SmartCLIP: Modular Vision-language Alignment with Identification Guarantees

Shaoan Xie*[1], Lingjing*[1], Yujia Zheng[1], Yu Yao[3], Zeyu Tang[1],
Eric P. Xing[1,2], Guangyi Chen[2,1], Kun Zhang[1,2]

[1] Carnegie Mellon University

[2] Mohamed bin Zayed University of Artificial Intelligence

[3] University of Sydney

* Equal contribution

Abstract

Contrastive Language-Image Pre-training (CLIP) [37] has emerged as a pivotal model in computer vision and multimodal learning, achieving state-of-the-art performance at aligning visual and textual representations through contrastive learning. However, CLIP struggles with potential information misalignment in many image-text datasets and suffers from entangled representation. On the one hand, short captions for a single image in datasets like MSCOCO may describe disjoint regions in the image, leaving the model uncertain about which visual features to retain or disregard. On the other hand, directly aligning long captions with images can lead to the retention of entangled details, preventing the model from learning disentangled, atomic concepts – ultimately limiting its generalization on certain downstream tasks involving short prompts.

In this paper, we establish theoretical conditions that enable flexible alignment between textual and visual representations across varying levels of granularity. Specifically, our framework ensures that a model can not only preserve *cross-modal semantic information in its entirety but also* disentangle *visual representations to capture fine-grained textual concepts. Building on this foundation, we introduce* ***SmartCLIP****, a novel approach that identifies and aligns the most relevant visual and textual representations in a modular manner. Superior performance across various tasks demonstrates its capability to handle information misalignment and supports our identification theory. The code is available at* https://github.com/MidPush/SmartCLIP.

1. Introduction

Contrastive Language-Image Pre-training (CLIP) [37] has been the cornerstone for many computer vision and machine learning tasks, such as text-to-image retrieval [2], image and video understanding [6, 16, 39, 42, 46, 58], and generative models [28, 38, 40]. It aligns the representations from different modalities with a contrastive learning loss [8, 35]. Specifically, each image-caption pair in the dataset is treated as a positive pair, while negative pairs are created by matching images with captions randomly drawn from the dataset. The image and text encoders are trained with a symmetric cross-entropy loss that draws the image and text representations in each positive pair together while pulling the negative pairs' representations apart.

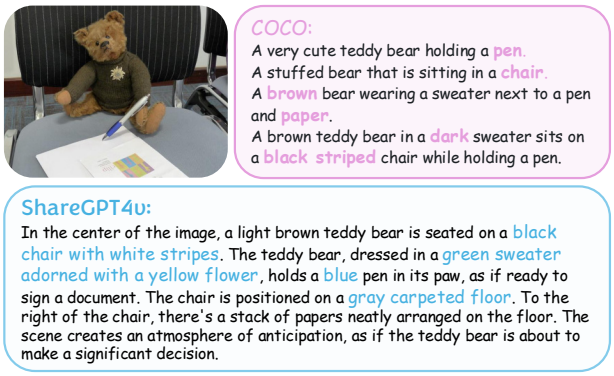

Figure 1. **Depiction of two primary challenges for CLIP.** (1) Information Misalignment: An image can be paired with multiple captions that describe disparate aspects - the first caption contains concepts "bear" and "pen" whereas the second only mentions "bear" and "paper". Aligning the image with both captions leads to the loss of key concepts "pen" and "paper" not shared across the captions. (2) Entangled Representations: Long, detailed captions involving multiple concepts (e.g., "chair", "pen", "flower", "floor") encourage the model to form entangled representations, hindering independent understanding of each individual concept.

Training CLIP requires vast amounts of image-text pairs, making it challenging to maintain dataset quality at such a large scale. In particular, the quality of text captions has been a key concern, prompting the development of various methods to enhance both their diversity and accu-

racy. ALIGN [15] shows that scaling up the dataset size can compensate for the noisy text supervision. BLIP models [23, 24] improve captions by incorporating additional captioning and filtering mechanisms. VE-CLIP [21] introduces a visual-enriched captioning approach to further refine caption quality. Similarly, LaCLIP [10] leverages language models to rewrite captions, while RecapCLIP [25] uses LLaMA-3 [29] to generate captions for 1.3 billion images. Despite these efforts, recent findings reveal that longer and seemingly higher-quality captions do not necessarily yield improved performance on many downstream tasks [20]. Li et al. [25] find that the text-to-image retrieval accuracy on Flickr30K drops from 84.2 to 74.1 when replacing the original captions with longer captions.

A key issue contributing to the observed performance degradation is the information misalignment between images and their captions, a problem that becomes more pronounced when multiple captions are paired with a single image. On one hand, an image may be paired with several captions, each capturing only a partial aspect of the image. In Figure 1, aligning the image with the first caption, "A very cute teddy bear holding a pen", risks forcing the model to discard other important concepts like "chair" and "paper", which are required to align with the second and third captions in the pink text box. This misalignment between image and text introduces conflicts during standard CLIP training, leading to the loss of key visual concepts.

On the other hand, training CLIP with long, detailed captions, as seen in recent approaches [10, 21, 25], encourages the model to learn entangled representations of multiple concepts bundled together in a single caption. Thus, it remains challenging to explicitly extract object/concept-centric representations from CLIP's visual representation. This entanglement is particularly problematic for tasks that require individual, atomic concepts or novel combinations of them, as empirically observed on short-text-to-image retrieval tasks [54]. In Figure 1, the long caption generated by ShareGPT4V [7] contains an exhaustive set of concepts such as "chair", "pen", "flower", and "floor". This aggregation can hinder the model's performance on tasks that demand the understanding of each concept individually.

In this paper, we propose a refined approach to representation alignment in vision-language models like CLIP [37]. We frame the alignment challenge as a latent-variable identification problem and develop theoretical conditions that enable flexible alignment between textual and visual representations at different levels of granularity. Our framework enables the model to *preserve* the complete cross-modal information while also *disentangling* representations to capture fine-grained concepts, effectively addressing the misalignment and disentanglement issues discussed earlier.

Building on these theoretical insights, we introduce **SmartCLIP**, a novel method that identifies and aligns visual and textual concepts in a modular manner. Specifically, we design a mask network that selects a subset of dimensions from the full representations, corresponding to only the concepts present in each specific caption. This allows the model to perform text-image alignment over the most relevant concepts modules, rather than the entire representation. We empirically demonstrate that **SmartCLIP** outperforms state-of-the-art models across a range of downstream tasks, showcasing its effectiveness in addressing alignment challenges. In particular, **SmartCLIP** significantly improves retrieval performance across text lengths, achieving 98.7% accuracy (up from 78.2%) on the ShareGPT4V long text-to-image retrieval tasks, while boosting short text-to-image retrieval R1 from 56.1% to 66.0%.

Our main contributions are summarized as follows.

i We identify critical issues of information misalignment and entangled representations within the CLIP framework. To overcome these challenges, we propose a latent-variable formulation and establish theoretical conditions that guarantee the recovery of the latent variables.

ii Building upon our theoretical findings, we propose **SmartCLIP**, featuring adaptive masking and a modular contrastive learning objective that facilitates the learning of disentangled, modular representations.

iii We perform extensive experiments on a variety of tasks, including long and short text-to-image retrieval, zero-shot classification, and text-to-image generation. **SmartCLIP** consistently outperforms or matches state-of-the-art models across these benchmarks, demonstrating its efficacy and validating our theoretical contributions.

2. Related Work

Vision-language models. The breakthrough of CLIP [37] has attracted significant attention from the community. SLIP [33] and DECLIP [26] propose to incorporate self-supervised learning techniques to improve the learned representation. Coca [51] introduces a decoder in addition to the contrastive learning branch. LiT [52] locks the image encoder and only finetunes the text encoder. SigLIP [53] adopts a simple sigmoid loss to handle large training batch sizes. LoTLIP [48] inserts corner tokens after the classification token to support long-text understanding. TULIP [34] replaces the absolute position embedding with the relative position embedding to support longer text understanding. ALIGN [15] demonstrates that increasing dataset size can mitigate the impact of noisy text supervision. Recent methods have been focusing on generating better captions [20, 21, 23–25, 59]. CLIP-MOE [56] introduces mixture-of-experts to CLIP. LLM2CLIP [12] augments CLIP with large language models. LongCLIP [54] extends the token constraint of CLIP from 77 to 248 and applies PCA to perform short text-to-image contrastive learning to preserve its short text capability. Llip [22] learns a text-dependent vi-

sual representation by mixing a set of learnable tokens with a cross-attention module. In contrast, **SmartCLIP** directly learns a single global representation that encodes all disentangled, interpretable concepts through masking.

Latent variable identification. Learning high-level, semantic information from low-level observational data (e.g., images and text) can often be formulated as latent-variable identification problems. Though appealing, such tasks are accompanied by substantial difficulties, especially for complex real-world data distributions involving nonlinear generating functions. Recently, a wealth of papers [1, 5, 9, 11, 14, 17, 31, 32, 41, 44, 49, 55, 57] propose to overcome such obstacles by leveraging auxiliary information, such as temporal information, multiple domains, and multiple views/modalities. Especially related to our work are those that tap into the paired multi-view data to identify the shared information across available views [9, 11, 31, 32, 43, 49]. Recent work [11, 31, 32] relies on specific forms of latent variable distributions (e.g., independence or exponential family). These constraints restrict their applicability for distributions entailing complex interactions among latent variables. Prior work [9, 43] adopt more flexible assumptions on the underlying distribution and identify blocks of latent variables directly shared by two views arising from data augmentations, which is extended to a multi-view setting [49]. The problem under investigation can be viewed as a form of this multi-view setting where the paired image and text captions are considered as views sharing the semantic latent variable. Existing works [9, 43, 49] assume that views are grouped over all the data pairs and this grouping information is known so that one can learn a designated encoder for each view group. However, this view grouping information is inaccessible for our problems – for any two text captions of different images, we cannot judge whether they belong to the same view group. In our theory section, we show that by properly utilizing the data-generating process, we can learn such information directly and further achieve the desired identification results, thus generalizing existing multi-view latent variable identification results.

3. Problem Formulation

As motivated previously, we aim to 1) preserve all the semantic information shared across modalities, and 2) learn a disentangled representation that corresponds to textual concepts at diverse granularity levels. To this end, we propose the following data-generating process underlying the visual-language data distribution.

Notations. We indicate the dimensionality of a vector with $d(\cdot)$. We denote a subset of dimensions of a vector $\mathbf{z}$ with $[\mathbf{z}]_{\mathcal{B}}$ with the index set $\mathcal{B}$. We define the set of indices whose corresponding values are nonzero in a vector $\mathbf{m}$ with $\mathcal{B}(\mathbf{m}) := \{i \in d(\mathbf{m}) : [\mathbf{m}]_i \neq 0\}$.

Data-generating processes. We depict the data-generating process in Figure 2 and in (1).

$$\mathbf{z}_{\mathrm{T}} := \mathbf{z}_{\mathrm{I}} \odot \mathbf{m};\ \mathbf{i} := g_{\mathrm{I}}(\mathbf{z}_{\mathrm{I}}, \boldsymbol{\epsilon}_{\mathrm{I}});\ \mathbf{t} := g_{\mathrm{T}}(\mathbf{z}_{\mathrm{T}}, \boldsymbol{\epsilon}_{\mathrm{T}}). \quad (1)$$

We assume that each pair of image $\mathbf{i} \in \mathcal{I} \subset \mathbb{R}^{d(\mathbf{i})}$ and text caption $\mathbf{t} \in \mathcal{T} \subset \mathbb{R}^{d(\mathbf{t})}$ originate from semantic information $\mathbf{z}_{\mathrm{I}} \in \mathcal{Z}_{\mathrm{I}} \subset \mathbb{R}^{d(\mathbf{z}_{\mathrm{I}})}$, together with modality-specific variations $\boldsymbol{\epsilon}_{\mathrm{I}}$ and $\boldsymbol{\epsilon}_{\mathrm{T}}$ (e.g., illumination for images, tenses for text), through generating functions $g_{\mathrm{I}} : (\mathbf{z}_{\mathrm{I}}, \boldsymbol{\epsilon}_{\mathrm{I}}) \mapsto \mathbf{i}$ and $g_{\mathrm{T}} : (\mathbf{z}_{\mathrm{T}}, \boldsymbol{\epsilon}_{\mathrm{T}}) \mapsto \mathbf{t}$ respectively. We treat the text caption as continuous variables, as each word can be represented with a continuous word embedding vector in practice [3, 30].

As demonstrated in Figure 1, text captions of the same image often convey partial information of the entire image semantics. Thus, we associate each text caption's representation $\mathbf{z}_{\mathrm{T}} := \mathbf{m} \odot \mathbf{z}_{\mathrm{I}}$ with a binary random mask $\mathbf{m} \in \mathcal{M} \subset \{0, 1\}^{d(\mathbf{z}_{\mathrm{I}})}$ that eliminates information absent in the specific caption $\mathbf{t}$.

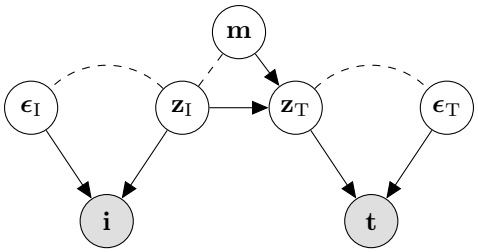

Figure 2. **The data-generating process.** The text representation $\mathbf{z}_{\mathrm{T}}$ contains partial information of the vision representation $\mathbf{z}_{\mathrm{I}}$, as indicated by masking $\mathbf{m}$. Modal-specific information is represented as $\boldsymbol{\epsilon}_{\mathrm{I}}$ and $\boldsymbol{\epsilon}_{\mathrm{T}}$. Dashed edges indicate potential statistical dependence.

Goals. Our two goals can be formalized as follows.

a *Preserving cross-modal information*: identifying the complete latent representation $\mathbf{z}_{\mathrm{I}}$.

b *Disentangling concepts*: identifying concepts $\mathbf{z}_{\mathrm{T}}$ associated with a given textual description $\mathbf{t}$ at different granularity levels potentially unseen during training.

Examples. In Figure 1, the image $\mathbf{i}$ contains concepts "bear", "chair", and "pen", which we assume correspond to three components in the representation $\mathbf{z}_{\mathrm{I}}$, say $[\mathbf{z}_{\mathrm{I}}]_1$, $[\mathbf{z}_{\mathrm{I}}]_2$, and $[\mathbf{z}_{\mathrm{I}}]_3$. The first two COCO captions $\mathbf{t}^{(1)}$ and $\mathbf{t}^{(2)}$ mention only a subset of these concepts, i.e., ("bear", "pen") and ("bear", "chair") separately. Thus, the masks for these captions are $\mathbf{m}^{(1)} = [1, 0, 1]$ and $\mathbf{m}^{(2)} = [1, 1, 0]$. The variables $\boldsymbol{\epsilon}_{\mathrm{I}}$ and $\boldsymbol{\epsilon}_{\mathrm{T}}$ represent modality-specific nuance factors such as illumination conditions in the image $\mathbf{i}$ and syntax in the text $\mathbf{t}$. For Goal a, we seek to *preserve* the complete information $\mathbf{z}_{\mathrm{I}}$. This involves retaining all relevant textual concepts present in the captions, namely "bear", "chair", and "pen" from captions $\mathbf{t}^{(1)}$ and $\mathbf{t}^{(2)}$. For Goal b, we intend to *disentangle* the representation $\mathbf{z}_{\mathrm{I}}$ into finer concept blocks potentially unseen in the training. This includes identifying individual concepts such as "bear" in the dimen-

sion $[\mathbf{z}_\mathrm{I}]_1$, even if the training captions only include "bear" in combination with other concepts.

4. Identification Theory

In this section, we present the theoretical results towards Goal a and Goal b. We show that under a suitable learning objective (2), the learned variables $(\hat{\mathbf{z}}_\mathrm{I}, \hat{\mathbf{z}}_\mathrm{T})$ can be identified with the corresponding true quantities $(\mathbf{z}_\mathrm{I}, \mathbf{z}_\mathrm{T})$ up to certain equivalent classes. In particular, we resort to the block-wise identifiability [18, 19, 43, 49] throughout this work. This suffices for our purpose since often several dimensions jointly (i.e., a block) comprise a meaningful concept while a single dimension may not be interpretable.

Definition 4.1 (Block-wise Identifiability)**.** The true variable $\mathbf{v}$ is block-wise identifiable if it is related to its estimate $\hat{\mathbf{v}}$ through an invertible map $\mathbf{v} \mapsto \hat{\mathbf{v}}$.

The learning objective. Our estimation model consists of vision/text encoders $(f_\mathrm{I}, f_\mathrm{T})$ (smooth, invertible functions), and a masking function $\hat{\mathbf{m}} : \mathcal{T} \to \mathcal{M}$ that estimates the true mask $\mathbf{m}$ underlying a given text caption $\mathbf{t}$.

$$\begin{aligned} &\underbrace{\arg\min_{f_\mathrm{I}, f_\mathrm{T}, \hat{\mathbf{m}}} \|\hat{\mathbf{m}}(\mathbf{t})\|_0}_{L_\mathrm{sparsity}}, \quad \text{subject to:} \\ &\underbrace{\arg\min_{f_\mathrm{I}, f_\mathrm{T}, \hat{\mathbf{m}}} \|f_\mathrm{I}(\mathbf{i}) \odot \hat{\mathbf{m}}(\mathbf{t}) - f_\mathrm{T}(\mathbf{t})\|}_{L_\mathrm{align}}, \forall (\mathbf{i}, \mathbf{t}). \end{aligned} \tag{2}$$

Our learning objective (2) consists of an alignment term L_align that draws the positive pairs across modalities. The negative pairs in regular contrastive losses [8, 35, 37] can be implemented through an entropy term at the sample limit [47]. This serves the same role as the invertibility condition on the encoder models [43], which we directly assume for theoretical convenience. In Section 5, we discuss practical considerations for constructing negative pairs. We enforce sparsity regularization L_sparsity on the inferred mask $\hat{\mathbf{m}}$ to select the simplest representation.

We introduce our key conditions in Condition 4.2 and theoretical results in Theorem 4.3.

Condition 4.2 (Identification Conditions)**.**

i [Smoothness & invertibility]: Generating functions g_I and g_T are smooth and have smooth inverses.

ii [Fully-supported $p(\mathbf{z}_\mathrm{I}, \mathbf{m})$]: The joint distribution over the semantic variable $\mathbf{z}_\mathrm{I}$ and the mask $\mathbf{m}$ is fully supported: $p(\mathbf{z}_\mathrm{I}, \mathbf{m}) > 0$ for any $(\mathbf{z}_\mathrm{I}, \mathbf{m}) \in \mathcal{Z}_\mathrm{I} \times \mathcal{M}$.

Discussion. Condition 4.2-*i* ensures the generating functions $(g_\mathrm{I}, g_\mathrm{T})$ preserve latent variables' information, without which the task of recovering such latent variables would be ill-posed. Practically, the high dimensionality of image data $\mathbf{i}$ offers sufficient capacity to hold all information, and the text variable $\mathbf{t}$ only contains information filtered through its mask $\mathbf{m}$. This condition is widely employed in the latent-variable identification literature [13, 17, 18, 43]. Condition 4.2-*ii* prescribes that the representation $\mathbf{z}_\mathrm{I}$ and the mask $\mathbf{m}$ that marginally appear in the training distribution should also be present jointly with non-zero probability density. Interpreting the mask $\mathbf{m}$ as a concept selector (e.g., selecting "bear" and "pen"), this condition ensures that each concept retains its full range of variations (such as different shapes of bears and lengths of pens) across various mask selections. To satisfy this requirement, one can restrict the joint support $\mathcal{Z}_\mathrm{I} \times \mathcal{M}$ to an appropriate subset, ensuring that only relevant combinations of $\mathbf{z}_\mathrm{I}$ and $\mathbf{m}$ are present. Alternatively, one can enrich the caption set for each image, thereby increasing the diversity and coverage of concept combinations and filling in the joint support. This aligns with recent caption-augmentation techniques [21, 23–25, 54] as discussed in Section 1, revealing the synergy between our framework and existing efforts in the community.

Theorem 4.3 (Concept Representation Identification)**.** *We assume the data-generating process in (1). Let $(f_\mathrm{I}, f_\mathrm{T}, \hat{\mathbf{m}})$ be an optimum of (2). Under Condition 4.2, the true representation $[\mathbf{z}]_{\tilde{\mathcal{B}}}$ is block-wise identifiable for any index set $\tilde{\mathcal{B}}$ such that $\tilde{\mathcal{B}} = \cup_{\mathbf{m}\in\mathcal{V}} \mathcal{B}(\mathbf{m})$ or $\tilde{\mathcal{B}} = \cap_{\mathbf{m}\in\mathcal{V}} \mathcal{B}(\mathbf{m})$ over any subset of masks $\mathcal{V} \subset \mathcal{M}$.*

Concept preservation. Theorem 4.3 states that one can recover the concept block $[\mathbf{z}_\mathrm{I}]_{\mathcal{B}(\mathbf{m})}$ associated with each individual text caption in the dataset $\mathcal{M}$. [1] Furthermore, it ensures that the union of concepts $[\mathbf{z}_\mathrm{I}]_{\cup_{\mathbf{m}\in\mathcal{V}}\mathcal{B}(\mathbf{m})}$ from any subset of text captions $\mathcal{V} \subset \mathcal{M}$ can be preserved. In the running example of Figure 1, our formulation allows us to preserve concepts ("bear", "pen", "chair") in the image representation by selectively matching them with the two captions, whereas existing models like CLIP may lose either "pen" and "chair" since they are only mentioned in one caption. Therefore, Theorem 4.3 effectively addresses Goal a.

Concept disentanglement. The intersection operation in Theorem 4.3 empowers us to *disentangle* representations into potentially atomic concepts. In the example of Figure 1, we can identify the concept "bear" as the intersection of the two text captions, despite the absence of a standalone caption containing only "bear" in the dataset. Consequently, this part of the statement tackles Goal b. Our results underscore the importance of associating each image with a diverse set of captions that share overlapping concepts.

Theoretical contribution. Theorem 4.3 extends existing theoretical frameworks [9, 43, 49]. Notably, Yao et al. [49] provide identification guarantees for shared representations over multiple views, generalizing earlier results confined to two views [9, 43]. This multi-view formulation

[1] We refer to a text caption with its mask $\mathbf{m}$ to simplify the notation.

is analogous to our setup when considering each group of text captions $\mathbf{t}$ associated with the same mask $\mathbf{m}$ as a distinct view group. However, prior work [49] relies on explicit knowledge of these groupings to train view-specific encoders. In contrast, our problem setting presents a greater challenge since we have no access to this grouping information. Specifically, given two captions of any different images, it is unclear whether they stem from the same mask (i.e., the same view group). Therefore, the identification guarantees in prior studies do not apply to our setting. Theorem 4.3 demonstrates that our estimation model, paired with the learning objective (2), can automatically infer the necessary grouping information (i.e., the masks). By doing so, our approach relaxes the identification conditions in previous work, enabling effective representation identification without explicit group knowledge.

5. SmartCLIP: Modular Vision-language Alignment

Drawing on the theoretical framework in Section 4, we present **SmartCLIP**, a modular alignment model designed to achieve Goal a and Goal b. We discuss the implementation of the learning objective (2) and the model architecture.

Modular alignment through adaptive masking. The masking function $\hat{\mathbf{m}}(\cdot)$ is instrumental in our modular alignment objective (2). **SmartCLIP** implements this function with a transformer block that ingests a caption representation $\hat{\mathbf{z}}_{\mathrm{T}}$ as its input and outputs a binary vector $\hat{\mathbf{m}}(\hat{\mathbf{z}}_{\mathrm{T}})$ via a straight-through estimator [4].

Modular contrast construction. As discussed in Section 4, the negative pairs in regular contrastive losses [8, 35] serves a similar role as the invertibility assumption (Condition 4.2-*i*) [43, 47]. We denote generic image, text representations as $\mathbf{I}$, $\mathbf{T}$, and positive, negative pairs with $\mathbf{P}_{\mathrm{pos}}$, $\mathbf{P}_{\mathrm{neg}}$ respectively. The canonical one-side contrastive loss $\mathcal{L}_{\mathrm{ctr}}$ [37] is defined as:

$$\begin{aligned}&\mathcal{L}_{\mathrm{ctr}}\Big(\underbrace{\left(\mathbf{I}^{(i)},\mathbf{T}^{(i)}\right)}_{\mathbf{P}_{\mathrm{pos}}},\underbrace{\left(\mathbf{I}^{(i)},\mathbf{T}^{(j)}\right)}_{\mathbf{P}_{\mathrm{neg}}}\Big)\\&=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\mathrm{sim}(\mathbf{I}^{(i)},\mathbf{T}^{(i)})}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\mathrm{sim}(\mathbf{I}^{(i)},\mathbf{T}^{(j)})}{\tau}\right)},\end{aligned}\tag{3}$$

where we denote the sample size, temperature, and cosine similarity with N, τ, and $\mathrm{sim}(\cdot)$ respectively.

Following the symmetric contrastive loss for CLIP [37], our alignment loss consists of two contrastive loss terms $\mathcal{L}_{\mathrm{ctrI}}$ and $\mathcal{L}_{\mathrm{ctrT}}$ that differ in the negative pairs:

$$\mathcal{L}_{\mathrm{ctrI}}:=\mathcal{L}_{\mathrm{ctr}}\left(\mathbf{P}_{\mathrm{pos}},\mathbf{P}_{\mathrm{negI}}\right),\ \mathcal{L}_{\mathrm{ctrT}}:=\mathcal{L}_{\mathrm{ctr}}\left(\mathbf{P}_{\mathrm{pos}},\mathbf{P}_{\mathrm{negT}}\right),\tag{4}$$

with the positive and negative pairs defined as follows:

$$\mathbf{P}_{\mathrm{pos}}:=\left(\hat{\mathbf{z}}_{\mathrm{I}}^{(i)}\odot\hat{\mathbf{m}}(\hat{\mathbf{z}}_{\mathrm{T}}^{(i)}),\ \hat{\mathbf{z}}_{\mathrm{T}}^{(i)}\right),\tag{5}$$

$$\mathbf{P}_{\mathrm{negI}}:=\left(\hat{\mathbf{z}}_{\mathrm{I}}^{(i)}\odot\hat{\mathbf{m}}(\hat{\mathbf{z}}_{\mathrm{T}}^{(j)}),\ \hat{\mathbf{z}}_{\mathrm{T}}^{(j)}\right),\tag{6}$$

$$\mathbf{P}_{\mathrm{negT}}:=\left(\hat{\mathbf{z}}_{\mathrm{I}}^{(j)}\odot\hat{\mathbf{m}}(\hat{\mathbf{z}}_{\mathrm{T}}^{(i)}),\ \hat{\mathbf{z}}_{\mathrm{T}}^{(i)}\right).\tag{7}$$

In particular, $\mathbf{P}_{\mathrm{negI}}$ contrasts the image representation $\hat{\mathbf{z}}_{\mathrm{I}}^{(i)}$ in the positive pair with randomly sampled caption representations $\hat{\mathbf{z}}_{\mathrm{T}}^{(j)}$ (see the green region in Figure 3), whereas $\mathbf{P}_{\mathrm{negT}}$ contrasts the text representation $\hat{\mathbf{z}}_{\mathrm{T}}^{(i)}$ in the positive pair with randomly sampled image representations $\hat{\mathbf{z}}_{\mathrm{I}}^{(j)}$ (see the orange region in Figure 3).

Sparsity penalty. We implement L_{sparsity} in (2) with a ℓ_1 term for its compatibility with deep-learning training:

$$\mathcal{L}_{\mathrm{sparsity}}=\|\hat{\mathbf{m}}(\mathbf{t})\|_1.\tag{8}$$

This term ensures that the textual concepts are encoded into a minimal number of latent dimensions, promoting the disentanglement of distinct concepts across text captions.

SmartCLIP training objective. In summary, the training objective of **SmartCLIP** is a weighted sum of loss terms in (4) and (8):

$$\mathcal{L}=\lambda_{\mathrm{align}}\cdot(\mathcal{L}_{\mathrm{ctrI}}+\mathcal{L}_{\mathrm{ctrT}})+\lambda_{\mathrm{sparsity}}\cdot\mathcal{L}_{\mathrm{sparsity}},\tag{9}$$

where λ_{align} and $\lambda_{\mathrm{sparsity}}$ denote the weighting coefficients.

6. Experiments

6.1. Setup

Implementation details. Following Long-CLIP [54], we finetune the CLIP model [37] on ShareGPT4V [7], which contains around 1 million image-text pairs. We employ the position encoding in long-CLIP to handle 248 tokens (c.f., the 77-token limit in the original CLIP). Compared to the baseline CLIP model, we introduce a mask network $\hat{\mathbf{m}}$. The masking network is designed as a single transformer block, which takes the text sequence embedding $\hat{\mathbf{z}}_{\mathrm{T}}$ from the text encoder. Then we add an attention-pooling layer to down-sample it to the same size as the CLIP representation, e.g., 768 in ViT-L/14. We tested including more transformer blocks in the mask network but did not observe significant improvements. Therefore, we stick to one block for faster training and inference. Unlike Long-CLIP [54] which processes all the captions for each image at each gradient step, we only sample one caption for each image, reducing our overall training time by half. Specifically, on 8 H100 GPUs, training a Vit-B/16 model for one epoch takes about 4 minutes with our model, whereas it takes around 7 minutes for Long-CLIP. After the pooling layer, we apply sigmoid to

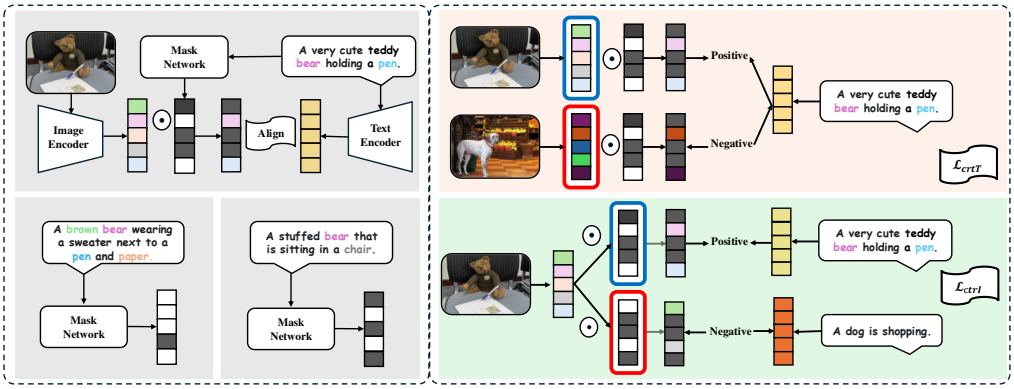

Figure 3. **The diagram of SmartCLIP**. On the left, we introduce adaptive masking for alignment with different text prompts. The mask network selects which part of the image representation to be used. On the right, we present our modular contrastive objectives (4).

Table 1. **Results of short-caption text-image retrieval on the 5k COCO2017 validation set and the whole 30k Flickr30K dataset.**

		COCO						Flickr30k					
		Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
B/16	CLIP	51.8	76.8	84.3	32.7	57.7	68.2	44.1	68.2	77.0	24.7	45.1	54.6
	Direct Fine-tuning	37.4	62.3	72.1	21.8	43.4	54.5	25.7	45.8	55.4	17.9	34.5	43.1
	Long-CLIP [54]	57.6	81.1	87.8	40.4	65.8	75.2	46.8	71.4	79.8	34.1	56.3	65.7
	SmartCLIP (Ours)	**61.9**	**83.3**	**89.7**	**42.4**	**68.2**	**77.8**	**55.6**	**78.2**	**85.0**	**36.3**	**58.8**	**67.8**
L/14	CLIP	56.1	79.5	86.8	35.4	60.1	70.2	48.5	72.6	80.8	28.0	49.3	58.7
	Direct Fine-tuning	37.9	63.1	72.2	23.1	45.1	55.9	26.0	46.3	55.6	17.9	34.9	43.5
	Long-CLIP [54]	62.8	85.1	91.2	46.3	70.8	79.8	53.4	77.5	85.3	41.2	64.1	72.6
	SmartCLIP (Ours)	**66.0**	**86.2**	**92.6**	**48.5**	**73.1**	**81.7**	**63.9**	**84.4**	**90.2**	**43.8**	**66.5**	**74.8**

restrict the output to the range $(0, 1)$ and employ straight-through estimation [4] to binarize the outputs. The training batch size is 1024 and the learning rate is 10^{-6} for the CLIP component and 10^{-3} for the mask network.

Evaluation. We evaluate the following datasets:

- Long text-to-image retrieval datasets: ShareGPT4V validation split [7] and Urban1k [54]. The captions for each image are long and describe details about the image. Both datasets contain 1000 text-to-image pairs.
- Short text-to-image retrieval datasets: COCO2017 validation split [27] and Flick30K [50]. Following Long-CLIP [54], we use 30K Flickr training dataset.
- Zero-shot image classification datasets. We use benchmark datasets: Country211, Fer2013, Fgvc-aircraft, GTSRB, ImageNet, ImagetNet-V2, VOC2007, VOC2007-Multi, and SUN397. [2]

Baselines. In this paper, we benchmark our approach against CLIP [37] and the recent state-of-the-art model long-CLIP [54].

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

6.2. Comparison with CLIP Models

We present our experimental results across three key tasks: long text-to-image retrieval, short text-to-image retrieval, and zero-shot classification.

Long text-to-image retrieval. Table 2 showcases our method's performance on long text-to-image retrieval tasks. **SmartCLIP** achieves substantial improvements over baseline models, particularly the SOTA Long-CLIP, which is designed to handle long text sequences. For example, on the Urban1k dataset, **SmartCLIP** elevates the performance from 78.9% to 90.0%, marking an impressive 14% boost.

Short text-to-image retrieval. Similarly, as illustrated in Table 1, **SmartCLIP** significantly outperforms all baseline models across various metrics and datasets in short text-to-image retrieval tasks. The encouraging performance gains show that **SmartCLIP** can capture detailed information within images while simultaneously emphasizing the main semantic content.

Zero-shot classification. To comprehensively evaluate our model's capabilities, we conduct zero-shot classification benchmarks in Table 3. Both the standard CLIP model and **SmartCLIP** demonstrate superior performance on different datasets. Notably, **SmartCLIP** exhibits a slight perfor-

Table 2. **The R@1 of long-caption text-image retrieval on 1k ShareGPT4V [2] validation set and Urban-1000 dataset.** The best results are **bold**. We cite the results from Long-CLIP [54].

		ShareGPT4V		Urban1k	
		I2T	T2I	I2T	T2I
B/16	CLIP [37]	78.2	79.6	68.1	53.6
	Direct Fine-tuning	94.1	93.6	-	-
	Long-CLIP [54]	94.6	93.3	78.9	79.5
	SmartCLIP (Ours)	**98.7**	**98.1**	**90.0**	**87.4**
L/14	CLIP [37]	81.8	84.0	68.7	52.8
	Direct Fine-tuning	95.3	95.4	-	-
	Long-CLIP [54]	95.8	95.6	82.7	86.1
	SmartCLIP (Ours)	**97.9**	**98.5**	**93.0**	**90.1**

Table 3. **Zero-shot classification performance on ViT-L/14 models.** When the class name is very short, i.e., a single word like ImageNet, CLIP model perform better. When the class name is a combination of several words, our method achieves better results, e.g., the road sign in GTSRB.

Dataset	CLIP	LongCLIP	**SmartCLIP**
Country211	**31.8**	28.1	26.9
Fer2013	49.0	57.8	**58.6**
Fgvc-aircraft	**31.7**	30.6	30.4
GTSRB	50.2	48.9	**52.4**
ImageNet	**75.3**	72.9	72.5
ImageNet-V2	**69.7**	66.9	66.6
VOC2007	78.3	77.5	**78.6**
VOC2007-Multi	79.0	82.1	**83.7**
SUN397	67.5	**72.5**	72.1

This playful food sculpture transforms cucumbers into a fearsome T-Rex dinosaur. The **cucumbers form the main body**, *with whole cucumbers creating the legs and tail, while sliced cucumbers make up the creature's midsection. More cucumbers are cleverly cut to shape the dinosaur's head, and additional cucumbers are arranged to suggest muscular limbs. From its mouth erupts a dramatic spray of carrots, with finely julienned* **carrots creating the effect of fire**. *These bright orange carrots provide a stunning contrast against the green vegetables. Shredded carrots cascade downward like flames, while more carrots are delicately cut to create a flame-like texture. The carrots' vibrant color makes the dinosaur appear truly animated. Fresh celery leaves crown the creation, with celery fronds adding a decorative touch around the body. More* **celery leaves create a natural backdrop**, *while additional celery pieces add texture throughout. The celery's feathered appearance provides an artistic flourish to the overall design.*

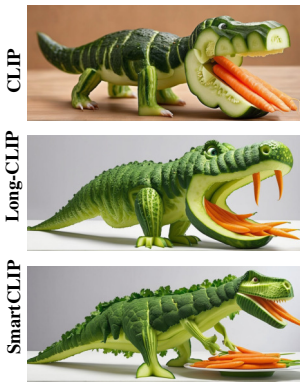

Figure 4. **Example of Long-text-to-image generation.** We replace the CLIP text encoder in SDXL with different finetuned CLIP models. Given a very long text, CLIP [37] truncates the input to 77 tokens, resulting in information loss in the image. Our model learns to generate details such as celery leaves on the back of the dinosaur while other models fail.

mance decline on ImageNet, which is expected since our model is fine-tuned on the ShareGPT4V dataset featuring long text captions, whereas ImageNet primarily consists of short, often single-word class names. However, **SmartCLIP** excels on datasets with multi-word class names, such as the GTSRB dataset, where it achieves the best performance in accurately classifying road sign descriptions.

6.3. Ablation Studies

We analyze the three key components in our model: the modular alignment module, the sparsity loss and the impact of caption diversity in the data.

Modular alignment. After introducing the mask network, we replace the standard contrastive learning with our selective alignment module. As shown in Figure 5, this change significantly improves performance. When using standard contrastive learning instead of the modular contrastive module (indicated by the purple lines), performance drops sharply. This happens because the mask information allows the network to easily separate positive pairs from negative ones, making the negative samples less informative. As a result, standard contrastive learning no longer effectively helps the model learn meaningful information.

Alignment coefficient λ_{align}. We test the impact of the alignment coefficient λ_{align}. The results in the right panels of Figure 5 show that our method performs consistently well across a wide range of λ_{align} values, from 0.1 to 20. This indicates that our approach is robust and does not require precise tuning of λ_{align} to achieve good performance.

Sparsity coefficient $\lambda_{\text{sparsity}}$. We also examine the sparsity coefficient $\lambda_{\text{sparsity}}$. The left panels of Figure 5 demonstrate that adding sparsity to the mask network improves performance. This supports our idea that promoting sparsity helps the model focus on the most relevant concepts, enhancing its ability to capture detailed information without being distracted by irrelevant details.

Caption diversity. We evaluate our model's performance under varying caption diversity conditions using the COCO dataset [27]. As shown in Table 4, increasing the number of captions per image enhances performance on the Flickr30K dataset, though at the expense of degraded performance on long-text-to-image retrieval tasks. Further improvements

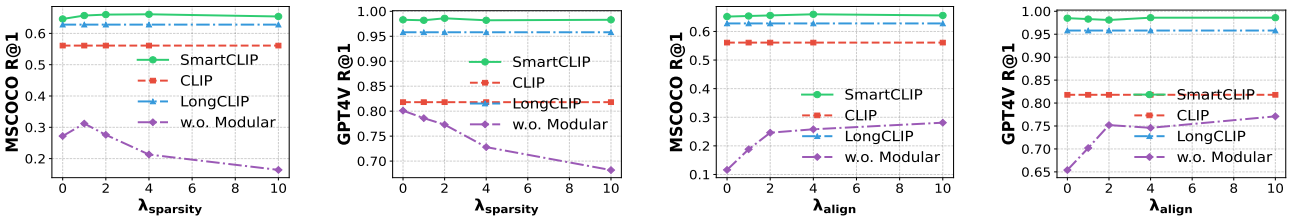

Figure 5. **Ablation Studies on two proposed modules: selective alignment and sparsity.** The baseline *w.o. Modular* means that we replace our modular alignment module with standard contrastive learning alignment.

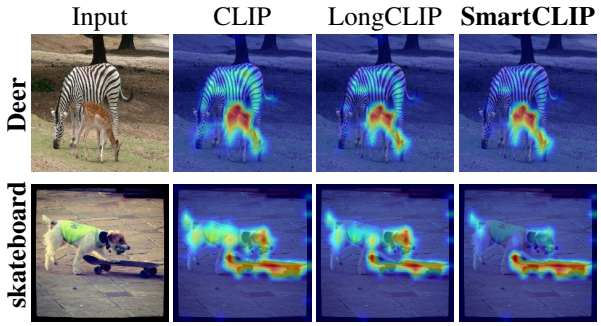

Figure 6. **Visualization of learned representations.** Given an image, we generate two captions (e.g., *a zebra* and *a zebra and a deer*) and compute cosine similarities with the same image embedding to perform binary classification for visualization using ScoreCAM [45]. Compared to baseline methods, our CLIP representations are more atomic and capture differences more effectively.

are achieved when we combine our training dataset with COCO. These results highlight both the importance of caption diversity and our method's capability to effectively handle complex text-image datasets.

6.4. Additional Results

Visualization. While our quantitative results demonstrate superior performance across various tasks, we also explore the qualitative aspects of our model by visualizing the learned representations. The visualization results are shown in Figure 6. We employ the ScoreCAM method [45] for this purpose. For each image, we generate two distinct captions, such as *"a zebra"* and *"a zebra and a deer"*. We then compute the cosine similarity between the image embedding and each of the two text embeddings. These similarity scores serve as logits for a classification task, which are then input into the ScoreCAM algorithm. **SmartCLIP** successfully learns modular representations, accurately capturing the relevant differences between captions.

Plug and play for text-to-image generation. One main advantage of **SmartCLIP** over other CLIP models trained from scratch is the low computational cost of finetuning. Additionally, our fine-tuned text encoder can replace the CLIP text encoders in large-scale models in a plug-and-play manner. Specifically, we substitute the text encoder in the SDXL [36] model with both Long-CLIP and **SmartCLIP**.

NumCapPerImg-COCO	T2I	I2T	LongT2I	LongI2T
1	53.6	39.3	85.3	89.3
3	53.6	40.9	85.3	88.6
5	56.4	41.2	85.2	86.5
Ours-ShareGPT	55.6	36.3	98.7	98.1
+COCO	57.0	38.4	97.8	98.5

Table 4. Retrieval results on Flickr30K [50] and ShareGPT-val [7] with models trained with different caption counts per image (top) and mixing long and short text-image datasets (bottom).

Table 5. **Long text to image generation performance.** We use the long captions (usually around 200 tokens, beyond the hard constraint 77 of the original CLIP model) to generate images with SDXL model [36], then we compare the generated images against the real images in ShareGPT4V validation split.

Method	KID ↓	Pr ↑	Re ↑	F1 ↑	DINO-L ↑
LongCLIP	1.05	0.238	0.768	0.363	0.401
SmartCLIP	**1.02**	**0.258**	**0.791**	**0.389**	**0.414**

As illustrated in Figure 4, our text encoder demonstrates a superior understanding of the long text, generating detailed elements such as celery leaves in the background. Furthermore, Table 5 presents quantitative results on image generation from long captions in the ShareGPT4V validation split. Our method consistently achieves better performance across various metrics, showcasing its effectiveness in handling complex, long text inputs.

7. Conclusion and Limitation

In this work, we address the information misalignment and representation entanglement issues in existing vision-language models (e.g., CLIP). We establish theoretical conditions for effectively connecting text representations to atomic-level visual features and propose **SmartCLIP**, a principled, refined vision-language model. Our experimental results validate both our theoretical results and the practical effectiveness of **SmartCLIP** in advancing multimodal learning. **Limitation.** As discussed earlier, Condition 4.2-*ii* could be violated for datasets in which a subset images are paired with a limited number of captions compared to others. In Section 4, we discuss practical strategies to mitigate such issues. Devising alternative theoretical conditions may provide additional insights into fully utilizing all pairing information, which we leave as future work.

Acknowledgement

We would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, and MBZUAI-WIS Joint Program. The work of L. Kong is supported in part by NSF DMS-2134080 through an award to Y. Chi.

References

[1] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR, 2023. 3

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023. 1

[3] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000. 3

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 5, 6

[5] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[6] Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. Enhancing multimodal understanding with clip-based image-to-text transformation. In *Proceedings of the 2023 6th International Conference on Big Data Technologies*, pages 414–418, 2023. 1

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2, 5, 6, 8

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 4, 5

[9] 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*, 2023. 3, 4

[10] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence*, pages 217–227. PMLR, 2020. 3

[12] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, et al. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024. 2

[13] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica, 2016. 4

[14] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019. 3

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[16] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 1

[17] 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*, pages 2207–2217. PMLR, 2020. 3, 4

[18] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 11455–11472. PMLR, 2022. 4

[19] Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation, 2023. 4

[20] Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Juan Lao Tebar, Wenze Hu, Zhe Gan, Peter Grasch, et al. Revisit large-scale image-caption data in pre-training multimodal foundation models. *arXiv preprint arXiv:2410.02740*, 2024. 2

[21] Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. Veclip: Improving clip training via visual-enriched captions. *ECCV. IEEE*, page 13, 2024. 2, 4

[22] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*, 2024. 2

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

tional conference on machine learning, pages 12888–12900. PMLR, 2022. 2, 4

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[25] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024. 2, 4

[26] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6, 7

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[29] AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*, 2024. 2

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[31] Hiroshi Morioka and Aapo Hyvarinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *International conference on artificial intelligence and statistics*, pages 3399–3426. PMLR, 2023. 3

[32] Hiroshi Morioka and Aapo Hyvarinen. Causal representation learning made identifiable by grouping of observational variables. In *Forty-first International Conference on Machine Learning*, 2024. 3

[33] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 2

[34] Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M Asano, Nanne van Noord, Marcel Worring, and Cees GM Snoek. Tulip: Token-length upgraded clip. *arXiv preprint arXiv:2410.10034*, 2024. 2

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 4, 5

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 8

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4, 5, 6, 7

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

[39] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 1

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[41] Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning, 2023. 3

[42] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 1

[43] 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. *arXiv preprint arXiv:2106.04619*, 2021. 3, 4, 5

[44] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[45] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 8

[46] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 1

[47] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 4, 5

[48] Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. Lotlip: Improving language-image pre-training for long text understanding. *arXiv preprint arXiv:2410.05249*, 2024. 2

[49] Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4, 5

[50] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6, 8

[51] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[52] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022. 2

[53] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 2

[54] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024. 2, 4, 5, 6, 7

[55] Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[56] Jihai Zhang, Xiaoye Qu, Tong Zhu, and Yu Cheng. Clip-moe: Towards building mixture of experts for clip with diversified multiplet upcycling. *arXiv preprint arXiv:2409.19291*, 2024. 2

[57] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024. 3

[58] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022. 1

[59] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. *arXiv preprint arXiv:2403.17007*, 2024. 2