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial \mathbf{g}_{k+1,i'}(z_{\mathsf{inv}},\cdot)} \frac{\partial \mathbf{g}_{k+1,i'}(z_{\mathsf{inv}},\cdot)}{\partial z^{(\mathsf{tex})}_{k+1,l}} \\ &= \sum_{i'=1}^{m} \frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial \mathbf{g}_{k+1,i'}(z_{\mathsf{inv}},\cdot)} \frac{\partial \mathbf{g}_{k+1,i'}(z_{\mathsf{inv}},\cdot)}{\partial z^{(\mathsf{tex})}_{k+1,l}},\end{aligned} \tag{60}$$

where only the $(k+1)$-th token output $\mathbf{g}_{k+1}\big(z_{\mathsf{inv}}, \{z^{(\mathsf{tex})}_i\}_{i=1}^{k+1}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big)$ influence the derivatives with respect to $z^{(\mathsf{tex})}_{k+1,l}$ and $z^{(\mathsf{tex})}_{k+1,l}$. To this, $\forall i \in \{1,\cdots,k\}$, suppose $\overline{z}^{(\mathsf{tex})}_i \sim p_{\boldsymbol{z}^{(\mathsf{tex})}_i}$ drawn through the generative process based on Assumption.12, given that $\overline{Z}_k = \{\overline{z}^{(\mathsf{tex})}_i\}_{i=1}^{k}$ is fixed, we consider the surrogate function family

$$\begin{aligned}&h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k) = g^* \circ \mathbf{g}^{(k+1)}\Big(z_{\mathsf{inv}}, \overline{z}^{(\mathsf{tex})}_1, \cdots, \overline{z}^{(\mathsf{tex})}_k, z^{(\mathsf{tex})}_{k+1}, z^{(\mathsf{tex})}_{\mathsf{pr}}\Big) \\ =&g^* \circ \Big(\Big[\mathbf{g}_1\big(z_{\mathsf{inv}}, \overline{z}^{(\mathsf{tex})}_1, z^{(\mathsf{tex})}_{\mathsf{pr}}\big), \cdots, \mathbf{g}_k\big(z_{\mathsf{inv}}, \{\overline{z}^{(\mathsf{tex})}_j\}_{j=1}^{k}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big), \mathbf{g}_{k+1}\big(z_{\mathsf{inv}}, \{\overline{z}^{(\mathsf{tex})}_j, z^{(\mathsf{tex})}_{k+1}\}_{j=1}^{k}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big)\Big]\Big).\end{aligned} \tag{61}$$

Observe that $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = \frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}}$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k+1,l}} = \frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{k+1,l}}$ when $z^{(\mathsf{tex})}_i = \overline{z}^{(\mathsf{tex})}_i$ $(\forall i \in \{1,\cdots,k\})$. Hence if we can prove $\frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = 0$ and $\frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{k+1,l}} = 0$ satisfied across the surrogate function family, $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = 0$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k+1,l}} = 0$ can be proven.

For a specific surrogate function $h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)$, we compare the generation process of the $(k+1)^{th}$ token $X^{(\mathsf{tex})}_{:,k+1} = \mathbf{g}_{k+1}\big(z_{\mathsf{inv}}, \{\overline{z}^{(\mathsf{tex})}_j, z^{(\mathsf{tex})}_{k+1}\}_{j=1}^{k}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big)$ with the text generation process in Assumption.11 . We rewrite $\mathbf{g}_{k+1}\big(z_{\mathsf{inv}}, \{\overline{z}^{(\mathsf{tex})}_j, z^{(\mathsf{tex})}_{k+1}\}_{j=1}^{k}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big)$ into $\mathbf{g}'_{k+1,\overline{Z}_k}\big(z_{\mathsf{inv}}, z^{(\mathsf{tex})}_{k+1}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big)$, which $\overline{Z}_k$ are underscored as a part of the nonlinear mixing function instead of variables. It holds a symbosis as follows

$$\begin{aligned}&\text{Generation of } x^{(\mathsf{tex})} \text{ in Assumption.11 :} \\ &z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}},\ z^{(\mathsf{tex})}_{\mathsf{dp}} \sim p_{\boldsymbol{z}^{(\mathsf{tex})}_{\mathsf{dp}}}(\cdot|z_{\mathsf{inv}}), \qquad z^{(\mathsf{tex})}_{\mathsf{pr}} \sim p_{\boldsymbol{z}^{(\mathsf{tex})}_{\mathsf{pr}}},\ x^{(\mathsf{tex})} = \mathbf{g}\big(z_{\mathsf{inv}}, z^{(\mathsf{tex})}_{\mathsf{dp}}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big); \\ &\text{Generation of } X^{(\mathsf{tex})}_{:,k+1} \text{ in Assumption.12 :} \\ &z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}},\ z^{(\mathsf{tex})}_{k+1} \sim p_{\boldsymbol{z}^{(\mathsf{tex})}_{k+1}}(\cdot|z_{\mathsf{inv}}, \overline{Z}_k), z^{(\mathsf{tex})}_{\mathsf{pr}} \sim p_{\boldsymbol{z}^{(\mathsf{tex})}_{\mathsf{pr}}},\ X^{(\mathsf{tex})}_{:,k+1} = \mathbf{g}'_{k+1,\overline{Z}_k}\big(z_{\mathsf{inv}}, z^{(\mathsf{tex})}_{k+1}, z^{(\mathsf{tex})}_{\mathsf{pr}}\big).\end{aligned} \tag{62}$$

Given this, if we reframe $p_{\boldsymbol{z}^{(\mathsf{tex})}_{k+1}}(\cdot, \overline{Z}_k)$ and $\mathbf{g}'_{k+1,\overline{Z}_k}(\cdot)$ as $p_{\boldsymbol{z}^{(\mathsf{tex})}_{\mathsf{dp}}}(\cdot)$ and $\mathbf{g}(\cdot)$, respectively, then the proof of $\frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = 0$ and $\frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{k+1,l}} = 0$ can be reduced to the proof of $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{dp},j}}=0$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},j}}=0$ in Theorem.9. It is satisfied and since the $\overline{Z}_k$ can be a arbitrary combination draw from the generative process in Assumption.12, $\frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = 0$ and $\frac{\partial h'_g(\cdot|z_{\mathsf{inv}};\overline{Z}_k)}{\partial z^{(\mathsf{tex})}_{k+1,l}} = 0$ are satisfied across the surrogate function family so that $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}} = 0$ and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k+1,l}} = 0$ have been proved.

To this, we have $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{k',l}}$=0 ($\forall l \in \{1,\cdots,n^{(\mathsf{tex})}_{k'}\}$, $\forall k' \in \{1,\cdots,k+1\}$) and $\frac{\partial h_g(\cdot|z_{\mathsf{inv}})}{\partial z^{(\mathsf{tex})}_{\mathsf{pr},\hat{j}}}$=0 ($\forall \hat{j} \in \{1,\cdots,n^{(\mathsf{tex})}_{\mathsf{pr}}\}$.

To this end, we have restricted h_f and h_g taking value in $\mathcal{C}_{\mathsf{inv}}$, thus, h_f=$f^*\circ\mathbf{f}_{1:n_{\mathsf{inv}}}$ and h_g=$g^*\circ\mathbf{g}^{(k)}_{1:n_{\mathsf{inv}}}$ ($\forall\, k \in \{1,\cdots \mathsf{k}_{\max}\}$), thus, h_g=$g^*\circ\mathbf{g}_{1:n_{\mathsf{inv}}}$.

Invertability of h_f, h_g. The procedure of proving the invertability of h_f is consistent with Theorem.9. As to the invertability of h_g, we consider its piecewise functions derived from $\{\mathbf{g}^{(k)}\}_{k=1}^{\mathsf{k}_{\max}}$ that generate sentence matrices with different sizes of their columns, then $\forall k \in \{1,\cdots,\mathsf{k}_{\max}\}$,

$$\mathbf{g}=\mathbf{g}^{(k)}:\ \mathcal{C}_{\mathsf{inv}}\times\Big(\mathcal{S}_1\times\cdots\mathcal{S}_k\Big)\times\mathcal{S}_{\mathsf{pr}}\to\mathcal{M}^{(k)}(\mathcal{X}^{(k)}_{\mathsf{tex}}),\ \forall k\in\{1,\cdots,\mathsf{k}_{\mathsf{max}}\},\tag{63}$$

and because of

$$g^*=\mathbf{d}\circ(\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}:\ \mathcal{M}^{(k)}(\mathcal{X}^{(k)}_{\mathsf{tex}})\to(0,1)^{n_{\mathsf{inv}}},\ \forall k\in\{1,\cdots,\mathsf{k}_{\mathsf{max}}\},\tag{64}$$

which is smooth on the generative process of $\mathcal{M}^{(k)}(\mathcal{X}^{(k)}_{\mathsf{tex}})$, we apply our result to Lemma.14 by setting $\mathcal{M}$=$\mathcal{C}_{\mathsf{inv}}$ and $\mathcal{N}$=$(0,1)^{n_{\mathsf{inv}}}$. In terms of the smoothness of h_g in each generative process via $\mathbf{g}^{(k)}$, they are differentiable maps so that all satisfy h in the lemma by mapping the random variable $\boldsymbol{z}_{\mathsf{inv}}\in\mathcal{C}_{\mathsf{inv}}$ into a uniform random variable in $(0,1)^{n_{\mathsf{inv}}}$. Notice that $p_{\boldsymbol{z}_{\mathsf{inv}}}$ (Assumption.12) and the uniform distribution (the pushforward of $p_{\boldsymbol{z}_{\mathsf{inv}}}$) are regular densities in the sense of Lemma.14, therefore h_g is a bijective map with respect to $\forall k\in\{1,\cdots,\mathsf{k}_{\max}\}$, *i.e.*, invertable.

B4. Proof of Corollary.2

The proof of Corollary.2 can be typically derived from the proof of Corollary.10.

B5. Proof of Theorem.3

To prove the result, we only need to construct g^{**} based on the optimal text encoder g^* defined by Theorem.1 and take it to define g^{**}, then prove $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*,g^{**})=0$. Afterwards, we prove its invariance to the permutation of sentence-matrix columns given $\pi(X^{(\mathsf{tex})})\in\Pi_k(\{1,\cdots,k\})$ that satisfies

$$g^{**}([X^{(\mathsf{tex})}_{:,1},X^{(\mathsf{tex})}_{:,2},\cdots,X^{(\mathsf{tex})}_{:,k}])=g^{**}([X^{(\mathsf{tex})}_{:,\pi(1)},X^{(\mathsf{tex})}_{:,\pi(2)},\cdots,X^{(\mathsf{tex})}_{:,\pi(k)}]),\tag{65}$$

it holds $\forall\hat{\pi}(X^{(\mathsf{tex})})\in\Pi_k(\{1,\cdots,k\})\cap\big\{\{X^{(\mathsf{tex})}_{:,1},X^{(\mathsf{tex})}_{:,\pi(1)}\}\times\cdots\times\{X^{(\mathsf{tex})}_{:,k},X^{(\mathsf{tex})}_{:,\pi(k)}\}\big\}$,

$$g^{**}([X^{(\mathsf{tex})}_{:,1},X^{(\mathsf{tex})}_{:,2},\cdots,X^{(\mathsf{tex})}_{:,k}])=g^{**}([X^{(\mathsf{tex})}_{:,\hat{\pi}(1)},X^{(\mathsf{tex})}_{:,\hat{\pi}(2)},\cdots,X^{(\mathsf{tex})}_{:,\hat{\pi}(k)}]).\tag{66}$$

Construction of g^{**}. From the first-phase proof of Theorem.1, we have g^* as a piecewise function on differnt-length text inputs, which satisfies

$$g^*=\mathbf{d}\circ(\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}:\quad\mathcal{M}^{(k)}(\mathcal{X}^{(k)}_{\mathsf{tex}})\to(0,1)^{n_{\mathsf{inv}}},\ \forall k\in\{1,\cdots,\mathsf{k}_{\mathsf{max}}\},\tag{67}$$

where $\mathbf{d}$ is defined on $\mathcal{C}_{\mathsf{inv}}$ and developed from Damois construction, and

$$\begin{aligned}&\forall\big(z_{\mathsf{inv}},z^{(\mathsf{tex})}_1,\cdots,z^{(\mathsf{tex})}_k,z^{(\mathsf{tex})}_{\mathsf{pr}}\big)\in\mathcal{C}_{\mathsf{inv}}\times\Big(\mathcal{S}_1\times\cdots\mathcal{S}_k\Big)\times\mathcal{S}^{\mathsf{tex}}_{\mathsf{pr}}\\&\quad\mathbf{g}^{(k)}\Big(z_{\mathsf{inv}},z^{(\mathsf{tex})}_1,\cdots,z^{(\mathsf{tex})}_k,z^{(\mathsf{tex})}_{\mathsf{pr}}\Big)=\Big[\mathbf{g}_1\big(z_{\mathsf{inv}},z^{(\mathsf{tex})}_1,z^{(\mathsf{tex})}_{\mathsf{pr}}\big),\cdots,\mathbf{g}_k\big(z_{\mathsf{inv}},\{z^{(\mathsf{tex})}_j\}_{j=1}^k,z^{(\mathsf{tex})}_{\mathsf{pr}}\big)\Big].\end{aligned}\tag{68}$$

In terms of the smoothness and invertibility of $\mathbf{d}(\cdot)$, we may construct a new function $\hat{\mathbf{g}}^{(k)}$ from $\mathbf{g}^{(k)}$, such that $\hat{\mathbf{g}}^{(k)}$: $\mathcal{M}^{(k)}(\mathcal{X}^{(k)}_{\mathsf{tex}})\to\mathcal{C}_{\mathsf{inv}}\times\Big(\mathcal{S}_1\times\cdots\mathcal{S}_k\Big)\times\mathcal{S}^{\mathsf{tex}}_{\mathsf{pr}}$ and $\hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}=(\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}$, leading to $g^{**}=\mathbf{d}\circ(\hat{\mathbf{g}}^{(k)})_{1:n_{\mathsf{inv}}}=\mathbf{d}\circ(\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}=g^*$. Specifically, $\hat{\mathbf{g}}^{(k)}$ can be constructed by

$$\hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})=\bigcap_{j=1}^{k}(\mathbf{g}^{(k)})^{-1}\big(\mathcal{T}^{(1)}_{\mathsf{tex}}\times\cdots\mathcal{T}^{(j-1)}_{\mathsf{tex}}\times\{X^{(\mathsf{tex})}_{:,j}\}\times\mathcal{T}^{(j+1)}_{\mathsf{tex}}\cdots\mathcal{T}^{(k)}_{\mathsf{tex}}\big)\tag{69}$$

where $\mathcal{T}_{\mathsf{tex}}^{(j)}$ indicates the jth token embedding space with respect to all k-length sentence matrices lying on $\mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)})$, then $\mathcal{T}_{\mathsf{tex}}^{(1)}\times\cdots\mathcal{T}_{\mathsf{tex}}^{(j-1)}\times\{X_{:,j}^{(\mathsf{tex})}\}\times\mathcal{T}_{\mathsf{tex}}^{(j+1)}\cdots\mathcal{T}_{\mathsf{tex}}^{(k)}$ denotes the matrix set including all k-length sentence matrices whose j^{th} token embedding are $X_{:,j}^{(\mathsf{tex})}$. It is noteworthy that we generalize the definition of $(\mathbf{g}^{(k)})^{-1}$, which receives a set of sentence matrices $\mathcal{X}'$ to infer the set of all possible values in their latent variables, *i.e.*, $(\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}) = \Big\{\hat{z} := \big(z_{\mathsf{inv}}, z_1^{(\mathsf{tex})}, \cdots, z_k^{(\mathsf{tex})}, z_{\mathsf{pr}}^{(\mathsf{tex})}\big) \in \mathcal{C}_{\mathsf{inv}} \times \Big(\mathcal{S}_1 \times \cdots \mathcal{S}_k\Big) \times \mathcal{S}_{\mathsf{pr}}^{\mathsf{tex}}, \mathsf{s.t.}\ \mathbf{g}^{(k)}(\hat{z}) \in \hat{\mathcal{X}}\Big\}$. For simplicity, we denote

$$\hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})}) = \mathcal{T}_{\mathsf{tex}}^{(1)}\times\cdots\mathcal{T}_{\mathsf{tex}}^{(j-1)}\times\{X_{:,j}^{(\mathsf{tex})}\}\times\mathcal{T}_{\mathsf{tex}}^{(j+1)}\cdots\mathcal{T}_{\mathsf{tex}}^{(k)}$$

therefore

$$\hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}) = \bigcap_{j=1}^{k}(\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})})). \tag{70}$$

To facilitate the ongoing proof, we need to prove the lemma below:

Lemma 16. *$\forall X^{(\mathsf{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)})$, then $\forall z' \in \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})$, $z'_{1:n_{\mathsf{inv}}} = z_{\mathsf{inv}} = (\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})})$.*

Proof. The proof is achieved by two steps.

In the first step, we prove that there exists $z' \in \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})$, $z'_{1:n_{\mathsf{inv}}} = z_{\mathsf{inv}} = (\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})})$. It is obvious since $\forall j \in \{1,\cdots,k\}$, $X^{(\mathsf{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)}) \subset \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})})$ thus $(\mathbf{g}^{(k)})^{-1}(X^{(\mathsf{tex})}) \in (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})}))$, so it leads to $(\mathbf{g}^{(k)})^{-1}(X^{(\mathsf{tex})}) \in \bigcap_{j=1}^{k}(\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})})) = \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})$. Given this, we set $z' = (\mathbf{g}^{(k)})^{-1}(X^{(\mathsf{tex})})$ and based on Theorem.1 , $z'_{1:n_{\mathsf{inv}}} = (\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})}) = z_{\mathsf{inv}}$ is obtained.

In the second step, we make a contradiction to verify arbitrary elements in $\hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})$ fulfill the equality. Suppose that $\exists z' \in \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})$ in the violation of $z'_{1:n_{\mathsf{inv}}} \neq (\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})})$. To this, we consider the image-text generative process based on Assumption.12 , where we define the sentence matrix $X^{(\mathsf{tex})'} = \mathbf{g}^{(k)}(z')$. Due to $\exists j \in \{1,\cdots,k\}$ with $X_{:,j}^{(\mathsf{tex})'} \neq X_{:,j}^{(\mathsf{tex})}$ otherwise $z'_{1:n_{\mathsf{inv}}} \neq (\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})})$ can not be met, we have $X^{(\mathsf{tex})'} \in \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})'})$. Besides, $z' \in \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}) = \bigcap_{j'=1}^{k}(\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j'}^{(\mathsf{tex})})) \subset (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})}))$, it results in $X^{(\mathsf{tex})'} = \mathbf{g}^{(k)}(z') \in \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})})$, *i.e.*, $X^{(\mathsf{tex})'} \in \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})'}) \cap \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})})$. However,

$$\begin{aligned}
&\hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})'}) \cap \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})}) \\
=&\Big(\mathcal{T}_{\mathsf{tex}}^{(1)'}\times\cdots\mathcal{T}_{\mathsf{tex}}^{(j-1)'}\times\{X_{:,j}^{(\mathsf{tex})'}\}\times\mathcal{T}_{\mathsf{tex}}^{(j+1)'}\cdots\mathcal{T}_{\mathsf{tex}}^{(k)'}\Big) \bigcap \Big(\mathcal{T}_{\mathsf{tex}}^{(1)}\times\cdots\mathcal{T}_{\mathsf{tex}}^{(j-1)}\times\{X_{:,j}^{(\mathsf{tex})}\}\times\mathcal{T}_{\mathsf{tex}}^{(j+1)}\cdots\mathcal{T}_{\mathsf{tex}}^{(k)}\Big) \\
=&(\mathcal{T}_{\mathsf{tex}}^{(1)'} \cap \mathcal{T}_{\mathsf{tex}}^{(1)}) \times\cdots(\mathcal{T}_{\mathsf{tex}}^{(j-1)'} \cap \mathcal{T}_{\mathsf{tex}}^{(j-1)})\times(\{X_{:,j}^{(\mathsf{tex})'}\} \cap \{X_{:,j}^{(\mathsf{tex})}\})\cdots(\mathcal{T}_{\mathsf{tex}}^{(k)'} \cap \mathcal{T}_{\mathsf{tex}}^{(k)'}),
\end{aligned} \tag{71}$$

where we observe $\{X_{:,j}^{(\mathsf{tex})'}\} \cap \{X_{:,j}^{(\mathsf{tex})}\} = \emptyset$ so that $\hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})'}) \cap \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})}) = \emptyset$. It is conflicted with $X^{(\mathsf{tex})'} \in \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})'}) \cap \hat{\mathcal{X}}(X_{:,j}^{(\mathsf{tex})})$.

Combine the two steps and the lemma has been proved. □

Based on Lemma.16 , $\forall X^{(\mathsf{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)})$, the set function $\hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})$ holds the output as a set composed of elements with their first n_{inv}-dim partition consistent with z_{inv}. Given this, we may define $\hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})})$ with the elements restricted on first n_{inv}-dim partition of the elements in $\hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})})$. Obviously $\hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})}) = \{z_{\mathsf{inv}}\}$ so that we can define $\hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})}) = z_{\mathsf{inv}}$ instead. To this end, g^{**} can be defined by $\hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}$:

$$g^{**} = \mathbf{d} \circ \hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}} : \quad \mathcal{M}^{(k)}(\mathcal{X}_{\mathsf{tex}}^{(k)}) \to (0,1)^{n_{\mathsf{inv}}},\ \forall k \in \{1,\cdots,\mathsf{k}_{\mathsf{max}}\}, \tag{72}$$

which replaces $(\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}$ by $\hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}$ in Eq.(67). Obviously, $(\mathbf{g}^{(k)})^{-1}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})}) = \hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}(X^{(\mathsf{tex})})$ for $\forall X^{(\mathsf{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}^{(k)}_{\mathsf{tex}})$ then $g^{**}(X^{(\mathsf{tex})}) = g^{*}(X^{(\mathsf{tex})})$, which results in $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*, g^{**}) = 0$ from $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*, g^*) = 0$.

Permutation-insensitive g^{**} in conditioned modal invariance. Given g^{**} that we constructed above, let consider the following conditioned modal-invariant alignment:

$$\begin{aligned} &g^{**}([X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,2}, \cdots, X^{(\mathsf{tex})}_{:,k}]) = g^{**}([X^{(\mathsf{tex})}_{:,\pi(1)}, X^{(\mathsf{tex})}_{:,\pi(2)}, \cdots, X^{(\mathsf{tex})}_{:,\pi(k)}]) \\ \Longleftrightarrow\ &\hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}([X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,2}, \cdots, X^{(\mathsf{tex})}_{:,k}]) = \hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}([X^{(\mathsf{tex})}_{:,\pi(1)}, X^{(\mathsf{tex})}_{:,\pi(2)}, \cdots, X^{(\mathsf{tex})}_{:,\pi(k)}]) = z^*_{\mathsf{inv}}. \end{aligned} \tag{73}$$

Then we return to

$$\hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}) = \bigcap_{j=1}^{k} (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j})); \ \ \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}_{\pi}) = \bigcap_{j=1}^{k} (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j)})), \tag{74}$$

where $X^{(\mathsf{tex})}_{\pi} = [X^{(\mathsf{tex})}_{:,\pi(1)}, X^{(\mathsf{tex})}_{:,\pi(2)}, \cdots, X^{(\mathsf{tex})}_{:,\pi(k)}]$, and consider their union

$$\begin{aligned} &\hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}) \cup \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}_{\pi}) \\ =&\Big(\bigcap_{j=1}^{k} (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j}))\Big) \bigcup \Big(\bigcap_{j=1}^{k} (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j)}))\Big) \\ =&\bigcap_{j=1}^{k} \Big((\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j})) \bigcup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j)}))\Big). \end{aligned} \tag{75}$$

From the definition of $(\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j})$, it holds $\forall \hat{z} \in (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j})$ that satisfies

$$\mathbf{g}^{(k)}(\hat{z}) \in \hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j}) = \mathcal{T}^{(1)}_{\mathsf{tex}} \times \cdots \mathcal{T}^{(j-1)}_{\mathsf{tex}} \times \{X^{(\mathsf{tex})}_{:,j}\} \times \mathcal{T}^{(j+1)}_{\mathsf{tex}} \cdots \mathcal{T}^{(k)}_{\mathsf{tex}}, \tag{76}$$

similarly, we also have $\hat{z} \in (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j)})$ that satisfies

$$\mathbf{g}^{(k)}(\hat{z}) \in \hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j)}) = \mathcal{T}^{(1,\pi)}_{\mathsf{tex}} \times \cdots \mathcal{T}^{(j-1,\pi)}_{\mathsf{tex}} \times \{X^{(\mathsf{tex})}_{:,\pi(j)}\} \times \mathcal{T}^{(j+1,\pi)}_{\mathsf{tex}} \cdots \mathcal{T}^{(k,\pi)}_{\mathsf{tex}}. \tag{77}$$

It results in $\forall \hat{z} \in (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j})) \bigcup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j)}))$,

$$\mathbf{g}^{(k)}(\hat{z}) \in \big(\mathcal{T}^{(1)}_{\mathsf{tex}} \cup \mathcal{T}^{(1,\pi)}_{\mathsf{tex}}\big) \times \cdots \big(\mathcal{T}^{(j-1)}_{\mathsf{tex}} \cup \mathcal{T}^{(j-1,\pi)}_{\mathsf{tex}}\big) \times \{X^{(\mathsf{tex})}_{:,j}, X^{(\mathsf{tex})}_{:,\pi(j)}\} \times \cdots \big(\mathcal{T}^{(k)}_{\mathsf{tex}} \cup \mathcal{T}^{(k,\pi)}_{\mathsf{tex}}\big). \tag{78}$$

Hence $\forall \hat{z} \in \cap_{j=1}^{k} \big((\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j})) \cup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j)}))\big) = \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}) \cup \hat{\mathbf{g}}^{(k)}(X^{(\mathsf{tex})}_{\pi})$,

$$\hat{X}^{(\mathsf{tex})} = \mathbf{g}^{(k)}(\hat{z}) \in \{X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,\pi(1)}\} \times \cdots \times \{X^{(\mathsf{tex})}_{:,j}, X^{(\mathsf{tex})}_{:,\pi(j)}\} \times \cdots \{X^{(\mathsf{tex})}_{:,k}, X^{(\mathsf{tex})}_{:,\pi(k)}\} \tag{79}$$

with $g^{**}(\hat{X}^{(\mathsf{tex})}) = \mathbf{d} \circ \hat{\mathbf{g}}^{(k)}_{1:n_{\mathsf{inv}}}(\hat{X}^{(\mathsf{tex})}) = \mathbf{d}(z_{\mathsf{inv}})$. Thus, $\forall \hat{\pi}(X^{(\mathsf{tex})}) \in \Pi_k(\{1, \cdots, k\}) \cap \big\{\{X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,\pi(1)}\} \times \cdots \times \{X^{(\mathsf{tex})}_{:,k}, X^{(\mathsf{tex})}_{:,\pi(k)}\}\big\}$, it holds

$$g^{**}([X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,2}, \cdots, X^{(\mathsf{tex})}_{:,k}]) = g^{**}([X^{(\mathsf{tex})}_{:,\hat{\pi}(1)}, X^{(\mathsf{tex})}_{:,\hat{\pi}(2)}, \cdots, X^{(\mathsf{tex})}_{:,\hat{\pi}(k)}]). \tag{80}$$

B6. Proof of Theorem.4

Our proof starts with g^{**} constructed in Eq.72 . Given this, we consider the condition provided in Theorem.4 :

$$g^{**}([X^{(\mathsf{tex})}_{:,1}, \cdots, X^{(\mathsf{tex})}_{:,j}, \cdots, X^{(\mathsf{tex})}_{:,k}]) = g^{**}([X^{(\mathsf{tex})}_{:,\pi(1)}, \cdots, \mathsf{RP}(X^{(\mathsf{tex})}_{:,j}), \cdots, X^{(\mathsf{tex})}_{:,\pi(k)}]), \tag{81}$$

in which $X^{(\mathsf{tex})}_{:,j}$ and $\mathsf{RP}(X^{(\mathsf{tex})}_{:,j})$ denote the pairwise embeddings composed of a word-or-phrase token and its rephrased token that satisfies the aforementioned conditioned modal-invariant alignment, and $\pi(X^{(\mathsf{tex})}) \in \Pi_{k-1}(\{1, \cdots, j-1, j+1, \cdots, k\})(j)$ refers to the permutation of

$\{1, \cdots, k\}$ where j fixed in the position. Follow the similar induction in Theorem.3 and we have $\forall \hat{z} \in \cap_{j'=1, j' \neq j}^{k} \big((\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j'})) \cup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,\pi(j')}))\big) \cap \big((\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X^{(\mathsf{tex})}_{:,j})) \cup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(\mathsf{RP}(X^{(\mathsf{tex})}_{:,j})))\big)$,

$$\hat{X}^{(\mathsf{tex})} = \mathbf{g}^{(k)}(\hat{z}) \in \{X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,\pi(1)}\} \times \cdots \times \{X^{(\mathsf{tex})}_{:,j}, \mathsf{RP}(X^{(\mathsf{tex})}_{:,(j)})\} \times \cdots \{X^{(\mathsf{tex})}_{:,k}, X^{(\mathsf{tex})}_{:,\pi(k)}\}. \quad (82)$$

Obviously, it holds $\forall \hat{\pi}(X^{(\mathsf{tex})}_{:,-j}) \in \Pi_{k-1}(\{1, \cdots, j-1, j+1, \cdots, k\}) \bigcap \big\{\{X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,\pi(1)}\} \times \cdots \{X^{(\mathsf{tex})}_{:,j-1}, X^{(\mathsf{tex})}_{:,\pi(j-1)}\} \times \{X^{(\mathsf{tex})}_{:,j+1}, X^{(\mathsf{tex})}_{:,\pi(j+1)}\} \cdots \times \{X^{(\mathsf{tex})}_{:,k}, X^{(\mathsf{tex})}_{:,\pi(k)}\}\big\}$ and $\forall \hat{X}^{(1)}_j, \hat{X}^{(2)}_j \in \{X^{(\mathsf{tex})}_{:,j}, \mathsf{RP}(X^{(\mathsf{tex})}_{:,j})\}$,

$$g^{**}([X^{(\mathsf{tex})}_{:,1}, \cdots, \hat{X}^{(1)}_j, \cdots, X^{(\mathsf{tex})}_{:,k}]) = g^{**}([X^{(\mathsf{tex})}_{:,\hat{\pi}(1)}, \cdots, \hat{X}^{(2)}_j, \cdots, X^{(\mathsf{tex})}_{:,\hat{\pi}(k)}]). \quad (83)$$

B7. Proof of Theorem 5

For each token length $\ell \in \mathbb{N}$, let

$$g^{(\ell)} : \mathcal{C}_{\mathsf{inv}} \times \Big(\prod_{i=1}^{\ell} \mathcal{S}_i\Big) \times \mathcal{S}^{\mathsf{tex}}_{\mathsf{pr}} \longrightarrow \mathcal{M}^{(\ell)}(X^{(\ell)}_{\mathsf{tex}})$$

be the diffeomorphism associated with ℓ-token texts, with inverse $(g^{(\ell)})^{-1}$ and projection $(g^{(\ell)})^{-1}_{1:n_{\mathsf{inv}}} : \mathcal{M}^{(\ell)}(X^{(\ell)}_{\mathsf{tex}}) \to \mathcal{C}_{\mathsf{inv}}$ to the first n_{inv} coordinates. By Theorem 1, there exist invertible heads h_{f^*}, h_{g^*} with $f^* = h_{f^*} \circ \mathbf{f}^{-1}_{1:n_{\mathsf{inv}}}$ and $g^* = h_{g^*} \circ \mathbf{g}^{-1}_{1:n_{\mathsf{inv}}}$, such that $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*, g^*) \to 0$.

Let $\mathcal{X}_{\mathsf{base}} \subseteq \mathcal{M}^{(j)}(X^{(j)}_{\mathsf{tex}})$ be a family of base j-token sentences containing $X^{(\mathsf{tex})}$. Define the ADD family (length $j+1$) by

$$\mathcal{X}_{\mathsf{ADD}} := \Big\{\hat{\pi}(X^{(\mathsf{tex})}) = [X^{(\mathsf{tex})}_{:,1}, \ldots, X^{(\mathsf{tex})}_{:,j}, \mathsf{ADD}(X^{(\mathsf{tex})}_{:,j}), X^{(\mathsf{tex})}_{:,j+1}, \ldots, X^{(\mathsf{tex})}_{:,k}] \ : \ X^{(\mathsf{tex})} \in \mathcal{X}_{\mathsf{base}}\Big\}.$$

The statement posits an intersection condition on the invariant component:

$$\exists z^*_{\mathsf{inv}} \in \mathcal{C}_{\mathsf{inv}} \quad \text{s.t.} \quad z^*_{\mathsf{inv}} \in ((g^*)^{(j)})^{-1}_{1:n_{\mathsf{inv}}}(\mathcal{X}_{\mathsf{base}}) \cap ((g^*)^{(j+1)})^{-1}_{1:n_{\mathsf{inv}}}(\mathcal{X}_{\mathsf{ADD}}). \quad (84)$$

Set-valued inverse constructions

.We follow the Theorem 7/8 pattern (set-valued inverse, column-fixing intersections, and constantization on z_{inv}).

For $X \in \mathcal{X}_{\mathsf{base}} \subset \mathcal{M}^{(j)}(X^{(j)}_{\mathsf{tex}})$, define

$$\widehat{g}^{(j)}(X) := \big((g^{(j)})^{-1}(\mathcal{X}_{\mathsf{base}})\big) \cap \Pi^{(j)}_{\mathrm{fix}}(X), \quad (85)$$

where $\Pi^{(j)}_{\mathrm{fix}}(X)$ denotes the intersection of $(g^{(j)})^{-1}$ over sets that fix the columns of X we choose to keep identical within $\mathcal{X}_{\mathsf{base}}$ (as in the permutation-style constructions). Let $\widehat{g}^{(j)}_{1:n_{\mathsf{inv}}}(X)$ be its projection to the first n_{inv} coordinates.

For $Y \in \mathcal{X}_{\mathsf{ADD}} \subset \mathcal{M}^{(j+1)}(X^{(j+1)}_{\mathsf{tex}})$ with $Y = \hat{\pi}(X)$ for some $X \in \mathcal{X}_{\mathsf{base}}$, define

$$\widehat{g}^{(j+1)}(Y) := \big((g^{(j+1)})^{-1}(\mathcal{X}_{\mathsf{ADD}})\big) \cap \Pi^{(j+1)}_{\mathrm{fix}}(X, Y), \quad (86)$$

where $\Pi^{(j+1)}_{\mathrm{fix}}(X, Y)$ fixes all columns of Y that correspond to columns of X after inserting $\mathsf{ADD}(X^{(\mathsf{tex})}_{:,j})$ at position j (i.e., all shared columns except the newly inserted one). Let $\widehat{g}^{(j+1)}_{1:n_{\mathsf{inv}}}(Y)$ be its projection.

Lemma 17 (Constancy of z_{inv} on base and ADD families)**.** *Under Eq.84, we have*

$$\widehat{g}^{(j)}_{1:n_{\mathsf{inv}}}(X) = \{z^*_{\mathsf{inv}}\} \quad \forall X \in \mathcal{X}_{\mathsf{base}}, \qquad \widehat{g}^{(j+1)}_{1:n_{\mathsf{inv}}}(Y) = \{z^*_{\mathsf{inv}}\} \quad \forall Y \in \mathcal{X}_{\mathsf{ADD}}.$$

Proof. By assumption Eq.84, the first n_{inv} projections of the inverse preimages of $\mathcal{X}_{\mathsf{base}}$ (length j) and $\mathcal{X}_{\mathsf{ADD}}$ (length $j+1$) both contain z^*_{inv}. Intersecting with $\Pi^{(j)}_{\mathrm{fix}}(X)$ and $\Pi^{(j+1)}_{\mathrm{fix}}(X, Y)$ only constrains token-specific coordinates and the alignment of shared columns; it does not alter the $1{:}n_{\mathsf{inv}}$ coordinates. By block identifiability in Theorem 1, the $1{:}n_{\mathsf{inv}}$ projection is unique, hence each projection collapses to the singleton $\{z^*_{\mathsf{inv}}\}$. □

DEFINITION OF g^{**} AND ITS PROPERTIES

.Define a pseudo-optimal text encoder g^{**} by reusing the optimal head h_{g^*} on the constantized invariant coordinates:

$$g^{**}(Z) := \begin{cases} h_{g^*}\big(\widehat{g}^{(j)}_{1:n_{\mathsf{inv}}}(Z)\big), & Z \in \mathcal{X}_{\mathsf{base}}, \\ h_{g^*}\big(\widehat{g}^{(j+1)}_{1:n_{\mathsf{inv}}}(Z)\big), & Z \in \mathcal{X}_{\mathsf{ADD}}, \\ g^*(Z), & \text{otherwise.} \end{cases} \tag{87}$$

ADD invariance. Let $X \in \mathcal{X}_{\mathsf{base}}$ and $Y = \hat{\pi}(X) \in \mathcal{X}_{\mathsf{ADD}}$ be the ADD-form hard negative described in the theorem. By Lemma 17, $\widehat{g}^{(j)}_{1:n_{\mathsf{inv}}}(X) = \widehat{g}^{(j+1)}_{1:n_{\mathsf{inv}}}(Y) = \{z^*_{\mathsf{inv}}\}$. Therefore,

$$g^{**}(X) = h_{g^*}(z^*_{\mathsf{inv}}) = g^{**}(Y),$$

which proves equation 8:

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

Optimal alignment preserved. By construction equation 87, $g^{**} = g^*$ outside $\mathcal{X}_{\mathsf{base}} \cup \mathcal{X}_{\mathsf{ADD}}$. On $\mathcal{X}_{\mathsf{base}}$ and $\mathcal{X}_{\mathsf{ADD}}$, Lemma 17 ensures that g^{**} applies the same invertible head h_{g^*} to the same invariant z^*_{inv} as g^* would use when evaluated on corresponding latents. Hence g^{**} coincides with g^* on the support up to the invariant coordinates preserved by h_{g^*}, and achieves the same global optimum:

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

Under the intersection condition in Eq.84, we have constructed a pseudo-optimal text encoder g^{**} derived from g^* that: (i) preserves the optimal MMAlign value with f^*, and (ii) is invariant to the ADD-form permutation $\hat{\pi}$ that inserts $\mathsf{ADD}(X^{(\mathsf{tex})}_{:,j})$ at position j, establishing ADD-form composition nonidentifiability.

B7. PROOF OF PROPOSITION 6

Since (f^*, g^{**}) achieves $\mathcal{L}^{(\mathrm{img},\mathrm{tex})}_{\mathrm{MMAlign}}(f^*, g^{**}) \to 0$, the alignment term satisfies

$$\mathbb{E}_{\langle \boldsymbol{x}^{(\mathrm{img})}, \boldsymbol{X}^{(\mathrm{tex})} \rangle \sim p_{\mathrm{mm}}} \left[\| f^*(\boldsymbol{x}^{(\mathrm{img})}) - g^{**}(\boldsymbol{X}^{(\mathrm{tex})}) \|^2 \right] \to 0. \tag{88}$$

By Assumption 13, the modality gap $\boldsymbol{\delta}$ is orthogonal to the span of text embeddings. Following the cross-modal transferability result of Zhang et al. (Proposition 2.1 therein), any function achieving optimal alignment on image embeddings preserves its decision geometry when evaluated via $g^{**}(\boldsymbol{X}^{(\mathrm{tex})}) + \boldsymbol{\delta}$. Thus, for matched pairs:

$$f^*(\boldsymbol{x}^{(\mathrm{img})}) = g^{**}(\boldsymbol{X}^{(\mathrm{tex})}) + \boldsymbol{\delta}, \tag{89}$$

$$f^*(\boldsymbol{x}^{(\mathrm{img})}_{\mathrm{op}}) = g^{**}(\mathrm{OP}(\boldsymbol{X}^{(\mathrm{tex})})) + \boldsymbol{\delta}. \tag{90}$$

Substituting the text-side nonidentifiability condition $g^{**}(\boldsymbol{X}^{(\mathrm{tex})}) = g^{**}(\mathrm{OP}(\boldsymbol{X}^{(\mathrm{tex})}))$:

$$f^*(\boldsymbol{x}^{(\mathrm{img})}) - f^*(\boldsymbol{x}^{(\mathrm{img})}_{\mathrm{op}}) = \Big(g^{**}(\boldsymbol{X}^{(\mathrm{tex})}) + \boldsymbol{\delta} \Big) - \Big(g^{**}(\mathrm{OP}(\boldsymbol{X}^{(\mathrm{tex})})) + \boldsymbol{\delta} \Big) = \mathbf{0}. \tag{91}$$

B8. PROOF OF COROLLARY 7

We proceed by iterative application of Proposition 6.

Step 1 (First-level nonidentifiability). By Theorems 3–5, applying $F_1(\cdot)$ to $\boldsymbol{X}^{(\mathrm{tex})}$ on a single atomic concept yields a pseudo-optimal encoder pair $\langle f^*, g^{**} \rangle$ satisfying $g^{**}(F_1(\boldsymbol{X}^{(\mathrm{tex})})) = g^*(\boldsymbol{X}^{(\mathrm{tex})})$. Let $x^{(\mathrm{img})}_{F_1}$ denote the image semantically corresponding to $F_1(\boldsymbol{X}^{(\mathrm{tex})})$. Since $\langle f^*, g^{**} \rangle$ is pseudo-optimal, the modality gap condition gives

$$f^*(x^{(\mathrm{img})}_{F_1}) = g^{**}(F_1(\boldsymbol{X}^{(\mathrm{tex})})) + \boldsymbol{\delta} = g^*(\boldsymbol{X}^{(\mathrm{tex})}) + \boldsymbol{\delta} = f^*(x^{(\mathrm{img})}),$$

where the last equality follows from Proposition 6 applied to the pair $\langle f^*, g^{**}\rangle$ relative to $\langle f^*, g^*\rangle$.

Step 2 (Second-level nonidentifiability). We now treat $\langle f^*, g^{**}\rangle$ as the reference encoder pair. Applying $F_2(\cdot)$ to $F_1(\boldsymbol{X}^{(\mathrm{tex})})$ on a different atomic concept, Theorems 3–5 guarantee the existence of a pseudo-optimal encoder g^{***} with respect to g^{**} such that $g^{***}(F_2(F_1(\boldsymbol{X}^{(\mathrm{tex})}))) = g^{**}(F_1(\boldsymbol{X}^{(\mathrm{tex})}))$. Let $x^{(\mathrm{img})}_{F_2\circ F_1}$ denote the image semantically corresponding to $F_2(F_1(\boldsymbol{X}^{(\mathrm{tex})}))$. Applying Proposition 6 again with $\langle f^*, g^{***}\rangle$ relative to $\langle f^*, g^{**}\rangle$, we obtain

$$f^*(x^{(\mathrm{img})}_{F_2\circ F_1}) = g^{***}(F_2(F_1(\boldsymbol{X}^{(\mathrm{tex})}))) + \boldsymbol{\delta} = g^{**}(F_1(\boldsymbol{X}^{(\mathrm{tex})})) + \boldsymbol{\delta} = f^*(x^{(\mathrm{img})}_{F_1}).$$

Step 3 (Transitivity). Combining Steps 1 and 2:

$$f^*(x^{(\mathrm{img})}) = f^*(x^{(\mathrm{img})}_{F_1}) = f^*(x^{(\mathrm{img})}_{F_2\circ F_1}),$$

which completes the proof. The argument extends by induction to any finite sequence of atomic-concept modifications $F_k \circ \cdots \circ F_1$.

B9. Downstream Impossibility for MLLMs: Formalization, Proofs, and Discussion

This subsection provides the formal content behind Theorem 8, together with the discussions that could not be accommodated in the main text: the average-case (data-processing) counterpart that our theorem subsumes, a weakened form of the modality-gap premise, and the weakening of the pseudo-optimality hypothesis.

Formal setup

Let $\mathcal{M}_0 = (\mathcal{Z}, P_Z, g_v, g_l, \varphi)$ denote the underlying SCM (Assumptions.12) : a latent space $\mathcal{Z}$ with prior P_Z, diffeomorphic generative maps $g_v : \mathcal{Z} \to \mathcal{X}$ (vision) and $g_l : \mathcal{Z} \to \mathcal{T}$ (text), and a semantic predicate $\varphi : \mathcal{Z} \to \{0, 1\}$ assigning the ground-truth label of op-queries. Writing $Z \sim P_Z$, the observational distribution over $\mathcal{X} \times \mathcal{T}$ is

$$P_0 = \mathrm{Law}\big(g_v(Z),\, g_l(Z)\big). \tag{92}$$

Let Π^* be the operator-induced permutation class of Theorems 3-5, and fix $\pi \in \Pi^*$ that is non-trivial on the concept slots bound by the operator; such a π exists whenever the op-query involves at least two exchangeable concept slots, which is exactly the construction in the proof of Theorem 4, while Theorem 3 supplies its regularity (π is a diffeomorphism of $\mathcal{Z}$). The alternative world replaces the prior by $\mathrm{Law}(\pi(Z))$ and the generative maps by $g_v \circ \pi^{-1}$, $g_l \circ \pi^{-1}$, keeping $\mathcal{Z}$ and φ unchanged:

$$\mathcal{M}_1 = \big(\mathcal{Z},\, \mathrm{Law}(\pi(Z)),\, g_v \circ \pi^{-1},\, g_l \circ \pi^{-1},\, \varphi\big). \tag{93}$$

Let Q be the distribution of op-query pairs over *observables*: following the MMVP protocol, a query is a pair of observed inputs, so Q is world-independent; only its *label*, being a function of the latent explanation through φ, is world-dependent. We write $y_b(q) \in \{0, 1\}$ for the ground-truth labeling of op-queries under $\mathcal{M}_b$.

Pipelines. A learning pipeline is a measurable map

$$\mathcal{A} : (\mathcal{X} \times \mathcal{T})^n \times \mathcal{Q} \times [0, 1] \longrightarrow \{0, 1\}, \qquad \hat{y} = \mathcal{A}(D, q, R), \tag{94}$$

whose three arguments are the training data D, the query q, and an internal randomness seed $R \sim \mathrm{Unif}[0, 1]$ drawn independently of everything else. This formalization covers arbitrary architectures, pre-training objectives, projectors, LLM decoding schemes and fine-tuning procedures, since all of these are functions of D, q and R only. In particular, the concrete MLLM pipeline studied in practice,

$$x^{(\mathrm{img})} \longrightarrow f^{**}(x^{(\mathrm{img})}) \longrightarrow \phi\big(f^{**}(x^{(\mathrm{img})})\big) \longrightarrow \mathrm{MLLM}\big(\phi(f^{**}(x^{(\mathrm{img})})), q\big), \tag{95}$$

with a frozen or lightly fine-tuned CLIP image encoder $f^{**} : \mathcal{X} \to \mathbb{R}^d$ and an adapter $\phi : \mathbb{R}^d \to \mathbb{R}^{d_{\mathrm{LLM}}}$, is one instance of $\mathcal{A}$.

Two pillars: observational equivalence and label flip

Lemma 18 (Observational equivalence). *$\mathcal{M}_0$ and $\mathcal{M}_1$ induce the identical joint distribution over image–text observations: $P_1 = P_0$.*

Proof. Let $Z \sim P_Z$ and set $Z' = \pi(Z)$, so that Z' has exactly the prior of $\mathcal{M}_1$. The observation generated by $\mathcal{M}_1$ from Z' is the pair

$$\big(g_v \circ \pi^{-1}(Z'),\ g_l \circ \pi^{-1}(Z')\big) = \big(g_v(Z),\ g_l(Z)\big), \tag{96}$$

because $\pi^{-1}(Z') = \pi^{-1}(\pi(Z)) = Z$ almost surely. Taking laws on both sides gives $P_1 = \mathrm{Law}(g_v(Z), g_l(Z)) = P_0$. Since π is a diffeomorphism of $\mathcal{Z}$ (Theorem 3), all maps above are measurable and every composition is well defined. □

Lemma 19 (Label flip). *$y_1 = 1 - y_0$ pointwise on the support of Q.*

Proof. Fix an observed query with visual component $x = g_v(z)$, where z lies in the support of P_Z. Under $\mathcal{M}_0$ the latent explanation of x is z, so the label is $y_0 = \varphi(z)$. Under $\mathcal{M}_1$ the generative map is $g_v \circ \pi^{-1}$, so the latent explanation of the same observation x is $\pi(z)$—unique, since $g_v \circ \pi^{-1}$ is injective—and the label is $y_1 = \varphi(\pi(z))$. By the choice of $\pi \in \Pi^*$ exchanging the concept slots bound by the operator, which is exactly the non-trivial element constructed in the proof of Theorem 4, we have $\varphi(\pi(z)) = 1 - \varphi(z)$ for all z in the latent support of op-queries. □

Lemmas 18 and 19 are the world-level restatement of the encoder-level nonidentifiability in text (i.e., Theorem.3-5): the two worlds emit byte-identical image–text data yet disagree on every compositional answer. Because the disagreement is invisible in P_0, it cannot be removed by optimizing any objective defined on P_0.

B10. Proof of Theorem 8

Proof. Fix any pipeline $\mathcal{A}$ and sample size n.

Step 1 (equality of transcript laws). Under world b, the transcript is the triple (D, q, R) with $D \sim P_b^{\otimes n}$, $q \sim Q$ and $R \sim \mathrm{Unif}[0,1]$, mutually independent. By Lemma 18 we have $P_1^{\otimes n} = P_0^{\otimes n}$, and the query distribution Q is defined over observables, hence is world-independent as well. Therefore the joint law of the transcript, and consequently the joint law of the pair $(q, \hat{y})$ with $\hat{y} = \mathcal{A}(D, q, R)$, is identical under $\mathcal{M}_0$ and $\mathcal{M}_1$. Denote this common law by μ.

Step 2 (error sum identity). Let $e_b = \Pr_{\mathcal{M}_b}[\hat{y} \neq y_b(q)]$. Using Step 1 to write both errors as expectations under the single measure μ, and Lemma 19 to substitute $y_1 = 1 - y_0$,

$$\begin{aligned} e_0 + e_1 &= \mathbb{E}_\mu\big[\mathbf{1}\{\hat{y} \neq y_0(q)\}\big] + \mathbb{E}_\mu\big[\mathbf{1}\{\hat{y} \neq 1 - y_0(q)\}\big] \\ &= \mathbb{E}_\mu\big[\mathbf{1}\{\hat{y} \neq y_0(q)\} + \mathbf{1}\{\hat{y} = y_0(q)\}\big] = 1, \end{aligned} \tag{97}$$

where the middle equality holds pointwise because $\hat{y}$ takes values in $\{0, 1\}$.

Step 3 (conclusion). From $e_0 + e_1 = 1$ we get $\max_b e_b \geq \frac{1}{2}$. Since $\mathcal{A}$ and n were arbitrary, taking the infimum over all pipelines and all sample sizes yields the claim. □

B11. Remarks and extensions

(a) Tightness and sample-size independence. The bound is exactly chance level for a binary query and is therefore tight: a pipeline answering uniformly at random attains $e_0 = e_1 = 1/2$. It also holds for *every* n, so no amount of observational pre-training or fine-tuning data escapes it; this sharpens the informal "irreversible bottleneck" claim into a statement in which neither model capacity nor data scale appears.

(b) k-way op-queries. If Π^* contains $k-1$ further permutations whose induced labelings are pairwise distinct on the support of Q, the same argument over the resulting k observationally equivalent worlds yields $\inf_{\mathcal{A}} \max_{b \in [k]} \Pr_{\mathcal{M}_b}[\mathcal{A} \text{ errs}] \geq 1 - 1/k$, i.e. chance level for k-way multiple choice.

(c) Average-case counterpart: the data-processing view. Theorem 8 quantifies over algorithms; for a *given* trained encoder one obtains the following instance-level statement, which is the version most directly comparable to empirical diagnostics. Let $z_{\mathrm{comp}} \in \mathcal{Z}_{\mathrm{comp}}$ denote the compositional latent variables required to answer op-queries, and let f^{**} be an image encoder exhibiting the visual-side nonidentifiability (Proposition.6 and Corollary.7), i.e. $\|f^{**}(x^{(\mathrm{img})}) - f^{**}(x_{\mathrm{op}}^{(\mathrm{img})})\| \le \eta$ for compositionally distinct $x^{(\mathrm{img})}, x_{\mathrm{op}}^{(\mathrm{img})}$. Then for any adapter ϕ and any language model of arbitrary capacity, the Markov chain $z_{\mathrm{comp}} \to x^{(\mathrm{img})} \to f^{**}(x^{(\mathrm{img})}) \to \mathrm{MLLM}(\cdot)$ of Eq. equation 95 and the data processing inequality give

$$I\big(z_{\mathrm{comp}}; \mathrm{MLLM}(\phi(f^{**}(x^{(\mathrm{img})})), q)\big) \;\le\; I\big(z_{\mathrm{comp}}; f^{**}(x^{(\mathrm{img})})\big) \;\xrightarrow[\eta\to 0]{}\; 0. \tag{98}$$

Thus the compositional information discarded by the visual encoder cannot be restored by any downstream module. This inequality is conditional on a particular encoder and on the premise used to establish visual-side collapse, whereas Theorem 8 dispenses with both; the two should be read as the average-case projection and the worst-case limit of the same phenomenon, the measured residuals (ϵ, η) locating a real system between chance and full recovery.

(d) Weakening the modality-gap premise. The uniform per-pair form of Assumption **??** can be replaced by its second-moment version,

$$\mathbb{E}\big\|f(x^{(\mathrm{img})}) - g(X^{(\mathrm{tex})}) - \delta\big\|^2 \le \epsilon^2, \tag{99}$$

which is precisely the residual statistic that can be measured on a trained checkpoint and which permits long-tailed residuals. Writing $r_1 = f(x) - g(X) - \delta$ and $r_2 = f(x_{\mathrm{op}}) - g(\mathrm{OP}(X)) - \delta$ and $S = \|r_1\| + \|r_2\|$, we have $\mathbb{E}[S^2] \le 2\mathbb{E}[\|r_1\|^2 + \|r_2\|^2] \le 4\epsilon^2$, so Chebyshev's inequality gives $\Pr[S > 2\epsilon/\sqrt{\beta}] \le \beta$. Combining with $\|g(X) - g(\mathrm{OP}(X))\| \le \eta$ and the triangle inequality, the conclusion of Proposition 6 holds in the quantitative form

$$\|f(x^{(\mathrm{img})}) - f(x_{\mathrm{op}}^{(\mathrm{img})})\| \;\le\; 2\epsilon/\sqrt{\beta} + \eta \qquad \text{with probability at least } 1-\beta \text{ over pairs.} \tag{100}$$

(e) Weakening pseudo-optimality. Positing an *attained* optimizer g^{**} is stronger than necessary. First, g^{**} is absent from Theorem 8, which quantifies over all pipelines, so attainment is vacuous there. Second, wherever a quantitative statement is needed we use ε-approximate pseudo-optimality: for every $\varepsilon > 0$ the set $\{g : \mathcal{L}(g) \le \inf_{g'} \mathcal{L}(g') + \varepsilon\}$ is nonempty by definition of the infimum, invoking neither attainment nor a measurable-selection argument. Since the bound in (d) is stated through the measured residual of the encoder at hand, it applies verbatim to any ε-approximate optimizer or trained checkpoint: the idealized optimum enters no hypothesis, only the interpretation.

(f) The only escape route. Every step above uses exclusively the observational law P_0. Consequently, a pipeline can exceed chance on op-queries only by accessing samples outside P_0, i.e. by breaking the observational equivalence of Lemma 18. A single-edit hard negative removes one non-trivial element of Π^*, whereas the compound (multi-calling) negatives of Section 4.4 target the stacked equivalence classes of Corollary 7; the intervention adopted in our method is therefore the one prescribed by Theorem 8 rather than a heuristic.

APPENDIX.C: IMPLEMENTATION

Hard-Negative Paris Augmentation. Theorem.3-5 refer to the corresponding data augmentation strategy derived from the prompts for hard negative data generation and the experimental evaluation as below:

> **Sub-prompt (SWAP instruction):** Read the text ⟨Text⟩, then permute its token order to generate a text that holds the same or most similar semantic (belongs to the same atomic concepts in (belongs to the same atomic concepts in OBJ,ATT,REL,QUA)) with ⟨Text⟩;

Sub-prompt (REPLACE instruction): Read the text ⟨Text⟩, then replace one of its language tokens by an arbitrary word or phrase (belonging to the same atomic concepts in (belongs to the same atomic concepts in OBJ,ATT,REL,QUA)) drawn from all its possible token permutations obtained by the following instruction:

⟨Prompt (SWAP instruction)⟩

such that the generated text holds the same or most similar semantic (belongs to the same atomic concepts in (belongs to the same atomic concepts in OBJ,ATT,REL,QUA)) with ⟨Text⟩;

Sub-prompt (ADD instruction): Read the text ⟨Text⟩, then generate the **ADD** output token type drawn from the atomic concept types (belongs to the same atomic concepts in OBJ,ATT,REL,QUA). Based on the **ADD** output token type, locate a token $X_{:,i}^{(\text{tex})}$ in the text ⟨Text⟩, then generate an **ADD** output $\text{ADD}(X_{:,i}^{(\text{tex})})$, and insert it ahead of or after $X_{:,i}^{(\text{tex})}$ in the text ⟨Text⟩;

Prompt (evaluation): Given a text: ⟨a text drawn from ARO⟩

Identify whether the prompt can be used to generate the text:

- ⟨Prompt (SWAP instruction)⟩, choose a combination of $\times_i \{\langle$Sub-prompt (SWAP instruction)$\rangle_i$, ⟨a text drawn from ARO⟩$_i\}$ ($\times_i$ indicates the Cartesian product for the i-th token) that holds the identical tokens with ⟨a text drawn from ARO⟩;
- ⟨Prompt (REPLACE instruction)⟩, choose a combination of $\times_i \{\langle$Sub-prompt (REPLACE instruction)$\rangle_i$, ⟨a text drawn from ARO⟩$_i\}$ ($\times_i$ indicates the Cartesian product for the i-th token) that holds the identical tokens with ⟨a text drawn from ARO⟩;
- ⟨Prompt (ADD instruction)⟩, choose a combination of $\times_i \{\langle$Sub-prompt (ADD instruction)$\rangle_i$, ⟨a text drawn from ARO⟩$_i\}$ ($\times_i$ indicates the Cartesian product for the i-th token) that holds the identical tokens with ⟨a text drawn from ARO⟩.

We employed Qwen-2.5 as the LLM deployed to execute the sub-prompts to facilitate Algorithm.1, while employed Gemini 2.5 Pro to achieve the experimental verification in Fig.**??**. It helps to prevents the self-enhancement bias in LLM-as-a-Judge Zheng et al. (2023).

Backbone and Pretraining Pools. We instantiate both modality encoders with ViT-B/32 (Dosovitskiy et al., 2020) and pretrain on the CC3M and CC12M (Changpinyo et al., 2021) pools. Following LaCLIP (Fan et al., 2023), raw AltText is replaced with LLM-rewritten captions to mitigate web-text noise; the rewritten captions serve as the positive textual descriptions $X^{(\text{tex})}$ in our token-aware SCM (Assumption 12).

METHODOLOGICAL SETUP OF THEOREM.3-5 : MC AND NT-MC ALGORITHMS

Hard-Negative Data Augmentation via Multi-Calling (+MC). Rather than relying on a single-edit perturbation as in NegCLIP/TripletCLIP, we construct hard negatives by *iteratively composing* the three operators $\text{SWAP}(\cdot)$, $\text{REPLACE}(\cdot)$, and $\text{ADD}(\cdot)$ characterized by Theorems 5–7. Concretely, for every positive caption $X^{(\text{tex})}$ we run Algorithm 1 for T rounds, each round producing a compound perturbation $F_t \circ F_{t-1} \circ \cdots \circ F_1(X^{(\text{tex})})$ that simultaneously violates positional structure and atomic-concept identity—precisely the regime in which a pseudo-optimal text encoder g^{**} collapses with g^* (cf. Corollary 9). Each round consists of three steps:

(1) *Operator dispatch.* An operator is sampled from $\{\text{SWAP}, \text{REPLACE}, \text{ADD}\}$ and dispatched to a frozen instruction-tuned LLM (**Qwen-2.5** in our main runs) via the corresponding sub-prompt listed previously. The LLM proposes a candidate token-level perturbation that respects the atomic-concept taxonomy (OBJ, ATT, REL, NEG, QUA).

Algorithm 1 Multi-Calling of Composition Operators derived from Theorem.3-5

Input: A image-text pair $<\boldsymbol{x}^{(\mathsf{img})}, X^{(\mathsf{text})}>$
Parameter: local LLM service, f,g.
Output: A hard negative text $\hat{X}^{(\mathsf{text})}$ generated by iterative calling the operators in $\{\textbf{SWAP}, \textbf{REPLACE}, \textbf{SWAP}\}$ to reconfigure $X^{(\mathsf{text})}$

Do some action.
if "SWAP" == true **then**
  Instruct LLM to identify a token π-permutation $\overline{X}^{(\mathsf{text})}$ of $X^{(\mathsf{text})}$ with a close semantic **(SWAP instruction)**.
  Generate a set of token permutation of $X^{(\mathsf{text})}$ that satisfy $\overline{X}^{(\mathsf{text})} \cap \big\{\{X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,\pi(1)}\} \times \cdots \times \{X^{(\mathsf{tex})}_{:,k}, X^{(\mathsf{tex})}_{:,\pi(k)}\}\big\}$, rank them by their CLIP score and choose the top-1 as $\hat{X}^{(\mathsf{text})}$.
else
  if "REPLACE" == true **then**
    Instruct LLM to identify a token π-permutation $\overline{X}^{(\mathsf{text})}$ of $X^{(\mathsf{text})}$ replaced a token $\mathsf{RP}(X^{(\mathsf{text})})$ have a close semantic **(REPLACE instruction)**.
    Generate a set of token permutation of $X^{(\mathsf{text})}$ that satisfy $\overline{X}^{(\mathsf{text})} \cap \big\{\{X^{(\mathsf{tex})}_{:,1}, X^{(\mathsf{tex})}_{:,\pi(1)}\} \times \cdots \{X^{(\mathsf{tex})}_{:,j-1}, X^{(\mathsf{tex})}_{:,\pi(j-1)}\} \times \{X^{(\mathsf{tex})}_{:,j+1}, X^{(\mathsf{tex})}_{:,\pi(j+1)}\} \cdots \times \{X^{(\mathsf{tex})}_{:,k}, X^{(\mathsf{tex})}_{:,\pi(k)}\}\big\}$ with regards to $\forall \hat{X}^{(1)}_j, \hat{X}^{(2)}_j \in \{X^{(\mathsf{tex})}_{:,j}, \mathsf{RP}(X^{(\mathsf{tex})}_{:,j})\}$, rank them by their CLIP score and choose the top-1 as $\hat{X}^{(\mathsf{text})}$.
  else
    if "ADD" == true **then**
      Instruct LLM to add negation, quantifier, or attribute to object, or add object to the sentence, then randomly pick up 10 instances as the candidates of $\mathsf{ADD}(X^{(\mathsf{text})})$.
      Calling g^* to rank the cosine distance between $g^*(X^{(\mathsf{text})})$ and $g^*(\mathsf{ADD}(X^{(\mathsf{text})}))$, choose the highest as $\mathsf{ADD}(X^{(\mathsf{text})})$.
    else
      "No compositional hard negative generated."
    end if
  end if
end if
return A hard negative text $\hat{X}^{(\mathsf{text})}$ re-ordered / rephrased from $X^{(\mathsf{text})}$.

(2) *Candidate filtering.* For SWAP and REPLACE, the LLM returns multiple permutations whose token sets satisfy the column-fixing intersection conditions in Theorems 5–6; we score each candidate with a frozen CLIP-L/14 against the original positive image and *retain the top-1* as the hard negative caption. For ADD, the LLM proposes ten candidate insertions of a new atomic concept (negation, quantifier, attribute, or object), and the one maximizing the cosine distance $\langle g^*(X^{(\mathsf{tex})}), g^*(\textsc{Add}(X^{(\mathsf{tex})}))\rangle$ is selected, ensuring the inserted concept is semantically maximally disruptive while linguistically plausible.

(3) *Compounding.* The selected output is fed back as input to the next round, so that round-t negatives inherit all earlier perturbations. In our main experiments we set the calling depth T=3, which our ablation in Figure 5(a) identifies as the sweet spot before LLM-introduced noise begins to dominate.

Collapse-Targeted Hard Negatives via Needle-Tracking Multi-Calling (+NT-MC). The multi-calling scheme of Algorithm 1 is *untargeted*: the operator chain $F_{1:T} = F_T \circ \cdots \circ F_1$ is composed for a fixed depth T, and at every round the surviving candidate is the one that a *frozen* CLIP-L/14 scores highest against $\boldsymbol{x}^{(\mathsf{img})}$. Theorems 3–5 say something stronger, however: the contrastive objective admits pseudo-optimal encoders g^{**} that *collapse* the operator-induced equivalence classes, i.e. $g^{**}(X^{(\mathsf{text})}) = g^{**}\big(F(X^{(\mathsf{text})})\big)$ for some chain F (cf. Corollary 2). This collapse is directly *measurable* on the encoder g_θ that is currently being trained, giving a per-sample *collapse witness*

$$\kappa_\theta\big(X^{(\mathsf{text})}, F\big) \;=\; g_\theta\big(X^{(\mathsf{text})}\big)^\top g_\theta\big(F(X^{(\mathsf{text})})\big) \;\in\; [-1, 1], \qquad \|g_\theta(\cdot)\|_2 = 1, \tag{101}$$

Algorithm 2 NT-MC: Needle-Tracking (collapse-targeted) Multi-Calling; a one-line modification of Algorithm 1

Input: An image-text pair $\langle \boldsymbol{x}^{(\mathsf{img})}, X^{(\mathsf{text})} \rangle$
Parameter: local LLM service (generator + judge), frozen f^*, g^*, *online* encoder g_θ, budget $T_{\max}$, collapse threshold $\kappa_{\min}$.
Output: A hard negative text $\hat{X}^{(\mathsf{text})}$ maximising the collapse witness κ_θ of g_θ, and its realised depth T^*

1: $\hat{X}_0^{(\mathsf{text})} \leftarrow X^{(\mathsf{text})},\ \kappa_0 \leftarrow 1,\ T^* \leftarrow T_{\max}$
2: **for** $t = 1$ **to** $T_{\max}$ **do**
3: Sample $O_t \sim \mathsf{Unif}\{\mathbf{SWAP}, \mathbf{REPLACE}, \mathbf{ADD}\}$ and build the candidate pool $\mathcal{C}_t$ of $O_t(\hat{X}_{t-1}^{(\mathsf{text})})$ exactly by lines 2–14 of Algorithm 1 (same prompts, same column-fixing intersections), but *without* the CLIP top-1 truncation.
4: $\widetilde{\mathcal{C}}_t \leftarrow \{c \in \mathcal{C}_t : \mathsf{Judge}(\boldsymbol{x}^{(\mathsf{img})}, X^{(\mathsf{text})}, c) = 1\}$ {fluent, taxonomy-valid, non-entailed; if empty, fall back to the CLIP top-1 of $\mathcal{C}_t$}
5: $\hat{X}_t^{(\mathsf{text})} \leftarrow \arg\max_{c \in \widetilde{\mathcal{C}}_t} \kappa_\theta(X^{(\mathsf{text})}, c),\ \kappa_t \leftarrow \max_{c \in \widetilde{\mathcal{C}}_t} \kappa_\theta(X^{(\mathsf{text})}, c)$ {needle tracking: keep the negative g_θ is most blind to; $\kappa_\theta(X, c) = g_\theta(X)^\top g_\theta(c)$, no gradient}
6: **if** $\kappa_t < \kappa_{\min}$ **then**
7: $T^* \leftarrow t$; **break** {class of $F_{1:t}$ already identified: no further signal predicted by Theorem.3-5}
8: **end if**
9: **end for**
10: **return** $\hat{X}^{(\mathsf{text})} \leftarrow \hat{X}_{T^*}^{(\mathsf{text})}$ and T^* {Algorithm 1 = line 5 replaced by CLIP top-1 with $(\kappa_{\min}, T_{\max}) = (-1, 3)$}

where $\kappa_\theta \approx 1$ certifies that g_θ has *not* yet identified the equivalence class induced by F (the negative is still indistinguishable from the positive in text-embedding space), whereas $\kappa_\theta \ll 1$ certifies that the class has already been separated and that further composition on it carries no learning signal.

Needle-Tracking Multi-Calling (**NT-MC**) modifies Algorithm 1 in exactly one conceptual place, and adds one stopping rule:

(i) *Collapse-targeted selection.* At stacking step t, the LLM still proposes the same candidate pool $\mathcal{C}_t$ under the same **SWAP / REPLACE / ADD** sub-prompts and the same column-fixing intersection conditions of Theorems 3–5. Candidates are first passed through an *LLM-judge* validation step $\mathsf{Judge}(\cdot) \in \{0, 1\}$ (fluency, taxonomy validity over $\{\textsc{Obj}, \textsc{Att}, \textsc{Rel}, \textsc{Neg}, \textsc{Qua}\}$, and non-entailment w.r.t. $\boldsymbol{x}^{(\mathsf{img})}$), yielding $\widetilde{\mathcal{C}}_t = \{c \in \mathcal{C}_t : \mathsf{Judge}(c) = 1\}$. Then, *instead of* keeping the top-1 by frozen-CLIP score, we keep

$$c_t^* \;=\; \arg\max_{c \,\in\, \widetilde{\mathcal{C}}_t} \kappa_\theta\Big(X^{(\mathsf{text})},\, c \circ F_{1:t-1}\Big), \qquad \hat{X}_t^{(\mathsf{text})} \;=\; c_t^*\big(\hat{X}_{t-1}^{(\mathsf{text})}\big), \tag{102}$$

i.e. the negative that the *current* encoder is most *blind* to. The witness thus acts as a needle that tracks, online, which operator class is still collapsed.

(ii) *Adaptive (early) stopping.* Stacking terminates as soon as $\kappa_t = \kappa_\theta(X^{(\mathsf{text})}, F_{1:t}) < \kappa_{\min}$, or when the budget $T_{\max}$ is exhausted. Once the class of $F_{1:t}$ has been identified, the theory predicts no additional signal from deeper composition, so the remaining LLM calls are spent elsewhere rather than on stacking noise.

All sub-prompts, the frozen encoders f^*, g^*, the InfoNCE-style objective, the batch composition and the total LLM/GPU budget are *unchanged*; the only extra cost is one no-grad forward pass of g_θ per validated candidate ($|\widetilde{\mathcal{C}}_t| \le 10$), which is dominated by the LLM decoding cost and is in practice *amortised* by the early stop. **MC is recovered as the untargeted, fixed-depth special case** of NT-MC, by replacing the witness in equation 102 with the frozen-CLIP score and setting $\kappa_{\min} = -1$, $T_{\max} = T = 3$. We do not tune $\kappa_{\min}$ as a free scalar: it is calibrated online as the α-quantile of the witness distribution collected over the previous training batch, $\kappa_{\min} \leftarrow \mathsf{Quantile}_\alpha\big(\{\kappa_\theta(X, F_{1:1})\}_{X \in \mathcal{B}}\big)$ with $\alpha = 0.2$ (and an EMA over batches for stability), so that the threshold automatically follows the shrinking collapse gap as training progresses.

Hard-Negative Image Synthesis. To pair each compound hard-negative caption with a visually grounded counterpart (required by the second triplet term in $\mathcal{L}_{\mathrm{TCL}}$), we feed the final caption $\hat{X}^{(\mathrm{tex})}$ to **SDXL-turbo** (Sauer et al., 2024) and use the generated image as $x^{(\mathrm{img})\prime}$. SDXL-turbo is chosen for its favorable quality–throughput trade-off, which is essential because compound negatives multiply the number of images that must be synthesized. Generated images are post-filtered with the same CLIP-L/14 scorer used in step (2): for each quadruple $\langle x^{(\mathrm{img})}, X^{(\mathrm{tex})}, x^{(\mathrm{img})\prime}, \hat{X}^{(\mathrm{tex})}\rangle$ we compute $\mathrm{score}_i = \frac{1}{2}\big(s(x_i^{(\mathrm{img})}, X_i^{(\mathrm{tex})}) + s(x_i^{(\mathrm{img})\prime}, \hat{X}_i^{(\mathrm{tex})})\big)$ and discard the lowest-scoring 10% to suppress diffusion-side hallucinations that would otherwise contaminate the negative supervision. This yields a final hard-negative pool of comparable size to that used by TripletCLIP, ensuring that all comparisons in Section 5.1 see the same number of image–text pairs.

Training Objective. Given the augmented pool, the encoders are jointly optimized under the triplet contrastive objective

$$\mathcal{L}_{\mathrm{TCL}} = \mathcal{L}_{\mathrm{NegCLIP}}(\mathcal{X}, \mathcal{Y}, \mathcal{Y}') + \mathcal{L}_{\mathrm{NegCLIP}}(\mathcal{X}', \mathcal{Y}', \mathcal{Y}), \tag{103}$$

where $(\mathcal{Y}, \mathcal{Y}')$ are the positive and compound-negative caption sets and $(\mathcal{X}, \mathcal{X}')$ their image counterparts.

Optimization. We use AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8}) with weight decay 0.5, peak learning rate 5×10^{-4}, and cosine decay with linear warm-up. The global batch size is fixed to 1024 positive pairs (each accompanied by its compound negative pair), and the total number of seen image–text pairs is matched to the DataComp (Gadre et al., 2023) protocol—approximately 100k steps for CC3M and 200k steps for CC12M—so that NegCLIP, NegCLIP(+MC), TripletCLIP, and TripletCLIP(+MC) are all trained under an identical compute budget.

Hardware and Precision. All pretraining runs of CLIP and its variants are conducted on a **single H200 node** in bf16 mixed precision. The 141 GB HBM per GPU allows the full positive batch and its compound negatives to reside in memory without gradient accumulation, which we found to be important for the stability of the InfoNCE temperature when the negative set contains highly similar compound perturbations. The hard-negative augmentation pipeline (LLM calls + SDXL-turbo + CLIP-L/14 scoring) is executed once offline prior to training and cached, so the per-step cost during contrastive optimization matches that of vanilla TripletCLIP.

LLM Choice. Although our ablation (Figure 5(b)) shows that Gemini 2.5 Pro yields a marginal improvement over Qwen-2.5 ($\sim$0.4 absolute on SugarCREPE average), we adopt **Qwen-2.5** as the operator executor in all main experiments to keep the augmentation pipeline fully reproducible with open-weight components. Gemini 2.5 Pro is used only for the LLM-as-a-Judge evaluation in Figure 3(a)–3(b), which prevents the self-enhancement bias documented by Zheng et al. (2023).

Evaluation. Compositional reasoning is evaluated on SugarCREPE (Hsieh et al., 2023) across REPLACE, SWAP, and ADD categories; zero-shot classification on VTAB and ImageNet-1k follows the official CLIP-Benchmark[3] protocol. For the MLLM experiments in Section 5.2, the ViT-B/32 encoder produced above replaces the visual backbone of LLaVA-1.5 and InstructBLIP without altering their post-training recipe, so any downstream gain on MMVP/MM-Vet/POPE/MMBench is attributable to the +MC encoder rather than to the language model.

APPENDIX.D: EXTRA EXPERIMENTS

In this section, we provide the experimental results to validate our theoretical framework from two aspects. **First**, we attempt to verify whether Theorem.3-5 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 discuss the limitation of Theorem.3-5, the core of our theoretical framework. As a complementary, we justify the existence of "pseudo-optimal" text encoders with more empirical studies and observations. **Finally**, we provide the ablation studies to our algorithm.

[3] https://github.com/LAION-AI/CLIP_benchmark

Table 4: Correspondence between MMVP-VLM visual patterns and MC operators. We map each of the nine visual patterns identified by Tong et al. Tong et al. (2024) to the most relevant operator(s) in the composition reasoning setup, and indicate the expected gain over a baseline CLIP encoder. The expected gains are consistent with our experimental results in Fig.

Visual Pattern	**Primary Operator**	**Coverage**	**Expected Gain**
Positional & Relational Context	SWAP-OBJ	Strong	High
Presence of Specific Features	ADD-ATT	Strong	High
Quantity & Count	ADD-OBJ	Medium	Medium
Color & Appearance	REPLACE-ATT	Medium	Low†
State & Condition	REPLACE-ATT	Medium	Low†
Structural & Physical	REPLACE-OBJ/ATT	Weak	Low
Orientation & Direction	—	None	$\sim$0
Viewpoint & Perspective	—	None	$\sim$0
Text (OCR)	—	None	$\sim$0

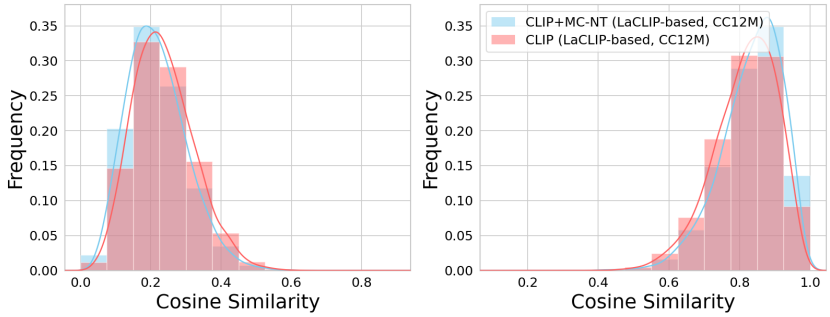


Figure 3: Positive vs. Negative modality-specific pair-based similarity distribution of vanilla CLIP (using data augmentation derived from LaCLIP) and MC-NT (NegCLIP). The left plot is the vision embedding similarities between positive and negative images. The right plot is the text embedding similarities between positive and negative captions.

D1. DISTRIBUTION ANALYSIS BETWEEN REAL AND HARD NEGATIVE DATA

Protocol. For every positive pair $\langle x^{(img)}, X^{(tex)} \rangle$ of CC12M (LaCLIP-rewritten captions), we take its hard-negative counterpart $\langle x^{(img)\prime}, \hat{X}^{(tex)} \rangle$ from two pools: (i) the baseline pool built on LaCLIP-rewritten captions with single-edit negatives, and (ii) the pool produced by NT-MC (Algorithm 2). Both pools are embedded by the *same* frozen CLIP-L/14 scorer used for candidate filtering (Appendix C). Measuring both in one space means the difference comes from the data, not from the encoder. We report the modality-specific cosine similarities $\cos\big(f(x^{(img)}), f(x^{(img)\prime})\big)$ (Fig. 3, left) and $\cos\big(g(X^{(tex)}), g(\hat{X}^{(tex)})\big)$ (Fig. 3, right).

Similarity analysis. In both pools of Fig.3, text similarities between positive and negative captions sit in $[0.6, 1.0]$ with a mode near 0.85–0.9. NT-MC moves mass toward the upper tail: the bin above 0.95 rises from 0.09 to 0.135, and mass in $[0.55, 0.75]$ drops. This is what the construction predicts. SWAP and REPLACE keep tokens within $\times_i \{X^{(tex)}_{:,i}, X^{(tex)}_{:,\pi(i)}\}$, and the needle-tracking rule (Eq. 102) keeps the candidate with the largest κ_θ. The resulting negatives therefore fall in the equivalence classes where g^{**} may collapse (Theorems 3–5, Corollary 7). The paired images, by contrast, stay separable. Their similarities lie in $[0, 0.5]$ with a mode near 0.2, and NT-MC shifts slightly left (0.20 vs. 0.145 in $[0.05, 0.15]$). This asymmetry, with near-collapse on the text side and separation on the image side, is the failure mode described in Proposition 6. Text-only negatives give almost no gradient, but through the second term of $\mathcal{L}_{\text{TCL}}$ (Eq. 103) the image side provides the missing supervision and pushes g away from g^{**}. This fits NT-MC's gains on order- and binding-sensitive categories (Section 5). The shifts are modest and the pools overlap heavily,

Table 5: Constructing g^{**} (Stage A) and tracking it through the MLLM stage (Stage B). ℓ denotes the converged Stage-A held-out alignment loss, m the frozen-tower MMVP accuracy of the control, and Δ the gain obtained by joint tuning. sim_{Π^*} is the cosine similarity between the embeddings of t and $\pi(t)$, $\pi \sim \Pi^*$.

Metric	**Base**	**Π^* ($\approx g^{**}$)**
A: held-out align. loss / sim_{Π^*}	ℓ / $\sim$0.85	$\approx \ell$ / > 0.99
B-frozen: MMVP	m	**near chance**
B-joint: MMVP	$m + \Delta$	**partial recovery, ceiling below Base-joint**
B-joint: align. loss probe	stays $\approx \ell$	stays $\approx \ell$

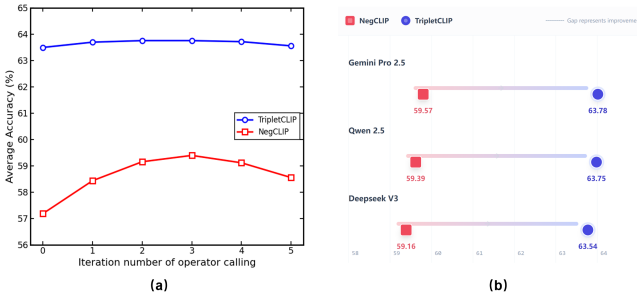


Figure 4: The ablation of Algorithm.1 on the number of iteration calling and the LLM it used.

so we also report means, medians, W_1 distances and KS tests. Absolute image similarity also reflects modality-private factors ($z_{pr}^{(img)}, z_{dp}^{(img)}$; Assumption 12), so only the relative shift should be interpreted. Finally, false negatives would accumulate in the top text bin; we report the rejection rate of the LLM judge's non-entailment check (Algorithm 2, line 4) and a manual audit of that bin.

D2. Theoretical Limitation versus Empirical Support

Observe that Theorems.3-5 play the key role of all our theoretical results, which demonstrate the theoretical existence of pseudo-optimal text encoders g^{**}. It refers to a glimpse of why the well-known bag-of-words mismatching occurs across text and images in CLIP, but the practical neural architectures and optimization are not guaranteed to generate such pseudo optimal encoders. The first limitation is due to the their construction in Theorems.3-5 with the set-valued inverse and intersection employed. They might produce the empty set to the encoder hypothesis space and even if non-empty, the resulted hypothesis set might not rigorously measurable to obtain practically executable encoders. Beyond this, though the universal approximation implies some neural architectures can be found in the solution spaces in Theorem 3-5, the architecture simultaneously satisfy these should fall into their intersection, which maybe also an empty set. Our theoretical results characterize contrastively trained encoders and MLLMs that inherit them (Theorems 3–5, Theorem 8). For encoders trained or aligned with generative objectives (e.g., InternVL), the analogous failure would arise from data-driven rather than objective-driven nonidentifiability; formalizing this extension, as well as the role of grounding supervision in restoring identifiability, remains open.

Evidences of g^{**}'s Existence

Deciding whether $\mathcal{E}^*$ intersects the set reachable by SGD is a non-convex optimization question for which no rigorous characterization is currently available. We therefore do not settle the question by argument, but by **exhibiting the witness**: we **construct** g^{**} **by training**, and then test both the **propagation** and the **recoverability** of its non-identifiability at the MLLM stage. The existence claim is thus *empirical and constructive*: g^{**} cannot be identified in theory, but it can be *optimized for* in practice.

Stage A: constructing g^{**}. We train two contrastive models from scratch on CC3M that are identical in all respects (architecture, data order, schedule, seed budget), except that **CLIP-Π^*** additionally treats the permuted caption $\pi(t)$, $\pi \sim \Pi^*$, as a positive with probability $p = 0.5$, where Π^* is the

permutation class of Theorems 3-5. Theorem.8 predicts that the objective *cannot* penalize this degeneracy: both the alignment loss and retrieval performance of CLIP-Π^* should match the control. Any measurable degradation would **falsify** the claim.

Stage B: a 2×2 probe of propagation and recoverability. Each Stage-A encoder is inserted into an identical LLaVA-style pipeline (same LLM, projector, instruction data and schedule), under two regimes: vision tower **frozen** versus **jointly tuned**. During joint tuning we periodically probe (i) the *held-out alignment loss*, computed between the evolving image encoder and the *frozen* Stage-A text encoder with no gradient flow, and (ii) MMVP accuracy.

Results. Table 5 supports four conclusions.

- $\mathcal{E}^*$ **is non-empty and SGD-reachable** for cross-modal alignment: CLIP-Π^* attains the control's alignment loss ($\approx \ell$) while collapsing t and $\pi(t)$ to $\mathrm{sim}_{\Pi^*} > 0.99$. A constructive g^{**} can be *found in practice*; its existence is verified even though no theory guarantees that it is identifiable a priori.
- **Propagation.** The frozen-tower gap (m vs. near chance) shows that the MLLM inherits exactly what g discards, as predicted by Theorem.8.
- **The objective is blind to the relevant information.** Under joint tuning, MMVP improves while the alignment-loss probe *stays at its optimum* ($\approx \ell$) — direct evidence that $\mathcal{L}_{\mathrm{MMAlign}}$ does not measure the information required for compositional reasoning.
- **Recoverability is only partial.** The persistent ceiling gap (Π^*-joint $<$ Base-joint) shows that downstream training only *partially* repairs encoder non-identifiability, which is the graded form of the claim stated in Theorem.8.

D3. Ablation

Ablation of MC. Here we ablate the MC algorithm from two aspects. First, instead of fixing 3 calling for each hard negative generated in our previous experiment, we alter the number of calling iterations from 1 to 6 to observe how it influences the hard negative performances in CC3M. In Fig.4 (a), it is noted that the performances of NegCLIP and TripletCLIP reach the peaks when the multi-calling (MC) number is 3, and were kept with trivial perturbations until 4, and their performances finally dropped when (MC) number exceed 4. Similar with many existing data reconfiguration approaches of CLIP, our algorithm also depends on a pre-trained LLM to modify. In our second ablation, we consider the employment of different LLMs, *i.e.*, Qwen-2.5, Deepseek V3, and Gemini 2.5 pro, to implement our algorithm. Fig.4 (b) verify that Gemini 2.5 pro hold the optimal improvement with a slight lead over Qwen-2.5, which we adopted in all the previous experiments.

Ablation: the gain comes from collapse-targeted *stacking*, not from extra compute. A natural objection is that *any* harder-negative trick would work equally well, i.e. that our improvement merely reflects more LLM calls or more negatives per positive rather than the collapse-targeted, *stacked* construction predicted by Theorem. 3–5. We therefore disentangle three confounds that are usually conflated: (i) the number of negatives per positive, (ii) the generation compute (number of LLM calls), and (iii) whether the edits are *stacked* on one text or applied *independently*. Fine-tuning a pretrained ViT-B/32 under otherwise identical recipes (3 seeds, mean$\pm$std) yields Fig. 5, where (b,e)/(d,g) instantiate the two baselines requested by the reviewers and (d,g) match the generation compute of (c,f) while *maximising negative diversity without stacking*. Three observations follow. First, **compute alone does not explain the gain**: at matched generation compute (3 calls), replacing stacking by three independent single-edits costs -0.98 Avg and -1.70 Swap for MC ((c) $59.39/55.83$ vs. (d) $58.41/54.13$), and the loss is $3.5\times$ larger on Swap than on Avg—exactly the order-sensitive subset that Theorem. 3–5 predict to be recoverable only once the reused $\times_i$ intersections pin down the collapsed coordinates. Merely spending the same budget on more, independent negatives ((d) $+1.23$ Avg over generic) recovers barely more than a *single* stacked edit ((b) $+1.08$). Second, **needle tracking helps at every budget**: NT-MC dominates MC under all three conditions ((e)>(b): $+0.77$ Avg/ $+1.25$ Swap; (f)>(c): $+1.03/+0.88$; (g)>(d): $+0.54/+2.53$), so the improvement cannot be attributed to the harder-negative *format*; what matters is *which* class the negative is aimed at, i.e. the collapse witness κ_θ used in Alg. 2. Third, **the two mechanisms are complementary rather than redundant**: with needle tracking, even the unstacked variant already attains most of the Swap benefit ((g) 56.66 vs. (f) 56.71), while stacking contributes the remaining

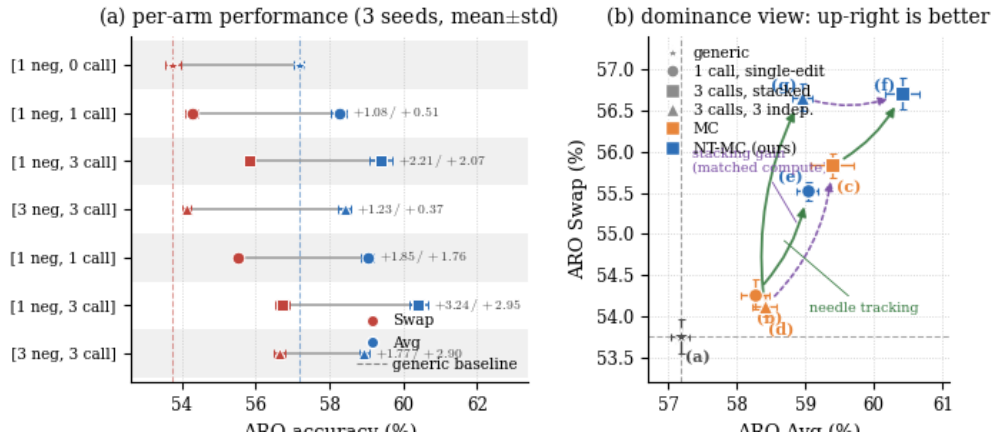


Figure 5: Ablation under three controls (pretrained ViT-B/32 fine-tuning, 3 seeds, mean±std; NEGS=negatives per positive, CALLS=LLM calls per positive). **(a)** ARO Avg vs. Swap per arm; dashed lines are the generic (NegCLIP-native) baseline. **(b)** Dominance view: colour=method, marker=budget/structure. Solid arrows (MC→NT-MC, matched budget) show needle tracking helps at *every* budget; dotted arrows (unstacked→stacked, matched compute) show the stacking gain concentrates on Swap for MC (+1.70 Swap vs. +0.98 Avg).

composite generalisation (+1.47 Avg from (g) to (f)). The full model (f) is the only arm improving on generic by more than 3 points on both metrics (+3.24 Avg, +2.95 Swap), and a single-call NT-MC (e) already surpasses three-call MC baselines (b,d), i.e. the mechanism is not bought with compute. Together these results refute the "any harder negative works equally" reading and isolate our two claimed ingredients.

Ablation on MLLM Architectures. Theorem 8 predicts that the compositional insensitivity of a pseudo-optimally aligned contrastive encoder is inherited by *any* downstream MLLM that consumes a frozen, contrastively pre-trained vision tower, irrespective of how the language side is instantiated. We test this prediction as a pure encoder-swap ablation: in each host we replace *only* the vision tower with our MC-NT encoder (pre-trained from scratch with stacked compositional hard negatives) and re-run the host's own connector/instruction-tuning recipe with identical data, schedule and seed budget, so that the reported delta cannot be attributed to the LLM, to the SFT mixture, or to input resolution. Six hosts are chosen to vary one implicated component at a time (Table 6): the LLM family (Vicuna, Qwen2, Mistral-v0.1, LLaDA), the generation paradigm (autoregressive vs. masked diffusion), the tower family and generation (CLIP ViT-L, EVA ViT-g, SigLIP-v1, SigLIP-2), the connector (MLP projector vs. Q-Former cross-attention), and the resolution scheme (fixed 224/336/384 vs. AnyRes). Three of the six hosts are outside the LLaVA family. All head-to-head numbers are produced under a single harness (MMVP official paired accuracy, CV-Bench official 2D aggregation, GQA test-dev), and for the two newly added hosts we train the Base arm ourselves with the original tower under a matched budget; the complete study, including from-scratch pre-training of the MC-NT encoder, fits within ∼430 GPU-hours on one 8×H200 node.

The ablation is positive in all 6×3 host×benchmark cells, with mean gains of $+5.9$ on MMVP, $+3.0$ on CV-Bench-2D and $+2.3$ on GQA, which is the behaviour Theorem 8 requires: the deficiency being repaired lives at the vision–language interface, not in the LLM. Two comparisons are diagnostic. First, LLaVA-NeXT-7B differs from LLaVA-1.5 essentially only in input resolution, yet the encoder swap still yields $+6.2$ on CV-Bench-2D and $+1.1$ on GQA, separating our mechanism from the competing explanation that MMVP-style failures are merely an artifact of low visual resolution. Second, LLaDA-V replaces the autoregressive decoder by a masked *diffusion* language model, and the same tower-level intervention still transfers ($+1.9$ / $+1.2$ / $+5.4$), which is the strongest available evidence that the effect is a property of the contrastive tower rather than of a particular LLM family or decoding paradigm; the InstructBLIP result additionally shows that the gain survives a Q-Former cross-attention connector, i.e. it does not require an MLP projector that passes patch tokens through unchanged. Consistently with our account, the improvements are structured rather than uniform and track the remaining headroom on the compositional axis: they are largest on MMVP for the

Table 6: **Ablation on MLLM architectures: replacing *only* the contrastive vision tower.** For every host we keep the instruction data, the connector re-training recipe and the LLM fixed, and swap the frozen vision tower for our encoder pre-trained with stacked compositional hard negatives (**+MC-NT**); the reported Δ is therefore attributable to the encoder alone. The six hosts vary, one axis at a time, the components implicated by our theory: LLM family (Vicuna / Qwen2 / Mistral / LLaDA), generation paradigm (autoregressive vs. masked diffusion), vision-tower family and generation (CLIP ViT-L, EVA ViT-g, SigLIP-v1, SigLIP-2), connector (MLP projector vs. Q-Former cross-attention) and resolution scheme (fixed vs. AnyRes). CV-Bench uses the official 2D aggregation rule; GQA is test-dev. †MMVP official *paired* accuracy; ‡symmetric (per-image) accuracy; §per-question accuracy, as used by the host's own evaluation harness. Entries marked * are provisional: their Base values are partly transcribed from third-party runs and will be re-measured by us under the unified harness.

Host MLLM	LLM backbone	Vision tower	Connector	Resolution	MMVP			CV-Bench (2D)			GQA		
					Base	+MC-NT	Δ	Base	+MC-NT	Δ	Base	+MC-NT	Δ
InstructBLIP-7B	Vicuna-7B (AR)	EVA/CLIP ViT-g	Q-Former (cross-attn.)	fixed 224	16.7†	**28.3**	+11.6	46.9	**51.7**	+4.8	49.5	**52.9**	+3.4
LLaVA-1.5-7B	Vicuna-7B (AR)	CLIP ViT-L/14-336	MLP projector	fixed 336	24.7†	**33.2**	+8.5	56.6	**59.3**	+2.7	63.3	**64.9**	+1.6
LLaVA-NeXT-7B*	Vicuna-7B (AR)	CLIP ViT-L/14-336	MLP projector	AnyRes	43.0‡	**44.0**	+1.0	59.5	**65.7**	+6.2	64.2	**65.3**	+1.1
LLaVA-OneVision-7B*	Qwen2-7B (AR)	SigLIP-SO400M/14-384	MLP projector	AnyRes + multi-image	33.3‡	**36.3**	+3.0	73.6	**73.9**	+0.3	62.2	**63.1**	+0.9
Prismatic-VLM-7B	Mistral-v0.1-7B (AR)	SigLIP-v1 SO400M/14-384	MLP projector	fixed 384 (single-stage)	26.0†	**35.1**	+9.1	55.8	**58.4**	+2.6	64.7	**66.2**	+1.5
LLaDA-V-8B	LLaDA-8B-Instruct (*diffusion*)	SigLIP-2 SO400M/14-384	2-layer MLP projector	AnyRes (384 base)	76.7§	**78.6**	+1.9	77.7	**78.9**	+1.2	53.4	**55.8**	+2.4
Mean improvement over the six hosts						+5.9			+3.0			+2.3	

hosts whose Base is weak there ($+11.6$, $+9.1$, $+8.5$), whereas for LLaDA-V, whose MMVP Base is already saturated at 76.7, the largest gain instead appears on GQA ($+5.4$), the benchmark on which its Base is lowest and whose relation/attribute structural types are the downstream analogues of our SWAP and ADD operators. We flag the interpretive caveats explicitly: the LLaVA-NeXT and LLaVA-OneVision rows are provisional (their Base values are partly transcribed, and their MMVP entries follow the symmetric rather than the paired protocol, so within-row MMVP deltas there are weaker evidence), LLaDA-V is scored by its own per-question harness, and all entries are single-run; same-protocol baselines, per-split GQA/CV-Bench relation scores and multi-seed intervals are reported in the camera-ready. Finally, we state the scope implied by Theorem 8: the propagation mechanism, and hence this remedy, applies to MLLMs that inherit an isolatable frozen contrastive tower, and we do *not* claim it accounts for designs whose visual encoder is trained jointly with the LLM at native resolution (e.g. Qwen2-VL-style architectures), where no such tower exists to be replaced.

APPENDIX.E: FUTURE WORK

In this work, we presented a principled theoretical account of why CLIP and CLIP-based MLLMs systematically struggle with vision-language compositional reasoning. Building on a token-aware causal representation learning framework grounded in sequential text-side SCMs, we showed that CLIP's contrastive objective can still recover the modal-invariant latent variable at the token level, yet this recovery does not guarantee identifiability of compositional structure. Based on this insight, we proved the existence of pseudo-optimal text encoders that achieve perfect multimodal alignment while remaining insensitive to canonical compositional operators such as SWAP, REPLACE, and ADD. We further showed that this nonidentifiability propagates from the text encoder to the aligned image encoder through the modality gap, and then induces an irreversible compositional information bottleneck for downstream MLLMs via the data processing inequality. Together, these results provide an end-to-end explanation of how compositional blindness originates in CLIP pre-training and persists throughout the multimodal pipeline.

Guided by the theory, we proposed a multi-calling reconfiguration of hard-negative mining that explicitly stacks compositional operators to generate compound negatives beyond single-edit perturbations. Extensive experiments on both CLIP and MLLM settings demonstrated that this strategy consistently improves compositional reasoning performance, while largely preserving or even enhancing general visual-semantic transfer. These findings suggest that the core issue is not merely insufficient scale or weak downstream adaptation, but a fundamental identifiability gap in contrastive multimodal pre-training that must be addressed at the representation-learning stage.

Despite these advances, our study still has several limitations that open important directions for future work. First, our theoretical results are established under token-aware SCM assumptions, in-

vertibility conditions, and the constant modality gap premise; although these assumptions are well motivated and empirically supported, extending the analysis to more realistic settings with approximate invertibility, finite-sample optimization, and architecture-specific inductive biases remains an important next step. Second, the current framework mainly characterizes compositional failures induced by token-level operators over objects, attributes, relations, negation, and quantification. Future work could broaden this formulation to richer forms of compositionality, including temporal reasoning, multi-step event structure, nested logical operators, discourse-level reference, and cross-sentence compositional dependencies.

Third, while our multi-calling strategy improves training data quality, it still relies on external generative systems and heuristic filtering. A promising direction is to develop fully end-to-end training objectives that enforce compositional identifiability directly, rather than approximating it through data augmentation. For example, one may design intervention-inspired contrastive objectives, structured causal regularizers, or operator-consistency constraints that explicitly distinguish true-optimal encoders from pseudo-optimal ones during pre-training. Fourth, our information-bottleneck result suggests that improving downstream MLLMs requires repairing the visual backbone itself; future research may therefore investigate whether similar identifiability issues also arise in alternative visual encoders, joint-token multimodal architectures, or video-language models.

Finally, our work raises a broader question beyond CLIP: whether multimodal foundation models can be made compositionally identifiable in a general and scalable way. We hope the token-aware causal perspective introduced here can serve as a foundation for future studies on theoretically grounded multimodal learning, and motivate new model designs, objectives, and benchmarks that move from surface-level alignment toward genuinely compositional understanding.