

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

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

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

## 1 INTRODUCTION

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

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

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

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

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

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

## 2 PRELIMINARIES

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

### 2.1 CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING (CLIP)

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

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

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

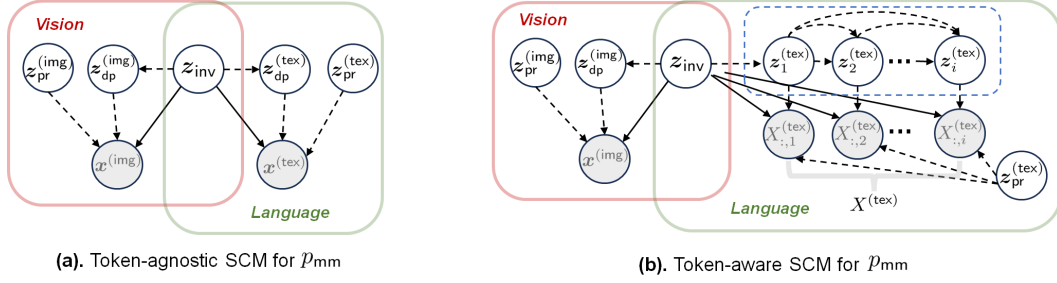


Figure 1: Latent-variable SCMs that represents the multimodal image-text data generation processes from the sentence-level aspect (Assumption 1 (a)) and the token-level aspect (Assumption 4 (b)). The goal of causal representation learning seeks for the unsupervised recovery of the modal-shared latent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

## 2.2 CONVENTIONAL CAUSAL REPRESENTATION FOR MULTIMODAL CONTRASTIVE TRAINING

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

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

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

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

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

$$\begin{aligned} \mathcal{L}_{MMAAlign}^{(img, tex)} &:= \mathbb{E}_{\langle x^{(img)}, x^{(tex)} \rangle \sim p_{mm}} \left[ \|f(x^{(img)}) - g(x^{(tex)})\| \right] \\ &\quad - H(f(x^{(img)})) - H(g(x^{(tex)})) \end{aligned} \quad (3)$$

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

$$f^* = h_f \circ \mathbf{f}_{1:n_{inv}}^{-1}, \quad g^* = h_g \circ \mathbf{g}_{1:n_{inv}}^{-1} \quad (4)$$

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

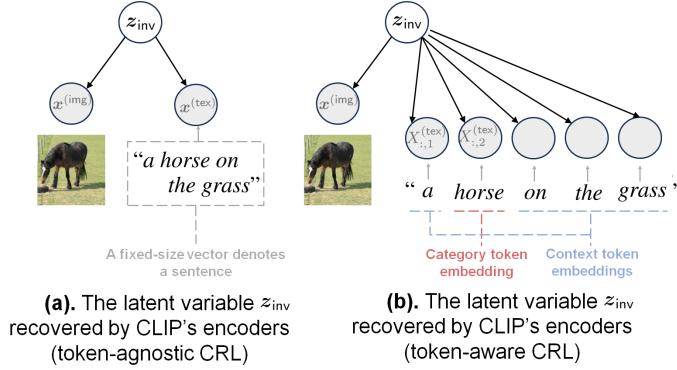


Figure 2: The comparison between (a) existing multimodal CRL theory (Daunhawer et al. (2022)) and (b) our CRL theory (Theorem.5 and Corollary.6). Our framework allows the analysis to CLIP with the word-and-phrase granularity, leading to our contributions to theoretically explain the CLIP weakness in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

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

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

### 3 LANGUAGE-TOKEN-AWARE CAUSAL REPRESENTATION: CORNERSTONE TO INTERPRET COMPOSITIONAL REASONING HARDNESS

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

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

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

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

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

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

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

$$\mathcal{L}_{\text{MMAAlign}}^{(\text{img}, \text{tex})} := \mathbb{E}_{\langle \mathbf{x}^{(\text{img})}, \mathbf{X}^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left[ \left\| f(\mathbf{x}^{(\text{img})}) - g(\mathbf{X}^{(\text{tex})}) \right\|^2 \right] - H(f(\mathbf{x}^{(\text{img})})) - H(g(\mathbf{X}^{(\text{tex})})), \quad (6)$$

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

$$f^* = h_f \circ \mathbf{f}_{1:n_{\text{inv}}}^{-1}, \quad g^* = h_g \circ \mathbf{g}_{1:n_{\text{inv}}}^{-1} \quad (7)$$

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

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

## 4 COMPOSITION NONIDENTIFIABILITY IN CLIP

As observed in existing research, CLIP is born vulnerably to identify the language compositional difference in an image-text pair. While such concrete definition could be shifted across specific literature. Our study focuses on the definition used to build CREPE (Ma et al. (2023)) and SUGARCREPE (Hsieh et al. (2023)): for an image-text pair  $\langle \mathbf{x}^{(\text{img})}, \mathbf{X}^{(\text{tex})} \rangle$ , they considered the tokenized word or phrase (i.e.,  $X_{i,:}^{(\text{tex})}$ , a column of token-embedding matrix  $\mathbf{X}^{(\text{tex})}$ ) 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  $\mathbf{X}^{(\text{tex})}$  can be categorized into three formats.

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

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

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

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

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

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

$$\mathcal{L}_{\text{MMAAlign}}^{(\text{img}, \text{tex})}(f^*, g^*) \rightarrow 0 \quad (8)$$

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

$$\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^{**}) \rightarrow 0 \quad (9)$$

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

$$g^{**}([X_{:,1}^{(\text{tex})}, X_{:,2}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\pi(1)}^{(\text{tex})}, X_{:,\pi(2)}^{(\text{tex})}, \dots, X_{:,\pi(k)}^{(\text{tex})}]), \quad (10)$$

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

$$g^{**}([X_{:,1}^{(\text{tex})}, X_{:,2}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\hat{\pi}(1)}^{(\text{tex})}, X_{:,\hat{\pi}(2)}^{(\text{tex})}, \dots, X_{:,\hat{\pi}(k)}^{(\text{tex})}]), \quad (11)$$

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

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

$$g^{**}([X_{:,1}^{(\text{tex})}, \dots, X_{:,j}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\pi(1)}^{(\text{tex})}, \dots, \text{RF}(X_{:,j}^{(\text{tex})}), \dots, X_{:,\pi(k)}^{(\text{tex})}]), \quad (12)$$

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

$$g^{**}([X_{:,1}^{(\text{tex})}, \dots, \hat{X}_j^{(1)}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\hat{\pi}(1)}^{(\text{tex})}, \dots, \hat{X}_j^{(2)}, \dots, X_{:,\hat{\pi}(k)}^{(\text{tex})}]). \quad (13)$$

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

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

$$\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^*) \rightarrow 0 \quad (14)$$

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

$$\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^{**}) \rightarrow 0 \quad (15)$$

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

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

then it holds

$$g^{**}([X_{:,1}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,1}^{(\text{tex})}, \dots, X_{:,j}^{(\text{tex})}, \text{ADD}(X_{:,j}^{(\text{tex})}), \dots, X_{:,k}^{(\text{tex})}]). \quad (16)$$

**Interpretation.** The statements and proof sketches in Theorems.7, 8, and 9 resemble the spirit of using Theorem. and Corollary.6 to construct a “pseudo-optimal” text encoder  $g^{**}$  that occur when the “true-optimal” text encoder  $g^*$  could be practically obtained by the causal representation of CLIP. In this situation,  $g^*$  and  $g^{**}$  can simultaneously achieve the modal-invariant alignment (i.e.,  $\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^*) \simeq 0$  and  $\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^{**}) \simeq 0$ ) 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.7-9. Since the encoders  $g^*$  and  $g^{**}$  share the same architecture and their parameters both achieve modal-invariant alignment during

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

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

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

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

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

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

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

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

## 5 EXPERIMENTS

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



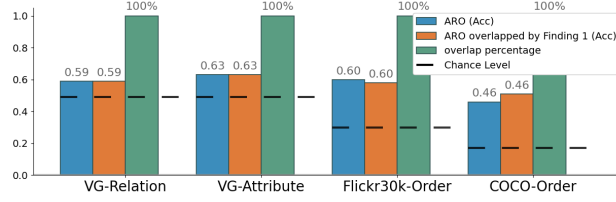


Figure 3: CLIP’s accuracy (ACC) on the negative samples generated by ARO and our Algorithm1. The overlap percentage indicates how many negative samples in ARO belong to the cases in Theorem.7-9.

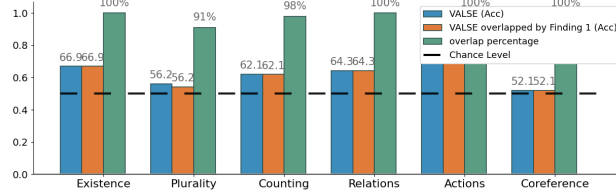


Figure 4: CLIP’s accuracy (ACC) on the negative samples generated by VALSE and our Algorithm1. The percentage indicates how many negative samples in VALSE belong to the cases in Theorem.7-9.

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

### 5.1 BRIDGING THEORETICAL-EMPIRICAL GAPS ON BENCHMARK DATA

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

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

### 5.2 EVIDENCES OF $g^{**}$ ’S EXISTENCE

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



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

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

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

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

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

### 5.3 MULTI-CALLING OF COMPOSITION OPERATORS

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

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

## REFERENCES

- Kartik Ahuja, Karthikeyan Shanmugam, and Kush R Varshney. Learning identifiable and interpretable latent models of high-dimensional data. *arXiv preprint arXiv:2002.02893*, 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.
- 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.
- Ashita Bhargava, Jize Zhou, Jieyu Zhang, Cheng-Yu Hsieh, and Ranjay Krishna. Attrprompt: A new data-centric paradigm for probing and improving attribute-object compositionality of vision-language models. *arXiv preprint arXiv:2303.14237*, 2023.
- Johann Brehmer, Julius Von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Weakly supervised causal representation learning. *arXiv preprint arXiv:2010.15794*, 2022.
- Simon Buchholz, Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Learning identifiable representations that support sample-efficient intervention. *arXiv preprint arXiv:2302.01828*, 2023.
- Colin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent concepts in language models with contrastive search. *arXiv preprint arXiv:2210.14922*, 2023.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.
- Ziliang Chen, Xin Huang, Quanlong Guan, Liang Lin, and Weiqi Luo. A retrospect to multi-prompt learning across vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22190–22201, 2023.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.
- 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.
- Rohith Gandikota, Mert Yükeşgönül, Yonatan Bisk, and Jacob Baldridge. Compositional learning of vision-language concepts. *arXiv preprint arXiv:2306.04833*, 2023.
- Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021.
- Luigi Gresele, Julius von Kügelgen, Ricardo P Monti, Bernhard Schölkopf, and Kun Zhang. Causal discovery in a binary setting with interventions. *arXiv preprint arXiv:2010.14241*, 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.
- Bohdan Kivva, Marc Vuffray, and Bryon Aragam. Identifiability of latent-variable models with arbitrarily many views. *arXiv preprint arXiv:2210.00063*, 2022.
- 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, Ari Holtzman, Rowan Zellers, and Yejin Choi. Aflite: A lightweight framework for adversarial filtering. *arXiv preprint arXiv:2009.09262*, 2020.
- Florian Leeb, Julius von Kügelgen, Bernhard Schölkopf, and Michel Besserve. Causal concept embedding models. *Advances in Neural Information Processing Systems*, 35:23668–23681, 2022.
- 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 P Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with hidden confounders using independent component analysis. *arXiv preprint arXiv:1906.08773*, 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.
- Zixian Pan, Jieyu Zhang, Cheng-Yu Hsieh, and Ranjay Krishna. The noun-verb ambiguity of open-domain images. *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, 2022.
- Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.
- Dong-Sub Park, Wilson Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Eduard Hovy, and Quoc V Le. Speech-t5: Unifying speech generation and speech recognition via a single t5-based model. *arXiv preprint arXiv:2110.07205*, 2021.
- Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, et al. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760, 2024.

- Sydney Pratt, Corentin Kervadec, Sébastien Gontier, Emmanuel Dellandrea, Thierry Robert, Anirudh Goyal, Yoshua Bengio, and Massimo Caccia. Swig: A benchmark for in-the-wild compositional generalisation. *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, 2022.
- 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.
- Aida Ravichander, Eduard Hovy, and Richard M Pang. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:2004.09384*, 2020.
- 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.
- 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.
- Sebastian Schwettmann, Florian Leeb, Bernhard Schölkopf, and Michel Besserve. Concept embedding models: A case study in toxicology. *arXiv preprint arXiv:2301.11823*, 2023a.
- Sebastian Schwettmann, Florian Leeb, Bernhard Schölkopf, and Michel Besserve. Towards a theoretical framework for concept discovery. *arXiv preprint arXiv:2305.18728*, 2023b.
- Anna Seigal and Yuesong Shen. Identifiability of deep generative models with structural constraints. *arXiv preprint arXiv:2006.07899*, 2021.
- Chandler Squires, Yue Wu, Kun Zhang, and Bryon Aragam. Causal-learn: Causal discovery in python. *The Journal of Machine Learning Research*, 24(1):14781–14787, 2023.
- 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.
- Kun Su and Dan Yu. Negation-aware contrastive learning for sentence representation. *arXiv preprint arXiv:2205.11581*, 2022.
- 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.
- Kush R Varshney. On the identifiability of nonlinear latent variable models. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2402–2406, 2017.
- Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Non-linear identifiability of causal representations from temporal sequences. *arXiv preprint arXiv:2006.15059*, 2021.
- Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.
- Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing Systems*, 34:16158–16170, 2021.
- Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

- Mert Yüsekşönül, Rohith Gandikota, Jacob Baldridge, Dilek Erhan, Yonatan Bisk, and Gokhan Tur. Vision-and-language models are not compositional'out-of-the-box'. *arXiv preprint arXiv:2210.03494*, 2022.
- 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.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.10703*, 2019.
- 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.
- 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.

## A APPENDIX.A

### A.1 RELATED WORK AND BACKGROUND

In this section, 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 compositionality 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 Yüsekşgönül et al. (2022); Bhargava et al. (2023); Hu et al. (2023); Gandikota et al. (2023); Su & Yu (2022). A prominent approach involves explicitly training models to differentiate correct captions from synthetically generated hard negatives Yüsekşgönül et al. (2022); Gandikota et al. (2023). 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 Pan et al. (2022); Zellers et al. (2019); Le Bras et al. (2020); Ross et al. (2017); Pratt et al. (2022). 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 Zellers et al. (2019); 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.

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

To pave the way for understanding the specific assumption for multimodal data, let’s first define a general SCM using a consistent LaTeX format. This will introduce the core components and notation, which are then specialized in the assumption you provided.

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

$\mathbf{V} = V_1, \dots, V_n$  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  $x^{(\text{img})}$  and a text description  $x^{(\text{tex})}$ , are considered endogenous.

$\mathbf{U} = U_1, \dots, U_n$  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 ran-

domness in the model. In your assumption, the latent variables  $z_{\text{inv}}$ ,  $z_{\text{dp}}^{(\text{img})}$ ,  $z_{\text{pr}}^{(\text{img})}$ ,  $z_{\text{dp}}^{(\text{tex})}$ , and  $z_{\text{pr}}^{(\text{tex})}$  can be thought of as being determined by exogenous sources of variation.

$\mathcal{F} = f_1, \dots, f_n$  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  $\text{pa}(V_i) \subseteq V \setminus V_i$  and its associated exogenous variable  $U_i$ :

$$V_i := f_i(\text{pa}(V_i), U_i) \quad (17)$$

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

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:  $z_{\text{inv}} \sim p_{z_{\text{inv}}}$ : The modal-invariant feature.  $z_{\text{pr}}^{(\text{img})} \sim p_{z_{\text{pr}}^{(\text{img})}}$ : The image-private feature.  $z_{\text{pr}}^{(\text{tex})} \sim p_{z_{\text{pr}}^{(\text{tex})}}$ : The text-private feature. The dependent partitions,  $z_{\text{dp}}^{(\text{img})}$  and  $z_{\text{dp}}^{(\text{tex})}$ , are also influenced by exogenous noise, but their generation is conditioned on  $z_{\text{inv}}$ . Endogenous Variables: These are the variables whose values are generated within the model. This includes the dependent latent variables and the final observed data:  $z_{\text{dp}}^{(\text{img})}$ : The image-dependent partition, generated based on  $z_{\text{inv}}$ .  $z_{\text{dp}}^{(\text{tex})}$ : The text-dependent partition, generated based on  $z_{\text{inv}}$ .  $x^{(\text{img})}$ : The generated image.  $x^{(\text{tex})}$ : The generated text. Structural Equations: The assumption provides the structural equations for the final observed variables,  $x^{(\text{img})}$  and  $x^{(\text{tex})}$ :

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

There are also implicit structural equations for the dependent partitions:

$$z_{\text{dp}}^{(\text{img})} \sim p_{z_{\text{dp}}^{(\text{img})}}(\cdot | z_{\text{inv}}) \quad z_{\text{dp}}^{(\text{tex})} \sim p_{z_{\text{dp}}^{(\text{tex})}}(\cdot | z_{\text{inv}}) \quad (19)$$

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

By laying out the SCM in this manner, we can clearly see the causal dependencies. The modal-invariant feature  $z_{\text{inv}}$  is a common cause of both the image and the text, which is what creates the "mutual semantics" between them. The private features,  $z_{\text{pr}}^{(\text{img})}$  and  $z_{\text{pr}}^{(\text{tex})}$ , account for the variability within each modality that is independent of the other. The dependent partitions,  $z_{\text{dp}}^{(\text{img})}$  and  $z_{\text{dp}}^{(\text{tex})}$ , 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  $z_{\text{inv}}$ .

**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); Seigal & Shen (2021); Lachapelle et al. (2021); Monti et al. (2019); Kivva et al. (2022); Squires et al. (2023); Buchholz et al. (2023); Gresele et al. (2021); Ahuja et al. (2022); Varshney (2017); Leeb 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); von Kügelgen et al. (2021). 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., Schwettmann et al. (2023b); Burns et al. (2023); Chen et al. (2020); Meng et al. (2022); Olah et al. (2020); Ravichander et al. (2020); Kim et al. (2018); Schwettmann et al. (2023a); Park et al. (2021); Squires et al. (2023).

## B APPENDIX.B

In this section, we provide the proofs to our main theoretical results in this paper.



## B.0.1 PROOF OF THEOREM.2

The proof sketch of our Theorem.2 can be derived into three parts. In the first part, we show how to construct the optimal  $f^*, g^*$  to fulfill 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  $z_{dp}^{(img)}, z_{pr}^{(img)}$  and the text-specific partitions  $z_{dp}^{(tex)}, z_{pr}^{(tex)}$ , thus, they only recover the modal-invariant partitions of the inverses  $f^{-1}, g^{-1}$ . 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  $\mathcal{L}_{MMAAlign}^{(img, tex)}$  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  $(0, 1)^{n_{inv}}$  is uniform distribution without extra moment constraint. To this, we show how to construct a pair of  $f, g$  that map  $x^{(img)}, x^{(tex)}$  into  $(0, 1)^{n_{inv}}$ , simultaneously attain the global minimization of  $\mathcal{L}_{MMAAlign}^{(img, tex)}$ . They would further lead to the construction of  $h_f, h_g$ .

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

$$d_i(z^{(inv)}) = F_i(z_i^{(inv)} | z_{1:i-1}^{(inv)}), i = 1, \dots, n_{inv} \quad (20)$$

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

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

Given this, we consider the following derivation:

$$\begin{aligned} \mathcal{L}_{MMAAlign}^{(img, tex)}(f^*, g^*) &= \mathbb{E}_{\langle x^{(img)}, x^{(tex)} \rangle \sim p_{mm}} \left[ \left\| f^*(x^{(img)}) - g^*(x^{(tex)}) \right\|_2^2 \right] - H(f^*(x^{(img)})) - H(g^*(x^{(tex)})) \\ &= \mathbb{E}_{\langle x^{(img)}, x^{(tex)} \rangle \sim p_{mm}} \left[ \left\| d(z^{(inv)}) - d(z^{(inv)}) \right\|_2^2 \right] - H(d(z^{(inv)})) - H(d(z^{(inv)})) \\ &= 0. \end{aligned} \quad (21)$$

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

$$\mathcal{L}_{MMAAlign}^{(img, tex)}(f^*, g^*) = \mathbb{E}_{\langle x^{(img)}, x^{(tex)} \rangle \sim p_{mm}} \left[ \left\| f^*(x^{(img)}) - g^*(x^{(tex)}) \right\|_2^2 \right] - H(f^*(x^{(img)})) - H(g^*(x^{(tex)})) \quad (22)$$

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

$$\begin{aligned} \mathbb{E}_{p_{mm}} \left[ \left\| h_f(z^{(img)}) - h_g(z^{(tex)}) \right\|_2^2 \right] &= 0 \\ \iff \mathbb{E}_{p_{mm}} \left[ \left\| h_f(z_{inv}, z_{dp}^{(img)}, z_{pr}^{(img)}) - h_g(z_{inv}, z_{dp}^{(tex)}, z_{pr}^{(tex)}) \right\|_2^2 \right] &= 0, \\ H(h_f(z^{(img)})) &= 0, \\ H(h_g(z^{(tex)})) &= 0. \end{aligned} \quad (23)$$

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.1.

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

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

$$h_f(z_{\text{inv}}, z_{\text{sp}}^{(\text{img})}) = h_f(z_{\text{inv}}, z_{\text{dp}}^{(\text{img})}, z_{\text{pr}}^{(\text{img})}); h_g(z_{\text{inv}}, z_{\text{sp}}^{(\text{tex})}) = h_g(z_{\text{inv}}, z_{\text{dp}}^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}). \quad (24)$$

In this way, we only need to prove

$$\begin{aligned} \frac{\partial h_f(\cdot|z_{\text{inv}})}{\partial z_{\text{sp},i}^{(\text{img})}} &= 0, \text{ s.t. } \forall i \in \{1, \dots, n_{\text{img}(\text{dp})} + n_{\text{img}(\text{pr})}\}; \\ \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{sp},j}^{(\text{tex})}} &= 0, \text{ s.t. } \forall j \in \{1, \dots, n_{\text{tex}(\text{dp})} + n_{\text{tex}(\text{pr})}\}, \end{aligned} \quad (25)$$

then the statements would be satisfied.

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

$$\exists l \in \{1, \dots, n_{\text{img}(\text{dp})} + n_{\text{img}(\text{pr})}\}, (\bar{z}_{\text{inv}}, \bar{z}_{\text{sp}}^{\text{img}}) \sim p_{z_{\text{sp}}^{(\text{img})}} \text{ s.t. } \frac{\partial h_f(\bar{z}_{\text{inv}}, \bar{z}_{\text{sp}}^{\text{img}})}{\partial z_{\text{sp},l}^{\text{img}}} \neq 0, \quad (26)$$

thus, we assume that the partial derivative of  $h_f$  with respect to the  $l^{\text{th}}$  image-private latent variable is non-zero at some point in the support of  $p_{z_{\text{sp}}^{(\text{img})}}$ , i.e.,  $\mathcal{Z}_{\text{img}} = \mathcal{C} \times \mathcal{Z}_{\text{img}^{\text{sp}}}$  ( $\mathcal{C}$  and  $\mathcal{Z}_{\text{img}^{\text{sp}}}$  are the subspaces that represent the supports of  $z_{\text{inv}}$  and  $z_{\text{sp}}^{(\text{img})}$ ). 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  $\frac{\partial h_f}{\partial z_{\text{sp},l}^{\text{img}}} \neq 0$  in a neighborhood of  $(\bar{z}_{\text{inv}}, \bar{z}_{\text{sp}}^{\text{img}})$ :

$$\exists \gamma > 0, \text{ s.t. } z'_l \mapsto h_f(\bar{z}_{\text{inv}}, (\bar{z}_{\text{sp}}^{\text{img}})_{-l}, z'_l) \text{ is strictly monotonic on } (z'_l - \gamma, z'_l + \gamma) \quad (27)$$

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

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

Under the condition in Eq.27, we separately consider two cases.

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

$$\mathbb{E}_{p_{\text{mm}}} \left[ \left\| h_f(\bar{z}_{\text{inv}}, \bar{z}_{\text{sp}}^{(\text{img})}) - h_g(\bar{z}_{\text{inv}}, \bar{z}_{\text{sp}}^{(\text{tex})}) \right\|_2^2 \right] = 0 \iff \mathbb{E}_{p_{\text{mm}}} \left[ \left\| h_f(\bar{z}_{\text{inv}}, \bar{z}_{\text{sp}}^{(\text{img})}) - \mathbf{v}_{z_{\text{inv}}} \right\|_2^2 \right] = 0 \quad (28)$$

Given this, the optima of  $h_f$  with respect to  $z_{sp}^{(img)}$  in the range of Eq.27 refers to a constant function, therefore

$$\forall \hat{z}_l \in (z'_l - \gamma, z'_l + \gamma), \quad h_f(\bar{z}_{inv}, (\bar{z}_{sp, -l}^{(img)}, \hat{z}_l)) = \mathbf{v}_{\bar{z}_{inv}}, \quad (29)$$

so that  $\frac{\partial h_f(\bar{z}_{inv}, (\bar{z}_{sp, -l}^{(img)}, \hat{z}_l))}{\partial z_{dp, l}^{img}} = 0$  on  $(z'_l - \gamma, z'_l + \gamma)$ . It violates the condition in Eq.27.

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

$$\Omega(z_{inv}, z_1, z_2) = \left| h_f(z_{inv}, z_1) - h_g(z_{inv}, z_2) \right| \geq 0. \quad (30)$$

Our goal is to show that  $\Omega$  is strictly positive with probability greater than zero with respect to  $p_{\text{mm}}$ .

Specifically, given  $\gamma$  defined from Eq.27, we may define  $\eta(\gamma) > 0$ , such that given  $z'_l \in (z'_l - \gamma, z'_l)$ , it holds

$$z'_l \mapsto h_g(\bar{z}_{inv}, (\bar{z}_{sp, -\hat{l}}^{\text{tex}}, z'_l)) \text{ is strictly monotonic on } (z'_l - \eta(\gamma), z'_l + \eta(\gamma)), \quad (31)$$

which is achieved due to the continuity of the first partial derivative of  $h_g$  w.r.t.  $z_{sp, \hat{l}}^{\text{tex}}$ . To achieve the

strict positivity of Eq.30, we are going to prove  $\frac{\partial h_f(\bar{z}_{inv}, z_{sp}^{\text{img}})}{\partial z_{sp, l}^{\text{img}}} - \frac{\partial h_g(\bar{z}_{inv}, z_{sp}^{\text{tex}})}{\partial z_{sp, \hat{l}}^{\text{tex}}} \neq 0$  in an open subset  $\mathcal{Z}' \subset \bar{z}_{inv} \times \left( \mathcal{Z}_{-l}^{\text{img}^{\text{sp}}} \times (z'_l - \gamma, z'_l) \right) \times \left( \mathcal{Z}_{-\hat{l}}^{\text{tex}^{\text{sp}}} \times (z'_l - \eta(\gamma), z'_l + \eta(\gamma)) \right)$  where  $\mathcal{Z}_{-l}^{\text{img}^{\text{sp}}}$  and  $\mathcal{Z}_{-\hat{l}}^{\text{tex}^{\text{sp}}}$  denote the subspaces of  $\mathcal{Z}_{\text{img}^{\text{sp}}}$  and  $\mathcal{Z}_{\text{tex}^{\text{sp}}}$  except for the  $l^{\text{th}}$  dimension and the  $\hat{l}^{\text{th}}$  dimension, respectively.

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

On the other hand, suppose that  $\bar{z}_{inv} \times (\hat{\mathbf{z}}_{-l} \times \hat{z}_l) \times (\hat{\mathbf{z}}_{-\hat{l}} \times \hat{z}_{\hat{l}}) \in \bar{z}_{inv} \times \left( \mathcal{Z}_{-l}^{\text{img}^{\text{sp}}} \times (z'_l - \gamma, z'_l) \right) \times \left( \mathcal{Z}_{-\hat{l}}^{\text{tex}^{\text{sp}}} \times (z'_l - \eta(\gamma), z'_l + \eta(\gamma)) \right)$  is a solution of  $\frac{\partial h_f(\bar{z}_{inv}, z_{sp}^{\text{img}})}{\partial z_{sp, l}^{\text{img}}} - \frac{\partial h_g(\bar{z}_{inv}, z_{sp}^{\text{tex}})}{\partial z_{sp, \hat{l}}^{\text{tex}}} = 0$ . Given this, let consider the ranges  $(z'_l - \gamma, \hat{z}_l)$  and  $(\hat{z}_l - \gamma, z'_l)$  to the  $l^{\text{th}}$  dimension of  $\mathcal{Z}_{\text{img}^{\text{dp}}}$ . According to the monotonicity of  $h_f$  and the continuity of  $\frac{\partial h_f}{\partial z_{sp, l}^{\text{img}}}$  with respect to  $z_{sp, l}^{\text{img}}$ , there is  $\mathcal{Z}_l^{(1)}(\gamma) \in \left\{ (z'_l - \gamma, \hat{z}_l), (\hat{z}_l - \gamma, z'_l) \right\}$  so that given  $z_l^{(1)} \in \mathcal{Z}_l^{(1)}(\gamma)$ , it holds:

$$\frac{\partial h_f(\bar{z}_{inv} \times (\hat{\mathbf{z}}_{-l} \times z_l^{(1)}) \times (\hat{\mathbf{z}}_{-\hat{l}} \times \hat{z}_{\hat{l}}))}{\partial z_{sp, l}^{\text{img}}} - \frac{\partial h_f(\bar{z}_{inv} \times (\hat{\mathbf{z}}_{-l} \times \hat{z}_l) \times (\hat{\mathbf{z}}_{-\hat{l}} \times \hat{z}_{\hat{l}}))}{\partial z_{sp, l}^{\text{img}}} > 0; \quad (32)$$

Similarly, according to the monotonicity of  $h_g$  and the continuity of  $\frac{\partial h_g}{\partial z_{sp,l}^{\text{tex}}}$  with respect to  $z_{sp,l}^{\text{tex}}$ , there is  $\mathcal{Z}_i^{(2)}(\eta(\gamma)) \in \left\{ (z'_i - \eta(\gamma), \hat{z}_i), (\hat{z}_i, z'_i + \eta(\gamma)) \right\}$  so that given  $z_i^{(2)} \in \mathcal{Z}_i^{(2)}(\eta(\gamma))$ , it holds:

$$\frac{\partial h_g(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times \hat{z}_l) \times (\hat{z}_{-i} \times \hat{z}_i))}{\partial z_{sp,l}^{\text{tex}}} - \frac{\partial h_g(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times \hat{z}_l) \times (\hat{z}_{-i} \times z_i^{(2)}))}{\partial z_{sp,l}^{\text{tex}}} > 0. \quad (33)$$

Combine Eq.32 and Eq.33, then we obtain

$$\begin{aligned} & \frac{\partial h_f(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times z_l^{(1)}) \times (\hat{z}_{-i} \times \hat{z}_i))}{\partial z_{sp,l}^{\text{img}}} - \frac{\partial h_f(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times \hat{z}_l) \times (\hat{z}_{-i} \times \hat{z}_i))}{\partial z_{sp,l}^{\text{img}}} \\ & + \frac{\partial h_g(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times \hat{z}_l) \times (\hat{z}_{-i} \times \hat{z}_i))}{\partial z_{sp,l}^{\text{tex}}} - \frac{\partial h_g(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times \hat{z}_l) \times (\hat{z}_{-i} \times z_i^{(2)}))}{\partial z_{sp,l}^{\text{tex}}} > 0 \\ \Leftrightarrow & \frac{\partial h_f(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times z_l^{(1)}) \times (\hat{z}_{-i} \times \hat{z}_i))}{\partial z_{sp,l}^{\text{img}}} - \frac{\partial h_g(\bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times \hat{z}_l) \times (\hat{z}_{-i} \times z_i^{(2)}))}{\partial z_{sp,l}^{\text{tex}}} > 0 \\ & \left( \bar{z}_{\text{inv}} \times (\hat{z}_{-l} \times \hat{z}_l) \times (\hat{z}_{-i} \times \hat{z}_i) \text{ is a solution of } \frac{\partial h_f(\bar{z}_{\text{inv}}, z_{\text{dp}}^{\text{img}})}{\partial z_{sp,l}^{\text{img}}} - \frac{\partial h_g(\bar{z}_{\text{inv}}, z_{\text{sp}}^{\text{tex}})}{\partial z_{sp,l}^{\text{tex}}} = 0 \right). \end{aligned} \quad (34)$$

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

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

$$\begin{aligned} \mathcal{Z}^{(1)}(\gamma, \eta) &:= \bar{z}_{\text{inv}} \times \left( \hat{z}_{-l} \times \left( \inf_{\mathcal{Z}_l^{\text{img}^{\text{sp}}}} \mathcal{Z}_l^{(1)}(\gamma), z_l^{(3)} \right) \right) \times \left( \hat{z}_{-i} \times \left( \inf_{\mathcal{Z}_i^{\text{tex}^{\text{sp}}}} \mathcal{Z}_i^{(2)}(\eta(\gamma)), z_i^{(4)} \right) \right); \\ \mathcal{Z}^{(2)}(\gamma, \eta) &:= \bar{z}_{\text{inv}} \times \left( \hat{z}_{-l} \times (z_l^{(3)}, \sup_{\mathcal{Z}_l^{\text{img}^{\text{sp}}}} \mathcal{Z}_l^{(1)}(\gamma)) \right) \times \left( \hat{z}_{-i} \times \left( \inf_{\mathcal{Z}_i^{\text{tex}^{\text{sp}}}} \mathcal{Z}_i^{(2)}(\eta(\gamma)), z_i^{(4)} \right) \right); \\ \mathcal{Z}^{(3)}(\gamma, \eta) &:= \bar{z}_{\text{inv}} \times \left( \hat{z}_{-l} \times (z_l^{(3)}, \sup_{\mathcal{Z}_l^{\text{img}^{\text{sp}}}} \mathcal{Z}_l^{(1)}(\gamma)) \right) \times \left( \hat{z}_{-i} \times (z_i^{(4)}, \sup_{\mathcal{Z}_i^{\text{tex}^{\text{sp}}}} \mathcal{Z}_i^{(2)}(\eta(\gamma))) \right); \\ \mathcal{Z}^{(4)}(\gamma, \eta) &:= \bar{z}_{\text{inv}} \times \left( \hat{z}_{-l} \times \left( \inf_{\mathcal{Z}_l^{\text{img}^{\text{sp}}}} \mathcal{Z}_l^{(1)}(\gamma), z_l^{(3)} \right) \right) \times \left( \hat{z}_{-i} \times (z_i^{(4)}, \sup_{\mathcal{Z}_i^{\text{tex}^{\text{sp}}}} \mathcal{Z}_i^{(2)}(\eta(\gamma))) \right). \end{aligned} \quad (35)$$

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

Therefore  $\mathcal{Z}''(\gamma, \eta)$  satisfy some conditions: 1). non-empty; 2). it is an open subset in the topological subspace of  $\mathcal{C} \times \mathcal{Z}_{\text{img}^{\text{sp}}} \times \mathcal{Z}_{\text{tex}^{\text{sp}}}$ ; 3).  $\forall (z_{\text{inv}}, z_1, z_2) \in \mathcal{Z}''(\gamma, \eta)$ , it holds  $\Omega(z_{\text{inv}}, z_1, z_2) = \left| h_f((z_{\text{inv}}, z_1) - h_g(z_{\text{inv}}, z_2)) \right| > 0$ ; 4). it is fully supported with respect to  $p_{\text{mm}}$  generated by Assumption.1. As a consequence,

$$p_{\text{mm}}(\Omega(z_{\text{inv}}, z_1, z_2) > 0) > p_{\text{mm}}(\mathcal{Z}''(\gamma, \eta)) > 0. \quad (36)$$

So it leads to

$$\begin{aligned} & \mathbb{E}_{p_{\text{mm}}} \left[ \left\| h_f \left( (z_{\text{inv}}, z_{\text{dp}}^{(\text{img})}, z_{\text{pr}}^{(\text{img})}) \right) - h_g \left( (z_{\text{inv}}, z_{\text{dp}}^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}) \right) \right\|_2^2 \right] \\ & \geq \mathbb{E}_{p_{\text{mm}}(\mathcal{Z}''(\gamma, \eta))} \left[ \left\| \Omega(z_{\text{inv}}, z_{\text{sp}}^{(\text{img})}, z_{\text{sp}}^{(\text{tex})}) \right\|_2^2 \right] \\ & > 0. \end{aligned} \quad (37)$$

It results in the contradiction to Eq.23 .

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

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

**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 10.** (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  $p\#h(z)$  of  $p$  through  $h$  is also a regular density, i.e.,  $0 < p\#h(z) < \infty$ , 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  $\mathcal{N}=(0, 1)^{n_{\text{inv}}}$ . 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  $z_{\text{inv}}$  into a uniform random variable. Notice that  $p_{z_{\text{inv}}}$  (Assumption.1) and the uniform distribution (the pushforward of  $p_{z_{\text{inv}}}$ ) are regular densities in the sense of Lemma.10, therefore  $h_f$  and  $h_g$  are bijective maps, i.e., invertable.

### B.1 PROOF OF COROLLARY.3

Here we derive the formal version of Corollary.3 with its proof:

**Proposition 11.** *For fixed  $\gamma > 0$ , as the number of negative samples  $K - 1 \rightarrow \infty$ , the (normalized)  $\mathcal{L}_{\text{InfoNCE}}^{(\text{img} \rightarrow \text{tex})} - \log(K - 1)$  and  $\mathcal{L}_{\text{InfoNCE}}^{(\text{tex} \rightarrow \text{img})} - \log(K - 1)$  converge to*

$$\begin{aligned} & -\frac{1}{\gamma} \mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left( f(x^{(\text{img})})^\top g(x^{(\text{tex})}) \right) \\ & + \mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left( \log \mathbb{E}_{\hat{x}^{(\text{tex})} \sim p(x^{(\text{tex})})} \left[ e^{f(x^{(\text{img})})^\top g(\hat{x}^{(\text{tex})})/\gamma} \right] \right); \end{aligned} \quad (38)$$

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

respectively, with the following results:

1. The first terms of Eq.38 and 39, are minimized iff  $f, g$  are perfectly aligned, i.e.,  $\|f(x^{(\text{img})}) - g(x^{(\text{tex})})\|_2^2 \rightarrow 0$  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.38 and 39 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  $\mathcal{L}_{\text{InfoNCE}}^{(\text{img} \rightarrow \text{tex})}$ , 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  $\mathcal{L}_{\text{InfoNCE}}^{(\text{tex} \rightarrow \text{img})}$ . Note that  $x^{(\text{img})} \stackrel{\text{i.i.d.}}{\sim} p_{\text{mm}}(x^{(\text{img})})$ ;  $\{\hat{x}_k^{(\text{img})}\}_{k=1}^{K-1} \stackrel{\text{i.i.d.}}{\sim} p_{\text{mm}}(x^{(\text{img})})$ , therefore we have

$$\lim_{K-1 \rightarrow \infty} \log \left( \frac{e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma}}{K-1} + \sum_{k=1}^{K-1} \frac{e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma}}{K-1} \right) = \log \mathbb{E}_{\hat{x}^{(\text{tex})} \sim p(x^{(\text{tex})})} \left[ e^{f(x^{(\text{img})})^\top g(\hat{x}^{(\text{tex})})/\gamma} \right] \quad (40)$$

with the strong law of large numbers and the Continuous Mapping Theorem. Eq.40 results in

$$\begin{aligned} & \mathcal{L}_{\text{InfoNCE}}^{(\text{img} \rightarrow \text{tex})} - \log(K-1) \\ &= \mathbb{E}_{\substack{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}} \\ \langle x^{(\text{img})}, \hat{x}_k^{(\text{tex})} \rangle \sim p_{\text{mm}}}} - \log \frac{(K-1)e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma}}{e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma} + \sum_{k=1}^{K-1} e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma}} \\ &= \mathbb{E}_{\substack{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}} \\ \langle x^{(\text{img})}, \hat{x}_k^{(\text{tex})} \rangle \sim p_{\text{mm}}}} \left( -\frac{f(x^{(\text{img})})^\top g(x^{(\text{tex})})}{\gamma} + \log \frac{(e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma} + \sum_{k=1}^{K-1} e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma})}{K-1} \right) \\ &= -\frac{1}{\gamma} \mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left( f(x^{(\text{img})})^\top g(x^{(\text{tex})}) \right) \\ & \quad + \mathbb{E}_{\substack{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}} \\ \langle x^{(\text{img})}, \hat{x}_k^{(\text{tex})} \rangle \sim p_{\text{mm}}}} \left( \log \frac{(e^{f(x^{(\text{img})})^\top g(x^{(\text{tex})})/\gamma} + \sum_{k=1}^{K-1} e^{f(x^{(\text{img})})^\top g(\hat{x}_k^{(\text{tex})})/\gamma})}{K-1} \right) \end{aligned} \quad (41)$$

So

$$\begin{aligned} & \lim_{K-1 \rightarrow \infty} \mathcal{L}_{\text{InfoNCE}}^{(\text{img} \rightarrow \text{tex})} - \log(K-1) \\ &= -\frac{1}{\gamma} \mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left( f(x^{(\text{img})})^\top g(x^{(\text{tex})}) \right) \\ & \quad + \mathbb{E}_{\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left( \log \mathbb{E}_{\hat{x}^{(\text{tex})} \sim p(x^{(\text{tex})})} \left[ e^{f(x^{(\text{img})})^\top g(\hat{x}^{(\text{tex})})/\gamma} \right] \right) \end{aligned} \quad (42)$$

Similarly we obtain

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

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.38 and Eq.39 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  $p_{\text{mm}}$ , it holds  $p_{\text{mm}} = p_{x^{(\text{img})}} \delta(f(x^{(\text{img})}) - g(x^{(\text{tex})})) = p_{x^{(\text{tex})}} \delta(f(x^{(\text{img})}) - g(x^{(\text{tex})}))$ .

$g(\mathbf{x}^{(\text{tex})})$ ). Therefore we have

$$\begin{aligned}
& \mathbb{E}_{\langle \mathbf{x}^{(\text{img})}, \mathbf{x}^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left( \log \mathbb{E}_{\hat{\mathbf{x}}^{(\text{tex})} \sim p(\mathbf{x}^{(\text{tex})})} \left[ e^{f(\mathbf{x}^{(\text{img})})^\top g(\hat{\mathbf{x}}^{(\text{tex})})/\gamma} \right] \right) \\
&= \mathbb{E}_{\mathbf{x}^{(\text{tex})} \sim p(\mathbf{x}^{(\text{tex})})} \mathbb{E}_{\mathbf{x}^{(\text{img})} \sim \delta(f(\mathbf{x}^{(\text{img})}) - g(\mathbf{x}^{(\text{tex})}))} \left( \log \mathbb{E}_{\hat{\mathbf{x}}^{(\text{tex})} \sim p(\mathbf{x}^{(\text{tex})})} \left[ e^{f(\mathbf{x}^{(\text{img})})^\top g(\hat{\mathbf{x}}^{(\text{tex})})/\gamma} \right] \right) \\
&= \mathbb{E}_{\mathbf{x}^{(\text{tex})} \sim p(\mathbf{x}^{(\text{tex})})} \left( \log \mathbb{E}_{\hat{\mathbf{x}}^{(\text{tex})} \sim p(\mathbf{x}^{(\text{tex})})} \left[ e^{g(\mathbf{x}^{(\text{tex})})^\top g(\hat{\mathbf{x}}^{(\text{tex})})/\gamma} \right] \right) \\
&\triangleq \frac{1}{N} \sum_{i=1}^N \log \left( \frac{1}{N} \sum_{i=1}^N \left[ e^{g(\mathbf{x}_i^{(\text{tex})})^\top g(\mathbf{x}_j^{(\text{tex})})/\gamma} \right] \right) \\
&= \frac{1}{N} \sum_{i=1}^N \log \hat{p}_{\text{vMF-KDE}}(g(\mathbf{x}_i^{(\text{tex})})) + \log Z_{\text{vMF}} \\
&\triangleq -H(g(\mathbf{x}^{(\text{tex})})) + \log Z_{\text{vMF}}
\end{aligned} \tag{44}$$

where  $H(f(\mathbf{x}^{(\text{tex})}))$  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  $\kappa = \gamma^{-1}$ ;  $Z_{\text{vMF}}$  denotes the normalization constant for vMF distribution with  $\kappa = \gamma^{-1}$ . Using the same proof technique, we also obtain

$$\mathbb{E}_{\langle \mathbf{x}^{(\text{img})}, \mathbf{x}^{(\text{tex})} \rangle \sim p_{\text{mm}}} \left( \log \mathbb{E}_{\hat{\mathbf{x}}^{(\text{img})} \sim p(\mathbf{x}^{(\text{img})})} \left[ e^{f(\hat{\mathbf{x}}^{(\text{img})})^\top g(\mathbf{x}^{(\text{tex})})/\gamma} \right] \right) \triangleq -H(f(\mathbf{x}^{(\text{img})})) + \log Z_{\text{vMF}} \tag{45}$$

The proposition has been proven.  $\square$

#### B.1.1 PROOF OF THEOREM.4

Similar with the proof of Theorem.1, our proof of Theorem.4 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  $\mathbf{f}^{-1}, \mathbf{g}_i^{-1}, \forall i \in \{1, \dots, k_{\text{max}}\}$ ); 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.4 is consistent with Assumption.1, it leads to the same construction process of  $f^*$  and  $h_f$  in the proof of Theorem.2 .

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

$$\begin{aligned}
& \forall (\mathbf{z}_{\text{inv}}, \mathbf{z}_1^{(\text{tex})}, \dots, \mathbf{z}_k^{(\text{tex})}, \mathbf{z}_{\text{pr}}^{(\text{tex})}) \in \mathcal{C}_{\text{inv}} \times (\mathcal{S}_1 \times \dots \times \mathcal{S}_k) \times \mathcal{S}_{\text{pr}}^{\text{tex}} \\
& \mathbf{g}^{(k)}(\mathbf{z}_{\text{inv}}, \mathbf{z}_1^{(\text{tex})}, \dots, \mathbf{z}_k^{(\text{tex})}, \mathbf{z}_{\text{pr}}^{(\text{tex})}) = [\mathbf{g}_1(\mathbf{z}_{\text{inv}}, \mathbf{z}_1^{(\text{tex})}, \mathbf{z}_{\text{pr}}^{(\text{tex})}), \dots, \mathbf{g}_k(\mathbf{z}_{\text{inv}}, \{\mathbf{z}_j^{(\text{tex})}\}_{j=1}^k, \mathbf{z}_{\text{pr}}^{(\text{tex})})]
\end{aligned} \tag{46}$$



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

$$(\mathbf{g}^{(k)})^{-1} \left( [\mathbf{g}_1(z_{\text{inv}}, z_1^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}), \dots, \mathbf{g}_k(z_{\text{inv}}, \{z_j^{(\text{tex})}\}_{j=1}^k, z_{\text{pr}}^{(\text{tex})})] \right) = (z_{\text{inv}}, z_1^{(\text{tex})}, \dots, z_k^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}). \quad (47)$$

Hence for each  $X^{\text{tex}}$  generated by Assumption.4, *i.e.*,  $X^{\text{tex}} = [\mathbf{g}_1(z_{\text{inv}}, z_1^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}), \dots, \mathbf{g}_k(z_{\text{inv}}, \{z_j^{(\text{tex})}\}_{j=1}^k, z_{\text{pr}}^{(\text{tex})})] \ (\forall k \in \{1, \dots, k_{\text{max}}\})$ , we can restrict their outputs in the first  $n_{\text{inv}}$  dimensions such that  $(\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{\text{tex}}) = z^{(\text{inv})}$ . Then we employ the same Damois construction technique used in first step of the proof in Theorem.2 to define the function  $\mathbf{d} : \mathcal{C}_{\text{inv}} \rightarrow (0, 1)^{n_{\text{inv}}}$  that map  $z_{\text{inv}}$  into a uniform random variable. Derived from such construction,  $\mathbf{d}(z^{(\text{inv})})$  is uniformly distributed on  $(0, 1)^{n_{\text{inv}}}$  Darnois (1951), and is also smooth due to the third premise.

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

$$g^*(X^{\text{tex}}) := \mathbf{d} \circ (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{\text{tex}}) \text{ if } X^{\text{tex}} \in \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)}), \forall i \in \{1, \dots, k_{\text{max}}\}, \quad (48)$$

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

$$\begin{aligned} \mathcal{L}_{\text{MMAlign}}^{\text{(img, tex)}}(f^*, g^*) &= \mathbb{E}_{\substack{\langle x^{(\text{img})}, X^{\text{tex}} \rangle \\ \sim p_{\text{mm}}}} \left[ \left\| f^*(x^{(\text{img})}) - g^*(X^{\text{tex}}) \right\|_2^2 - H(f^*(x^{(\text{img})})) - H(g^*(X^{\text{tex}})) \right] \\ &= \sum_{k=1}^{k_{\text{max}}} p(\dim_{\text{col}}(X^{\text{tex}}) = k) \mathbb{E}_{\substack{\langle x^{(\text{img})}, X^{\text{tex}} \rangle \\ \sim p_{\text{mm}}(\mid \dim_{\text{col}}(X^{\text{tex}}) = k)}} \left[ \left\| f^*(x^{(\text{img})}) - g^*(X^{\text{tex}}) \right\|_2^2 \right. \\ &\quad \left. - H(f^*(x^{(\text{img})})) - H(g^*(X^{\text{tex}})) \right] \\ &= \sum_{k=1}^{k_{\text{max}}} p(\dim_{\text{col}}(X^{\text{tex}}) = k) \mathbb{E}_{\substack{\langle x^{(\text{img})}, X^{\text{tex}} \rangle \\ \sim p_{\text{mm}}(\mid \dim_{\text{col}}(X^{\text{tex}}) = k)}} \left[ \left\| f^*(x^{(\text{img})}) - \mathbf{d} \circ (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{\text{tex}}) \right\|_2^2 \right. \\ &\quad \left. - H(f^*(x^{(\text{img})})) - H(g^*(X^{\text{tex}})) \right] \\ &= \sum_{k=1}^{k_{\text{max}}} p(\dim_{\text{col}}(X^{\text{tex}}) = k) \mathbb{E}_{\substack{\langle x^{(\text{img})}, X^{\text{tex}} \rangle \\ \sim p_{\text{mm}}(\mid \dim_{\text{col}}(X^{\text{tex}}) = k)}} \left[ \left\| \mathbf{d}(z_{\text{inv}}) - \mathbf{d}(z_{\text{inv}}) \right\|_2^2 \right. \\ &\quad \left. - H(\mathbf{d}(z_{\text{inv}})) - H(\mathbf{d}(z_{\text{inv}})) \right] \\ &= 0, \end{aligned} \quad (49)$$

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

$$\mathcal{L}_{\text{MMAlign}}^{\text{(img, tex)}}(f^*, g^*) = \mathbb{E}_{\substack{\langle x^{(\text{img})}, X^{\text{tex}} \rangle \\ \sim p_{\text{mm}}}} \left[ \left\| f^*(x^{(\text{img})}) - g^*(X^{\text{tex}}) \right\|_2^2 - H(f^*(x^{(\text{img})})) - H(g^*(X^{\text{tex}})) \right], \quad (50)$$

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

$$h_g(z) = g^* \circ \mathbf{g}^{(k)}(z), \text{ if } z \in \mathcal{C}_{\text{inv}} \times (\mathcal{S}_1 \times \dots \times \mathcal{S}_k) \times \mathcal{S}_{\text{pr}}, \forall k \in \{1, \dots, k_{\text{max}}\}. \quad (51)$$

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

$$\begin{aligned} &\mathbb{E}_{p_{\text{mm}}} \left[ \left\| h_f(z^{(\text{img})}) - h_g(z^{(\text{tex})}) \right\|_2^2 \right] = 0 \\ &= \sum_{k=1}^{k_{\text{max}}} p(\dim_{\text{col}}(X^{\text{tex}}) = k) \mathbb{E}_{\substack{\langle x^{(\text{img})}, X^{\text{tex}} \rangle \\ \sim p_{\text{mm}}(\mid \dim_{\text{col}}(X^{\text{tex}}) = k)}} \left[ \left\| h_f(z_{\text{inv}}, z_{\text{dp}}^{(\text{img})}, z_{\text{pr}}^{(\text{img})}) - h_g(z_{\text{inv}}, \{z_j^{(\text{tex})}\}_{j=1}^k, z_{\text{pr}}^{(\text{tex})}) \right\|_2^2 \right] = 0, \\ &H(h_f(z^{(\text{img})})) = 0, \quad H(h_g(X^{\text{tex}})) = 0. \end{aligned} \quad (52)$$

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.4.

**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.2, it satisfies that given  $z_{\text{inv}} \sim p_{z_{\text{inv}}}$ , for all  $i \in \{1, \dots, n_{\text{img}(\text{dp})}\}$  and  $j \in \{1, \dots, n_{\text{img}(\text{pr})}\}$ , it results in  $\frac{\partial h_f(\cdot|z_{\text{inv}})}{\partial z_{\text{dp},i}^{(\text{img})}}=0$ ,  $\frac{\partial h_f(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{img})}}=0$ ; for all  $i \in \{1, \dots, n_{\text{tex}(\text{dp})}\}$  and  $j \in \{1, \dots, n_{\text{tex}(\text{pr})}\}$ , it results in  $\frac{\partial h_f(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}}=0$ ,  $\frac{\partial h_f(\cdot|z_{\text{inv}})}{\partial z_{\text{dp},i}^{(\text{tex})}}=0$ . They are consistent with the proof of Theorem.1.

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

When  $k = 1$ , it can be reduced to prove  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{dp},j}^{(\text{tex})}}=0$  and  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}}=0$  in the proof of Theorem.2, so it is satisfied. Regarding this as the first step, we construct a mathematical induction procedure to prove the rest.

Specifically, suppose that  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k,l}^{(\text{tex})}}=0$  ( $\forall l \in \{1, \dots, n_k^{(\text{tex})}\}$ ) and  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}}=0$  ( $\forall j \in \{1, \dots, n_{\text{pr}}^{(\text{tex})}\}$ ) with

$$h_g(\cdot) = g^* \circ \mathbf{g}^{(k)}(\cdot) = g^* \circ ([\mathbf{g}_1(\cdot), \mathbf{g}_2(\cdot), \dots, \mathbf{g}_k(\cdot)]), \quad (53)$$

thus,

$$\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}} = \frac{\partial (g^* \circ ([\mathbf{g}_1(z_{\text{inv}}, \cdot), \dots, \mathbf{g}_k(z_{\text{inv}}, \cdot)])}{\partial z_{\text{pr},j}^{(\text{tex})}} = \sum_{i=1}^k \sum_{i'=1}^m \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)}{\partial z_{\text{pr},j}^{(\text{tex})}} = 0 \quad (54)$$

and

$$\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k,l}^{(\text{tex})}} = \frac{\partial (g^* \circ ([\mathbf{g}_1(z_{\text{inv}}, \cdot), \dots, \mathbf{g}_k(z_{\text{inv}}, \cdot)])}{\partial z_{k,l}^{(\text{tex})}} = \sum_{i=1}^k \sum_{i'=1}^m \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)}{\partial z_{k,l}^{(\text{tex})}} = 0. \quad (55)$$

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  $\forall k' < k$ ,  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k',l'}^{(\text{tex})}}=0$  ( $\forall l' \in \{1, \dots, n_{k'}^{(\text{tex})}\}$ ).

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

$$\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k',l'}^{(\text{tex})}} = \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k-2,l'}^{(\text{tex})}} = \sum_{l=1}^{n_{k-1}^{(\text{tex})}} \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k-1,l}^{(\text{tex})}} \frac{\partial z_{k-1,l}^{(\text{tex})}}{\partial z_{k-2,l'}^{(\text{tex})}} + \sum_{l=1}^{n_k^{(\text{tex})}} \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k,l}^{(\text{tex})}} \frac{\partial z_{k,l}^{(\text{tex})}}{\partial z_{k-2,l'}^{(\text{tex})}} = 0. \quad (56)$$

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

$$\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k',l'}^{(\text{tex})}} = \sum_{t=1}^{k-k'} \sum_{l=1}^{n_{k-t+1}^{(\text{tex})}} \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k-t+1,l}^{(\text{tex})}} \frac{\partial z_{k-t+1,l}^{(\text{tex})}}{\partial z_{k',l'}^{(\text{tex})}} \quad (57)$$

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

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

$$\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}} = \sum_{i=1}^{k+1} \sum_{i'=1}^m \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)}{\partial z_{\text{pr},j}^{(\text{tex})}} = \sum_{i'=1}^m \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial \mathbf{g}_{k+1,i'}(z_{\text{inv}}, \cdot)} \frac{\partial \mathbf{g}_{k+1,i'}(z_{\text{inv}}, \cdot)}{\partial z_{\text{pr},j}^{(\text{tex})}} \quad (58)$$

and

$$\begin{aligned} \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k+1,l}^{(\text{tex})}} &= \sum_{i=1}^k \sum_{i'=1}^m \sum_{k'=1}^k \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)} \frac{\partial \mathbf{g}_{i,i'}(z_{\text{inv}}, \cdot)}{\partial z_{k',l}^{(\text{tex})}} \frac{\partial z_{k',l}^{(\text{tex})}}{\partial z_{k+1,l}^{(\text{tex})}} + \sum_{i'=1}^m \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial \mathbf{g}_{k+1,i'}(z_{\text{inv}}, \cdot)} \frac{\partial \mathbf{g}_{k+1,i'}(z_{\text{inv}}, \cdot)}{\partial z_{k+1,l}^{(\text{tex})}} \\ &= \sum_{i'=1}^m \frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial \mathbf{g}_{k+1,i'}(z_{\text{inv}}, \cdot)} \frac{\partial \mathbf{g}_{k+1,i'}(z_{\text{inv}}, \cdot)}{\partial z_{k+1,l}^{(\text{tex})}}, \end{aligned} \quad (59)$$

where only the  $(k + 1)$ -th token output  $\mathbf{g}_{k+1}(z_{\text{inv}}, \{z_i^{(\text{tex})}\}_{i=1}^{k+1}, z_{\text{pr}}^{(\text{tex})})$  influence the derivatives with respect to  $z_{k+1,l}^{(\text{tex})}$  and  $z_{k+1,l}^{(\text{tex})}$ . To this,  $\forall i \in \{1, \dots, k\}$ , suppose  $\bar{z}_i^{(\text{tex})} \sim p_{z_i^{(\text{tex})}}$  drawn through the generative process based on Assumption.4, given that  $\bar{Z}_k = \{\bar{z}_i^{(\text{tex})}\}_{i=1}^k$  is fixed, we consider the surrogate function family

$$\begin{aligned} h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k) &= g^* \circ \mathbf{g}^{(k+1)}(z_{\text{inv}}, \bar{z}_1^{(\text{tex})}, \dots, \bar{z}_k^{(\text{tex})}, z_{k+1}^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}) \\ &= g^* \circ \left( [\mathbf{g}_1(z_{\text{inv}}, \bar{z}_1^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}), \dots, \mathbf{g}_k(z_{\text{inv}}, \{\bar{z}_j^{(\text{tex})}\}_{j=1}^k, z_{\text{pr}}^{(\text{tex})}), \mathbf{g}_{k+1}(z_{\text{inv}}, \{\bar{z}_j^{(\text{tex})}\}_{j=1}^k, z_{k+1}^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})})] \right). \end{aligned} \quad (60)$$

Observe that  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}} = \frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{\text{pr},j}^{(\text{tex})}}$  and  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k+1,l}^{(\text{tex})}} = \frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{k+1,l}^{(\text{tex})}}$  when  $z_i^{(\text{tex})} = \bar{z}_i^{(\text{tex})}$  ( $\forall i \in \{1, \dots, k\}$ ).

Hence if we can prove  $\frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{\text{pr},j}^{(\text{tex})}} = 0$  and  $\frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{k+1,l}^{(\text{tex})}} = 0$  satisfied across the surrogate function family,  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}} = 0$  and  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k+1,l}^{(\text{tex})}} = 0$  can be proven.

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

Generation of  $x^{(\text{tex})}$  in Assumption.1 :

$$z_{\text{inv}} \sim p_{z_{\text{inv}}}, z_{\text{dp}}^{(\text{tex})} \sim p_{z_{\text{dp}}^{(\text{tex})}}(\cdot|z_{\text{inv}}), z_{\text{pr}}^{(\text{tex})} \sim p_{z_{\text{pr}}^{(\text{tex})}}, x^{(\text{tex})} = \mathbf{g}(z_{\text{inv}}, z_{\text{dp}}^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}); \quad (61)$$

Generation of  $X_{:,k+1}^{(\text{tex})}$  in Assumption.4 :

$$z_{\text{inv}} \sim p_{z_{\text{inv}}}, z_{k+1}^{(\text{tex})} \sim p_{z_{k+1}^{(\text{tex})}}(\cdot|z_{\text{inv}}, \bar{Z}_k), z_{\text{pr}}^{(\text{tex})} \sim p_{z_{\text{pr}}^{(\text{tex})}}, X_{:,k+1}^{(\text{tex})} = \mathbf{g}'_{k+1, \bar{Z}_k}(z_{\text{inv}}, z_{k+1}^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}).$$

Given this, if we reframe  $p_{z_{k+1}^{(\text{tex})}}(\cdot, \bar{Z}_k)$  and  $\mathbf{g}'_{k+1, \bar{Z}_k}(\cdot)$  as  $p_{z_{\text{dp}}^{(\text{tex})}}(\cdot)$  and  $\mathbf{g}(\cdot)$ , respectively, then the proof of  $\frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{\text{pr},j}^{(\text{tex})}} = 0$  and  $\frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{k+1,l}^{(\text{tex})}} = 0$  can be reduced to the proof of  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{dp},j}^{(\text{tex})}} = 0$  and  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k+1,l}^{(\text{tex})}} = 0$  in Theorem.2. It is satisfied and since the  $\bar{Z}_k$  can be a arbitrary combination draw from the generative process in Assumption.4,  $\frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{\text{pr},j}^{(\text{tex})}} = 0$  and  $\frac{\partial h'_g(\cdot|z_{\text{inv}}; \bar{Z}_k)}{\partial z_{k+1,l}^{(\text{tex})}} = 0$  are satisfied across the surrogate function family so that  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}} = 0$  and  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k+1,l}^{(\text{tex})}} = 0$  have been proved.

To this, we have  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{k',l}^{(\text{tex})}}=0$  ( $\forall l \in \{1, \dots, n_{k'}^{(\text{tex})}\}, \forall k' \in \{1, \dots, k+1\}$ ) and  $\frac{\partial h_g(\cdot|z_{\text{inv}})}{\partial z_{\text{pr},j}^{(\text{tex})}}=0$  ( $\forall j \in \{1, \dots, n_{\text{pr}}^{(\text{tex})}\}$ ).

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

**Invertability of  $h_f, h_g$ .** The procedure of proving the invertability of  $h_f$  is consistent with Theorem.2. As to the invertability of  $h_g$ , we consider its piecewise functions derived from  $\{\mathbf{g}^{(k)}\}_{k=1}^{k_{\text{max}}}$  that generate sentence matrices with different sizes of their columns, then  $\forall k \in \{1, \dots, k_{\text{max}}\}$ ,

$$\mathbf{g} = \mathbf{g}^{(k)} : \mathcal{C}_{\text{inv}} \times \left( \mathcal{S}_1 \times \dots \times \mathcal{S}_k \right) \times \mathcal{S}_{\text{pr}} \rightarrow \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)}), \forall k \in \{1, \dots, k_{\text{max}}\}, \quad (62)$$

and because of

$$g^* = \mathbf{d} \circ (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1} : \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)}) \rightarrow (0, 1)^{n_{\text{inv}}}, \forall k \in \{1, \dots, k_{\text{max}}\}, \quad (63)$$

which is smooth on the generative process of  $\mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)})$ , we apply our result to Lemma.10 by setting  $\mathcal{M}=\mathcal{C}_{\text{inv}}$  and  $\mathcal{N}=(0, 1)^{n_{\text{inv}}}$ . 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  $z_{\text{inv}} \in \mathcal{C}_{\text{inv}}$  into a uniform random variable in  $(0, 1)^{n_{\text{inv}}}$ . Notice that  $p_{z_{\text{inv}}}$  (Assumption.4) and the uniform distribution (the pushforward of  $p_{z_{\text{inv}}}$ ) are regular densities in the sense of Lemma.10, therefore  $h_g$  is a bijective map with respect to  $\forall k \in \{1, \dots, k_{\text{max}}\}$ , i.e., invertable.

## B.2 PROOF OF COROLLARY.6

The proof of Corollary.6 can be typically derived from the proof of Corollary.3.

## B.3 PROOF OF THEOREM.7

To prove the result, we only need to construct  $g^{**}$  based on the optimal text encoder  $g^*$  defined by Theorem.5 and take it to define  $g^{**}$ , then prove  $\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^{**}) = 0$ . Afterwards, we prove its invariance to the permutation of sentence-matrix columns given  $\pi(X^{(\text{tex})}) \in \Pi_k(\{1, \dots, k\})$  that satisfies

$$g^{**}([X_{:,1}^{(\text{tex})}, X_{:,2}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\pi(1)}^{(\text{tex})}, X_{:,\pi(2)}^{(\text{tex})}, \dots, X_{:,\pi(k)}^{(\text{tex})}]), \quad (64)$$

it holds  $\forall \hat{\pi}(X^{(\text{tex})}) \in \Pi_k(\{1, \dots, k\}) \cap \{\{X_{:,1}^{(\text{tex})}, X_{:,\pi(1)}^{(\text{tex})}\} \times \dots \times \{X_{:,k}^{(\text{tex})}, X_{:,\pi(k)}^{(\text{tex})}\}\}$ ,

$$g^{**}([X_{:,1}^{(\text{tex})}, X_{:,2}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\hat{\pi}(1)}^{(\text{tex})}, X_{:,\hat{\pi}(2)}^{(\text{tex})}, \dots, X_{:,\hat{\pi}(k)}^{(\text{tex})}]). \quad (65)$$

**Construction of  $g^{**}$ .** From the first-phase proof of Theorem.5, we have  $g^*$  as a piecewise function on differnt-length text inputs, which satisfies

$$g^* = \mathbf{d} \circ (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1} : \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)}) \rightarrow (0, 1)^{n_{\text{inv}}}, \forall k \in \{1, \dots, k_{\text{max}}\}, \quad (66)$$

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

$$\begin{aligned} \forall (z_{\text{inv}}, z_1^{(\text{tex})}, \dots, z_k^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}) \in \mathcal{C}_{\text{inv}} \times \left( \mathcal{S}_1 \times \dots \times \mathcal{S}_k \right) \times \mathcal{S}_{\text{pr}}^{\text{tex}} \\ \mathbf{g}^{(k)}(z_{\text{inv}}, z_1^{(\text{tex})}, \dots, z_k^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}) = \left[ \mathbf{g}_1(z_{\text{inv}}, z_1^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}), \dots, \mathbf{g}_k(z_{\text{inv}}, \{z_j^{(\text{tex})}\}_{j=1}^k, z_{\text{pr}}^{(\text{tex})}) \right]. \end{aligned} \quad (67)$$

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  $\hat{\mathbf{g}}^{(k)}: \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)}) \rightarrow \mathcal{C}_{\text{inv}} \times \left( \mathcal{S}_1 \times \dots \times \mathcal{S}_k \right) \times \mathcal{S}_{\text{pr}}^{\text{tex}}$  and  $\hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)} = (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}$ , leading to  $g^{**} = \mathbf{d} \circ (\hat{\mathbf{g}}^{(k)})_{1:n_{\text{inv}}} = \mathbf{d} \circ (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1} = g^*$ . Specifically,  $\hat{\mathbf{g}}^{(k)}$  can be constructed by

$$\hat{\mathbf{g}}^{(k)}(X^{(\text{tex})}) = \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1} (\mathcal{T}_{\text{tex}}^{(1)} \times \dots \times \mathcal{T}_{\text{tex}}^{(j-1)} \times \{X_{:,j}^{(\text{tex})}\} \times \mathcal{T}_{\text{tex}}^{(j+1)} \times \dots \times \mathcal{T}_{\text{tex}}^{(k)}) \quad (68)$$

where  $\mathcal{T}_{\text{tex}}^{(j)}$  indicates the  $j^{\text{th}}$  token embedding space with respect to all  $k$ -length sentence matrices lying on  $\mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)})$ , then  $\mathcal{T}_{\text{tex}}^{(1)} \times \dots \times \mathcal{T}_{\text{tex}}^{(j-1)} \times \{X_{:,j}^{(\text{tex})}\} \times \mathcal{T}_{\text{tex}}^{(j+1)} \dots \mathcal{T}_{\text{tex}}^{(k)}$  denotes the matrix set including all  $k$ -length sentence matrices whose  $j^{\text{th}}$  token embedding are  $X_{:,j}^{(\text{tex})}$ . 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.,  $(\mathbf{g}^{(k)})^{-1}(\mathcal{X}') = \{\hat{z} := (z_{\text{inv}}, z_1^{(\text{tex})}, \dots, z_k^{(\text{tex})}, z_{\text{pr}}^{(\text{tex})}) \in \mathcal{C}_{\text{inv}} \times (\mathcal{S}_1 \times \dots \times \mathcal{S}_k) \times \mathcal{S}_{\text{pr}}^{\text{tex}}, \text{s.t. } \mathbf{g}^{(k)}(\hat{z}) \in \mathcal{X}'\}$ . For simplicity, we denote

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

therefore

$$\hat{\mathbf{g}}^{(k)}(X^{(\text{tex})}) = \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})). \quad (69)$$

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

**Lemma 12.**  $\forall X^{(\text{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)})$ , then  $\forall z' \in \hat{\mathbf{g}}^{(k)}(X^{(\text{tex})})$ ,  $z'_{1:n_{\text{inv}}} = z_{\text{inv}} = (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{(\text{tex})})$ .

*Proof.* The proof is achieved by two steps.

In the first step, we prove that there exists  $z' \in \hat{\mathbf{g}}^{(k)}(X^{(\text{tex})})$ ,  $z'_{1:n_{\text{inv}}} = z_{\text{inv}} = (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{(\text{tex})})$ . It is obvious since  $\forall j \in \{1, \dots, k\}$ ,  $X^{(\text{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)}) \subset \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})$  thus  $(\mathbf{g}^{(k)})^{-1}(X^{(\text{tex})}) \in (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})}))$ , so it leads to  $(\mathbf{g}^{(k)})^{-1}(X^{(\text{tex})}) \in \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})) = \hat{\mathbf{g}}^{(k)}(X^{(\text{tex})})$ . Given this, we set  $z' = (\mathbf{g}^{(k)})^{-1}(X^{(\text{tex})})$  and based on Theorem.5,  $z'_{1:n_{\text{inv}}} = (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{(\text{tex})}) = z_{\text{inv}}$  is obtained.

In the second step, we make a contradiction to verify arbitrary elements in  $\hat{\mathbf{g}}^{(k)}(X^{(\text{tex})})$  fulfill the equality. Suppose that  $\exists z' \in \hat{\mathbf{g}}^{(k)}(X^{(\text{tex})})$  in the violation of  $z'_{1:n_{\text{inv}}} \neq (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{(\text{tex})})$ . To this, we consider the image-text generative process based on Assumption.4, where we define the sentence matrix  $X^{(\text{tex})'} = \mathbf{g}^{(k)}(z')$ . Due to  $\exists j \in \{1, \dots, k\}$  with  $X_{:,j}^{(\text{tex})'} \neq X_{:,j}^{(\text{tex})}$  otherwise  $z'_{1:n_{\text{inv}}} \neq (\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{(\text{tex})})$  can not be met, we have  $X^{(\text{tex})'} \in \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})'})$ . Besides,  $z' \in \hat{\mathbf{g}}^{(k)}(X^{(\text{tex})}) = \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})) \subset (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})}'))$ , it results in  $X^{(\text{tex})'} = \mathbf{g}^{(k)}(z') \in \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})$ , i.e.,  $X^{(\text{tex})'} \in \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})'}) \cap \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})$ . However,

$$\begin{aligned} & \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})'}) \cap \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})}) \\ &= (\mathcal{T}_{\text{tex}}^{(1)'} \times \dots \times \mathcal{T}_{\text{tex}}^{(j-1)'} \times \{X_{:,j}^{(\text{tex})'}\} \times \mathcal{T}_{\text{tex}}^{(j+1)'} \dots \mathcal{T}_{\text{tex}}^{(k)'}) \cap (\mathcal{T}_{\text{tex}}^{(1)} \times \dots \times \mathcal{T}_{\text{tex}}^{(j-1)} \times \{X_{:,j}^{(\text{tex})}\} \times \mathcal{T}_{\text{tex}}^{(j+1)} \dots \mathcal{T}_{\text{tex}}^{(k)}) \\ &= (\mathcal{T}_{\text{tex}}^{(1)} \cap \mathcal{T}_{\text{tex}}^{(1)}) \times \dots \times (\mathcal{T}_{\text{tex}}^{(j-1)'} \cap \mathcal{T}_{\text{tex}}^{(j-1)}) \times (\{X_{:,j}^{(\text{tex})'}\} \cap \{X_{:,j}^{(\text{tex})}\}) \times \dots \times (\mathcal{T}_{\text{tex}}^{(k)'} \cap \mathcal{T}_{\text{tex}}^{(k)}), \end{aligned} \quad (70)$$

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

Combine the two steps and the lemma has been proved.  $\square$

Based on Lemma.12,  $\forall X^{(\text{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)})$ , the set function  $\hat{\mathbf{g}}^{(k)}(X^{(\text{tex})})$  holds the output as a set composed of elements with their first  $n_{\text{inv}}$ -dim partition consistent with  $z_{\text{inv}}$ . Given this, we may define  $\hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}(X^{(\text{tex})})$  with the elements restricted on first  $n_{\text{inv}}$ -dim partition of the elements in  $\hat{\mathbf{g}}^{(k)}(X^{(\text{tex})})$ . Obviously  $\hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}(X^{(\text{tex})}) = \{z_{\text{inv}}\}$  so that we can define  $\hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}(X^{(\text{tex})}) = z_{\text{inv}}$  instead. To this end,  $g^{**}$  can be defined by  $\hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}$ :

$$g^{**} = \mathbf{d} \circ \hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)} : \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)}) \rightarrow (0, 1)^{n_{\text{inv}}}, \forall k \in \{1, \dots, k_{\text{max}}\}, \quad (71)$$

which replaces  $(\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}$  by  $\hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}$  in Eq.(66). Obviously,  $(\mathbf{g}^{(k)})_{1:n_{\text{inv}}}^{-1}(X^{(\text{tex})}) = \hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}(X^{(\text{tex})})$  for  $\forall X^{(\text{tex})} \in \mathcal{M}^{(k)}(\mathcal{X}_{\text{tex}}^{(k)})$  then  $g^{**}(X^{(\text{tex})}) = g^*(X^{(\text{tex})})$ , which results in  $\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^{**}) = 0$  from  $\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^*) = 0$ .

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

$$\begin{aligned} g^{**}([X_{:,1}^{(\text{tex})}, X_{:,2}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) &= g^{**}([X_{:,\pi(1)}^{(\text{tex})}, X_{:,\pi(2)}^{(\text{tex})}, \dots, X_{:,\pi(k)}^{(\text{tex})}]) \\ \iff \hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}([X_{:,1}^{(\text{tex})}, X_{:,2}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) &= \hat{\mathbf{g}}_{1:n_{\text{inv}}}^{(k)}([X_{:,\pi(1)}^{(\text{tex})}, X_{:,\pi(2)}^{(\text{tex})}, \dots, X_{:,\pi(k)}^{(\text{tex})}]) = z_{\text{inv}}^*. \end{aligned} \quad (72)$$

Then we return to

$$\hat{\mathbf{g}}^{(k)}(X^{(\text{tex})}) = \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})); \quad \hat{\mathbf{g}}^{(k)}(X_{\pi}^{(\text{tex})}) = \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,\pi(j)}^{(\text{tex})})), \quad (73)$$

where  $X_{\pi}^{(\text{tex})} = [X_{:,\pi(1)}^{(\text{tex})}, X_{:,\pi(2)}^{(\text{tex})}, \dots, X_{:,\pi(k)}^{(\text{tex})}]$ , and consider their union

$$\begin{aligned} &\hat{\mathbf{g}}^{(k)}(X^{(\text{tex})}) \cup \hat{\mathbf{g}}^{(k)}(X_{\pi}^{(\text{tex})}) \\ &= \left( \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})) \right) \cup \left( \bigcap_{j=1}^k (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,\pi(j)}^{(\text{tex})})) \right) \\ &= \bigcap_{j=1}^k \left( (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})) \cup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,\pi(j)}^{(\text{tex})})) \right). \end{aligned} \quad (74)$$

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

$$\mathbf{g}^{(k)}(\hat{z}) \in \hat{\mathcal{X}}(X_{:,j}^{(\text{tex})}) = \mathcal{T}_{\text{tex}}^{(1)} \times \dots \times \mathcal{T}_{\text{tex}}^{(j-1)} \times \{X_{:,j}^{(\text{tex})}\} \times \mathcal{T}_{\text{tex}}^{(j+1)} \dots \mathcal{T}_{\text{tex}}^{(k)}, \quad (75)$$

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

$$\mathbf{g}^{(k)}(\hat{z}) \in \hat{\mathcal{X}}(X_{:,\pi(j)}^{(\text{tex})}) = \mathcal{T}_{\text{tex}}^{(1,\pi)} \times \dots \times \mathcal{T}_{\text{tex}}^{(j-1,\pi)} \times \{X_{:,\pi(j)}^{(\text{tex})}\} \times \mathcal{T}_{\text{tex}}^{(j+1,\pi)} \dots \mathcal{T}_{\text{tex}}^{(k,\pi)}. \quad (76)$$

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

$$\mathbf{g}^{(k)}(\hat{z}) \in (\mathcal{T}_{\text{tex}}^{(1)} \cup \mathcal{T}_{\text{tex}}^{(1,\pi)}) \times \dots \times (\mathcal{T}_{\text{tex}}^{(j-1)} \cup \mathcal{T}_{\text{tex}}^{(j-1,\pi)}) \times \{X_{:,j}^{(\text{tex})}, X_{:,\pi(j)}^{(\text{tex})}\} \times \dots \times (\mathcal{T}_{\text{tex}}^{(k)} \cup \mathcal{T}_{\text{tex}}^{(k,\pi)}). \quad (77)$$

Hence  $\forall \hat{z} \in \bigcap_{j=1}^k ((\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})) \cup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,\pi(j)}^{(\text{tex})}))) = \hat{\mathbf{g}}^{(k)}(X^{(\text{tex})}) \cup \hat{\mathbf{g}}^{(k)}(X_{\pi}^{(\text{tex})})$ ,

$$\hat{X}^{(\text{tex})} = \mathbf{g}^{(k)}(\hat{z}) \in \{X_{:,1}^{(\text{tex})}, X_{:,\pi(1)}^{(\text{tex})}\} \times \dots \times \{X_{:,j}^{(\text{tex})}, X_{:,\pi(j)}^{(\text{tex})}\} \times \dots \times \{X_{:,k}^{(\text{tex})}, X_{:,\pi(k)}^{(\text{tex})}\} \quad (78)$$

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

$$g^{**}([X_{:,1}^{(\text{tex})}, X_{:,2}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\hat{\pi}(1)}^{(\text{tex})}, X_{:,\hat{\pi}(2)}^{(\text{tex})}, \dots, X_{:,\hat{\pi}(k)}^{(\text{tex})}]). \quad (79)$$

#### B.4 PROOF OF THEOREM.8

Our proof starts with  $g^{**}$  constructed in Eq.71. Given this, we consider the condition provided in Theorem.8:

$$g^{**}([X_{:,1}^{(\text{tex})}, \dots, X_{:,j}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\pi(1)}^{(\text{tex})}, \dots, \text{RF}(X_{:,j}^{(\text{tex})}), \dots, X_{:,\pi(k)}^{(\text{tex})}]), \quad (80)$$

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

$\{1, \dots, k\}$  where  $j$  fixed in the position. Follow the similar induction in Theorem.7 and we have  $\forall \hat{z} \in \cap_{j'=1, j' \neq j}^k ((\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j'}^{(\text{tex})})) \cup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,\pi(j')}^{(\text{tex})}))) \cap ((\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(X_{:,j}^{(\text{tex})})) \cup (\mathbf{g}^{(k)})^{-1}(\hat{\mathcal{X}}(\text{RF}(X_{:,j}^{(\text{tex})}))))$ ,

$$\hat{X}^{(\text{tex})} = \mathbf{g}^{(k)}(\hat{z}) \in \{X_{:,1}^{(\text{tex})}, X_{:,\pi(1)}^{(\text{tex})}\} \times \dots \times \{X_{:,j}^{(\text{tex})}, \text{RF}(X_{:,j}^{(\text{tex})})\} \times \dots \times \{X_{:,k}^{(\text{tex})}, X_{:,\pi(k)}^{(\text{tex})}\}. \quad (81)$$

Obviously, it holds  $\forall \hat{\pi}(X_{:,-j}^{(\text{tex})}) \in \Pi_{k-1}(\{1, \dots, j-1, j+1, \dots, k\}) \cap \{\{X_{:,1}^{(\text{tex})}, X_{:,\pi(1)}^{(\text{tex})}\} \times \dots \times \{X_{:,j-1}^{(\text{tex})}, X_{:,\pi(j-1)}^{(\text{tex})}\} \times \{X_{:,j+1}^{(\text{tex})}, X_{:,\pi(j+1)}^{(\text{tex})}\} \times \dots \times \{X_{:,k}^{(\text{tex})}, X_{:,\pi(k)}^{(\text{tex})}\}\}$  and  $\forall \hat{X}_j^{(1)}, \hat{X}_j^{(2)} \in \{X_{:,j}^{(\text{tex})}, \text{RF}(X_{:,j}^{(\text{tex})})\}$ ,

$$g^{**}([X_{:,1}^{(\text{tex})}, \dots, \hat{X}_j^{(1)}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,\hat{\pi}(1)}^{(\text{tex})}, \dots, \hat{X}_j^{(2)}, \dots, X_{:,\pi(k)}^{(\text{tex})}]). \quad (82)$$

## B.5 PROOF OF THEOREM 9

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

$$g^{(\ell)} : \mathcal{C}_{\text{inv}} \times \left( \prod_{i=1}^{\ell} \mathcal{S}_i \right) \times \mathcal{S}_{\text{pr}}^{\text{tex}} \longrightarrow \mathcal{M}^{(\ell)}(X_{\text{tex}}^{(\ell)})$$

be the diffeomorphism associated with  $\ell$ -token texts, with inverse  $(g^{(\ell)})^{-1}$  and projection  $(g^{(\ell)})_{1:n_{\text{inv}}}^{-1} : \mathcal{M}^{(\ell)}(X_{\text{tex}}^{(\ell)}) \rightarrow \mathcal{C}_{\text{inv}}$  to the first  $n_{\text{inv}}$  coordinates. By Theorem 5, there exist invertible heads  $h_{f^*}, h_{g^*}$  with  $f^* = h_{f^*} \circ \mathbf{f}_{1:n_{\text{inv}}}^{-1}$  and  $g^* = h_{g^*} \circ \mathbf{g}_{1:n_{\text{inv}}}^{-1}$ , such that  $\mathcal{L}_{\text{MMAlign}}^{(\text{img}, \text{tex})}(f^*, g^*) \rightarrow 0$ .

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

$$\mathcal{X}_{\text{ADD}} := \left\{ \hat{\pi}(X^{(\text{tex})}) = [X_{:,1}^{(\text{tex})}, \dots, X_{:,j}^{(\text{tex})}, \text{ADD}(X_{:,j}^{(\text{tex})}), X_{:,j+1}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}] : X^{(\text{tex})} \in \mathcal{X}_{\text{base}} \right\}.$$

The statement posits an intersection condition on the invariant component:

$$\exists z_{\text{inv}}^* \in \mathcal{C}_{\text{inv}} \quad \text{s.t.} \quad z_{\text{inv}}^* \in ((g^*)^{(j)})_{1:n_{\text{inv}}}^{-1}(\mathcal{X}_{\text{base}}) \cap ((g^*)^{(j+1)})_{1:n_{\text{inv}}}^{-1}(\mathcal{X}_{\text{ADD}}). \quad (83)$$

### B.5.1 SET-VALUED INVERSE CONSTRUCTIONS

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

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

$$\hat{g}^{(j)}(X) := ((g^{(j)})^{-1}(\mathcal{X}_{\text{base}})) \cap \Pi_{\text{fix}}^{(j)}(X), \quad (84)$$

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

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

$$\hat{g}^{(j+1)}(Y) := ((g^{(j+1)})^{-1}(\mathcal{X}_{\text{ADD}})) \cap \Pi_{\text{fix}}^{(j+1)}(X, Y), \quad (85)$$

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

**Lemma 13** (Constancy of  $z_{\text{inv}}$  on base and ADD families). *Under Eq.83, we have*

$$\hat{g}_{1:n_{\text{inv}}}^{(j)}(X) = \{z_{\text{inv}}^*\} \quad \forall X \in \mathcal{X}_{\text{base}}, \quad \hat{g}_{1:n_{\text{inv}}}^{(j+1)}(Y) = \{z_{\text{inv}}^*\} \quad \forall Y \in \mathcal{X}_{\text{ADD}}.$$

*Proof.* By assumption Eq.83, the first  $n_{\text{inv}}$  projections of the inverse preimages of  $\mathcal{X}_{\text{base}}$  (length  $j$ ) and  $\mathcal{X}_{\text{ADD}}$  (length  $j+1$ ) both contain  $z_{\text{inv}}^*$ . Intersecting with  $\Pi_{\text{fix}}^{(j)}(X)$  and  $\Pi_{\text{fix}}^{(j+1)}(X, Y)$  only constrains token-specific coordinates and the alignment of shared columns; it does not alter the  $1:n_{\text{inv}}$  coordinates. By block identifiability in Theorem 5, the  $1:n_{\text{inv}}$  projection is unique, hence each projection collapses to the singleton  $\{z_{\text{inv}}^*\}$ .  $\square$



### B.5.2 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:

$$g^{**}(Z) := \begin{cases} h_{g^*}(\hat{g}_{1:n_{\text{inv}}}^{(j)}(Z)), & Z \in \mathcal{X}_{\text{base}}, \\ h_{g^*}(\hat{g}_{1:n_{\text{inv}}}^{(j+1)}(Z)), & Z \in \mathcal{X}_{\text{ADD}}, \\ g^*(Z), & \text{otherwise.} \end{cases} \quad (86)$$

**ADD invariance.** Let  $X \in \mathcal{X}_{\text{base}}$  and  $Y = \hat{\pi}(X) \in \mathcal{X}_{\text{ADD}}$  be the ADD-form hard negative described in the theorem. By Lemma 13,  $\hat{g}_{1:n_{\text{inv}}}^{(j)}(X) = \hat{g}_{1:n_{\text{inv}}}^{(j+1)}(Y) = \{z_{\text{inv}}^*\}$ . Therefore,

$$g^{**}(X) = h_{g^*}(z_{\text{inv}}^*) = g^{**}(Y),$$

which proves equation 16:

$$g^{**}([X_{:,1}^{(\text{tex})}, \dots, X_{:,k}^{(\text{tex})}]) = g^{**}([X_{:,1}^{(\text{tex})}, \dots, X_{:,j}^{(\text{tex})}, \text{ADD}(X_{:,j}^{(\text{tex})}), \dots, X_{:,k}^{(\text{tex})}]).$$

**Optimal alignment preserved.** By construction equation 86,  $g^{**} = g^*$  outside  $\mathcal{X}_{\text{base}} \cup \mathcal{X}_{\text{ADD}}$ . On  $\mathcal{X}_{\text{base}}$  and  $\mathcal{X}_{\text{ADD}}$ , Lemma 13 ensures that  $g^{**}$  applies the same invertible head  $h_{g^*}$  to the same invariant  $z_{\text{inv}}^*$  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:

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

Under the intersection condition in Eq.83, 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  $\text{ADD}(X_{:,j}^{(\text{tex})})$  at position  $j$ , establishing ADD-form composition nonidentifiability.

## C APPENDIX.C

### C.1 IMPLEMENTATION OF THEOREM.7,8

**Algorithms.** Theorem.7,8 refer to the corresponding data augmentation algorithms illustrated in Algo.1. We present the prompts for hard negative data generation and the experimental evaluation as below:

**Prompt (re-ordering instruction):** Read the text  $\langle \rangle$ , then permute its token order to generate a text that holds the same or most similar semantic with  $\langle \rangle$ ;

**Prompt (rephrasing instruction):** Read the text  $\langle \rangle$ , then replace one of its language token by an arbitrary word or phrase from its all possible token permutation obtained by the following instruction:  
 <Prompt (re-ordering instruction)>  
 such that the generated text holds the same or most similar semantic with  $\langle \rangle$ ;

**Prompt (evaluation):** Given a text: <a text drawn from ARO>  
 Identify whether the prompt can be used to generate the text:  
 1. <Prompt (re-ordering instruction)>, choose a combination of  $\times_i \{ \text{<Prompt (re-ordering instruction)>}_i, \text{<a text drawn from ARO>}_i \}$  ( $\times_i$  indicates Cartesian product for the  $i$ -th token.) that holds the identical tokens with <a text drawn from ARO>;  
 2. <Prompt (rephrasing instruction)>, choose a combination of  $\times_i \{ \text{<Prompt (re-ordering instruction)>}_i, \text{<a text drawn from ARO>}_i \}$  ( $\times_i$  indicates Cartesian product for the  $i$ -th token.) that holds the identical tokens with <a text drawn from ARO>.

**Algorithm 1** Hard negative text generation derived from Theorem.7-9**Input:** A image-text pair  $\langle x^{(\text{img})}, X^{(\text{text})} \rangle$ **Parameter:** local LLM service,  $f, g$ .**Output:** A hard negative text  $\hat{X}^{(\text{text})}$  re-ordered / rephrased from  $X^{(\text{text})}$ 

```

1: Do some action.
2: if "SWAP" == true then
3:   Instruct LLM to identify a token  $\pi$ -permutation  $\bar{X}^{(\text{text})}$  of  $X^{(\text{text})}$  with a close semantic.
4:   Generate a set of token permutation of  $X^{(\text{text})}$  that satisfy  $\bar{X}^{(\text{text})} \cap \{X_{:,1}^{(\text{tex})}, X_{:,\pi(1)}^{(\text{tex})}\} \times \dots \times$ 
    $\{X_{:,k}^{(\text{tex})}, X_{:,\pi(k)}^{(\text{tex})}\}$ , rank them by their CLIP score and choose the top-1 as  $\hat{X}^{(\text{text})}$ .
5: else
6:   if "REPLACE" == true then
7:     Instruct LLM to identify a token  $\pi$ -permutation  $\bar{X}^{(\text{text})}$  of  $X^{(\text{text})}$  replaced a token
      $\text{RF}(X^{(\text{text})})$  have a close semantic.
8:     Generate a set of token permutation of  $X^{(\text{text})}$  that satisfy  $\bar{X}^{(\text{text})} \cap \{X_{:,1}^{(\text{tex})}, X_{:,\pi(1)}^{(\text{tex})}\} \times$ 
      $\dots \{X_{:,j-1}^{(\text{tex})}, X_{:,\pi(j-1)}^{(\text{tex})}\} \times \{X_{:,j+1}^{(\text{tex})}, X_{:,\pi(j+1)}^{(\text{tex})}\} \dots \times \{X_{:,k}^{(\text{tex})}, X_{:,\pi(k)}^{(\text{tex})}\}$  with regards to
      $\forall \hat{X}_j^{(1)}, \hat{X}_j^{(2)} \in \{X_{:,j}^{(\text{tex})}, \text{RF}(X_{:,j}^{(\text{tex})})\}$ , rank them by their CLIP score and choose the top-1
     as  $\hat{X}^{(\text{text})}$ .
9:   else
10:    if "ADD" == true then
11:      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  $\text{ADD}(X^{(\text{text})})$ .
12:      Calling  $g^*$  to rank the cosine distance between  $g^*(X^{(\text{text})})$  and  $g^*(\text{ADD}(X^{(\text{text})}))$ , choose
      the highest as  $\text{ADD}(X^{(\text{text})})$ .
13:    else
14:      "No compositional hard negative generated."
15:    end if
16:  end if
17: end if
18: return A hard negative text  $\hat{X}^{(\text{text})}$  re-ordered / rephrased from  $X^{(\text{text})}$ .

```

We employed Deepseek R1 to execute the first and the second prompt to facilitate our algorithm, while employed Gemini 2.5 Pro to achieve the experimental verification in Fig.3. It helps to prevents the self-enhancement bias in LLM-as-a-Judge Zheng et al. (2023).