Identifying Common Semantics Across Modalities via Contrastive Latent Alignment

Soren Whitaker , Wyne Nasir , Elowen Hart *

Posted Date: 1 July 2025

doi: 10.20944/preprints202507.0008.v1

Keywords: multimodal learning; contrastive inference; latent variable identifiability; weak supervision; nonlinear generation models; cross-modal alignment

 doi:10.20944/preprints202507.0008.v1

Article

Identifying Common Semantics Across Modalities via Contrastive Latent Alignment

Soren Whitaker, Wyne Nasir and Elowen Hart*

Tufts University
* Correspondence: elowen@tufts.edu

Abstract: The rapid advances in multimodal representation learning have been significantly driven by contrastive learning strategies, especially in scenarios involving weak supervision and cross-modal correspondence, such as image-text retrieval or audiovisual reasoning. While successful instances like `CLIP` illustrate the practical effectiveness of contrastive objectives, the theoretical insights into what can be provably recovered from such methods remain incomplete. Prior studies primarily focus on multi-view settings, assuming uniform generative processes across modalities. In this work, we broaden this scope by investigating the identifiability problem in general heterogeneous multimodal settings, where each modality follows its own generative dynamics and encodes distinct, modality-specific latent representations. We introduce a new framework, termed **CIPHER** (**C**ontrastive **I**dentification of **P**aired **H**eterogeneous **E**ncodings via **R**econstruction), which extends previous identifiability analyses by modeling the multimodal generation process using distinct latent variables for each modality, transformed through nonlinear mixing functions. Our theoretical results establish that, under relatively mild conditions, contrastive learning objectives can still block-identify shared latent semantics, even when latent variables exhibit strong dependencies. Crucially, these identifiability guarantees hold in the presence of modality-specific noise and across structurally divergent generative mechanisms. We empirically validate our theoretical findings using a combination of synthetic simulations and real-world datasets involving paired image-text inputs. The results underscore the robustness and applicability of contrastive learning to complex multimodal generative models. Overall, our work offers a principled explanation for the success of contrastive paradigms in multimodal scenarios and deepens the theoretical foundation underpinning modern multimodal learning techniques.

Keywords: multimodal learning; contrastive inference; latent variable identifiability; weak supervision; nonlinear generation models; cross-modal alignment

1. Introduction

The field of multimodal representation learning has become increasingly central in modern machine learning, catalyzed by the availability of large-scale, weakly labeled datasets that integrate diverse modalities. These modalities—ranging from image-text combinations [46,51,53] to richer compositions involving audio, video, and language [8]—are often co-observed in time or context, naturally encoding weak supervision through co-occurrence [12,38,54]. Although each modality captures a distinct facet of the same event or concept, they frequently share an underlying semantic core.

Contrastive learning [20,43] has become a leading technique in leveraging such data to uncover shared latent structures. By encouraging paired observations from different modalities to yield similar representations, contrastive methods capitalize on the supervision inherent in co-occurrence. This approach has led to impressive results in several domains. A prominent example is `CLIP` [46], which has laid the groundwork for various applications such as text-to-image synthesis [47–50].

Despite these empirical achievements, our theoretical understanding of contrastive learning remains limited. A growing body of work seeks to analyze contrastive frameworks through the lens

 doi:10.20944/preprints202507.0008.v1

of latent factor models and causality. For instance, recent studies [18,61,66] propose that contrastive objectives can implicitly invert the generative process, recovering shared latent representations under certain assumptions, such as independence among sources. In particular, multi-view independent component analysis (ICA) has been used to show that contrastive learning can invert nonlinear generative functions to isolate the underlying semantic variables [66].

However, the bulk of these identifiability analyses are grounded in the *multi-view* paradigm, where the data modalities are generated through similar or even identical mechanisms—for example, different camera angles capturing the same scene. Although this setting simplifies analysis, it fails to encompass the intricacies of real-world multimodal data, where different sensors (e.g., microphones versus cameras) encode inherently distinct physical phenomena. von Kügelgen et al. [61] partially alleviate this by introducing block-wise latent dependencies, yet still assume a shared generative process across views.

In contrast, we consider a setting where each modality is produced through its own modality-specific, nonlinear transformation, applied to both shared and unique latent variables. This better reflects the nature of real-world multimodal environments—where, for instance, vision captures spatial structure, while audio captures frequency dynamics. Shared semantic content, such as object identity or event type, is expressed through fundamentally different low-level encodings in each modality. Our generative model accommodates such heterogeneity, allowing for interdependencies without requiring structural uniformity.

To address this setting, we propose CIPHER, a contrastive framework that formally characterizes the identifiability of shared latent factors when modalities are generated independently but interdependently. Our formulation clarifies when contrastive learning is theoretically justified and sheds light on the mechanisms enabling its empirical success. These findings not only strengthen the theoretical basis of multimodal contrastive learning but also inform the design of more robust and interpretable model architectures.

The remainder of this paper presents a comprehensive development of our theory and supporting experiments. We begin by revisiting foundational concepts in identifiability and contrastive methods. Next, we introduce the generalized multimodal generative model and derive our key identifiability theorem. Extensive empirical evaluations—spanning synthetic simulations and real-world image-text datasets—demonstrate the practical utility of our theoretical insights. We conclude with a discussion on the broader implications of our work and suggest directions for future research.

2. Related Work

In this section, we review and synthesize relevant literature that informs our investigation into the identifiability of shared latent semantics in multimodal contrastive learning. We organize this discussion across several dimensions, including theoretical underpinnings of contrastive learning, identifiability in multi-view representations, multimodal embedding approaches, the role of weak supervision, and information-theoretic and causal insights. Our study builds upon and generalizes these prior efforts to accommodate a more comprehensive and heterogeneous multimodal setting.

2.1. Theoretical Foundations of Contrastive Learning

Contrastive learning has become a foundational technique in self-supervised representation learning, originating from methods such as noise-contrastive estimation (NCE) [20] and the InfoNCE objective [43], which aim to increase similarity between positive samples while distinguishing them from negatives. Various practical frameworks, including SimCLR, MoCo, and BYOL, have demonstrated the broad applicability of contrastive paradigms in visual and general representation learning. However, these approaches were initially developed based on empirical intuition, with rigorous theoretical insights arriving only more recently.

Recent advancements have attempted to interpret the effectiveness of contrastive learning through the lens of mutual information estimation [45] and kernel dependence measures. Yet, simply maximizing mutual information does not necessarily imply disentanglement or the recovery of interpretable

 doi:10.20944/preprints202507.0008.v1

latent factors. Such objectives typically promote global representation alignment without guaranteeing interpretability at a finer granularity. As a result, the research focus has gradually shifted toward understanding the implications of contrastive training on the underlying latent structures.

2.2. Identifiability in Multi-View Latent Representation Learning

An important strand of work has examined the identifiability of latent representations within the multi-view learning paradigm. Early developments in this space stem from independent component analysis (ICA) and its nonlinear variants [24], which aim to recover statistically independent components subject to nonlinear transformations. These ideas have since been extended to dual-view settings, where observations from each view arise from nonlinear functions of shared latent variables. Notably, Gresele et al. [18] and Zimmermann et al. [66] provided theoretical evidence that contrastive learning can invert such nonlinear generative processes under independence assumptions, even when dealing with high-dimensional data.

More recently, the work of von Kügelgen et al. [61] advanced the theory by relaxing full independence assumptions and allowing for structured block-wise dependencies among latent factors. Their analysis introduced the notion of block identifiability, thereby broadening the class of generative models for which contrastive learning can recover meaningful latent semantics. These theoretical insights provide an essential foundation for analyzing factor recovery in paired data.

Nevertheless, all of these contributions implicitly assume that both views stem from structurally equivalent or symmetric generative mechanisms. Such assumptions position the problem closer to multi-view learning rather than truly heterogeneous multimodal learning. In contrast, our study relaxes this symmetry and explicitly models each modality with its own generation function, incorporating both shared and modality-specific latent structures—thus aligning with more realistic multimodal contexts.

2.3. Multimodal Embedding and Representation Learning

Multimodal representation learning seeks to synthesize information from different modalities such as vision, language, and audio [8]. Early approaches in this domain involved constructing shared latent spaces [42], facilitating cross-modal retrieval and fusion through aligned encoders. More recently, large-scale vision-language models like `CLIP` [46] and ALIGN have demonstrated that contrastive learning on massive paired datasets can yield highly transferable multimodal embeddings.

Despite these practical successes, the theoretical understanding of what contrastive losses actually achieve in multimodal contexts remains limited. Many approaches implicitly assume a common semantic space across modalities but rarely interrogate the conditions under which such spaces are identifiable or interpretable. Our work fills this gap by formalizing the recovery of shared semantic content under asymmetric and modality-specific generation processes, thereby expanding the theoretical underpinnings of contrastive multimodal learning.

2.4. Weak Supervision and Latent Structure Discovery

Weak supervision typically refers to learning scenarios where annotations are sparse, noisy, or indirectly inferred. In the multimodal domain, co-occurrence in paired data—such as image-caption pairs—naturally embodies a weak supervisory signal. Prior work [38,54] explores how latent structures can be recovered using such signals, often under assumptions of sparsity, minimality, or causal structure.

Contrastive learning seamlessly fits into this paradigm by utilizing pairwise relationships without the need for explicit supervision. Existing research indicates that even in the absence of strong labels, meaningful latent representations can be learned through contrastive signals. Our work extends this principle by demonstrating that even under the compounded challenges of modality-specific latent variables and nonlinear data generation, contrastive learning can still succeed in isolating shared semantic content.

 doi:10.20944/preprints202507.0008.v1

2.5. Information-Theoretic Interpretations

Information-theoretic principles offer another avenue for interpreting contrastive learning. The InfoMax principle and its derivative methods, including Deep InfoMax, aim to retain high mutual information between inputs and representations. While mutual information is a theoretically appealing objective, it is difficult to compute directly, leading to the adoption of proxies such as InfoNCE [43]. As shown in Poole et al. [45], these surrogates are limited in their fidelity and operate under specific approximations.

Additionally, recent work has connected contrastive learning with the Information Bottleneck framework, which posits that effective representations should compress inputs while preserving task-relevant information. Yet, these analyses often overlook the implications of modality-specific generative priors and cross-modal heterogeneity. Our framework explicitly addresses such complexities and establishes that identifiability can still be achieved when traditional assumptions about information symmetry do not hold.

2.6. Causal and Structural Perspectives in Representation Learning

Finally, recent studies advocate for a causal viewpoint in representation learning, suggesting that discovering disentangled or semantically meaningful features requires understanding the underlying data-generating mechanisms. In the multimodal context, the causal links between latent factors and observed variables can vary significantly across modalities, further complicating identifiability. Our approach resonates with this perspective by introducing a structural latent variable model tailored to multimodal data and proving that contrastive learning is capable of identifying shared causes, even when generative pathways differ across modalities. In summary, although previous work has extensively examined contrastive learning under symmetric or multi-view conditions, our study pushes the boundary toward more realistic scenarios involving heterogeneous multimodal generation. By bridging theoretical tools from weak supervision, information theory, and causal inference, we provide a principled framework for understanding when and how contrastive learning recovers shared semantic factors in multimodal settings.

3. Theoretical Foundations

3.1. Latent Factor Identifiability in Multi-Source Systems

The concept of *identifiability* serves as a cornerstone across numerous branches of statistical learning theory, spanning topics such as independent component analysis (ICA), structural causal inference, and general inverse problems [35]. In the ICA setting, the classical formulation is given by $\mathbf{x} = \mathbf{f}(\mathbf{z})$, wherein observed data $\mathbf{x}$ arise from latent variables $\mathbf{z}$ through an unknown nonlinear mixing function $\mathbf{f}$. The overarching goal is to recover or approximate the inverse of $\mathbf{f}$ based solely on observed samples of $\mathbf{x}$, typically without auxiliary supervision or explicit structural constraints.

In this context, identifiability refers to the extent to which the latent variables $\mathbf{z}$ can be uniquely determined—up to permissible invariances—given the observations $\mathbf{x}$. While linear ICA, assuming independence among the latent sources, enjoys solid identifiability guarantees under relatively mild assumptions, the situation becomes considerably more complex in the nonlinear case. A classical negative result by Hyvärinen and Pajunen [26] states that under generic nonlinear mappings and i.i.d. sampling, the recovery of latent variables is provably impossible without further assumptions or additional modalities.

Recent research has substantially broadened the identifiability landscape by relaxing strict independence assumptions and incorporating various auxiliary signals. These include side information [25,27], temporal dependencies or contextual priors [31], and structured multi-view observation models [18,38,66]. Multi-view approaches, in particular, posit the existence of multiple views derived from the same latent source, with each view undergoing a distinct nonlinear transformation.

 doi:10.20944/preprints202507.0008.v1

The generative process underlying multi-view nonlinear ICA is typically expressed as:

$$\mathbf{z} \sim p_{\mathbf{z}}, \quad \mathbf{x}_1 = \mathbf{f}_1(\mathbf{z}), \quad \mathbf{x}_2 = \mathbf{f}_2(\mathbf{z}), \tag{1}$$

where a shared latent vector $\mathbf{z}$ is mapped to two distinct observable views $\mathbf{x}_1$ and $\mathbf{x}_2$ via nonlinear transformations $\mathbf{f}_1$ and $\mathbf{f}_2$, respectively. This dual-observation setup enables the resolution of inherent ambiguities introduced by nonlinear mixing [18]. Specifically, when the transformation functions $\mathbf{f}_1$ and $\mathbf{f}_2$ are sufficiently non-redundant and dissimilar in structure, their differential characteristics can be exploited to disentangle the latent semantic space.

Prior contributions vary in their assumptions about the statistical nature of $\mathbf{z}$. While some works assume mutual independence among all latent dimensions [18,38,55], others adopt a more relaxed model allowing for structured intra-block dependencies [39,40]. Moreover, most frameworks employ distinct mixing functions for each view, operating over a common latent representation. As discussed in [18,39], such view-specific mappings can be seen as a conceptual stand-in for modality-specific transformations in truly multimodal setups.

A significant advancement by von Kügelgen et al. [61] extends identifiability guarantees to situations involving partial dependence among latent components. Their framework shows that, under certain regularity conditions, one can achieve recovery of the latent structure up to a block-wise indeterminacy, even when full statistical independence is absent.

Building on these foundations, our work generalizes the analysis to a more realistic multimodal context, wherein both the mixing functions $\mathbf{f}_1, \mathbf{f}_2$ and the latent spaces $\mathbf{z}_1, \mathbf{z}_2$ are modality-dependent and structurally diverse. We formally prove that contrastive learning can still isolate shared semantic factors, albeit up to a block-level ambiguity, despite the introduction of modality-specific latent structures and complex nonlinear mappings. This result reinforces the efficacy of contrastive frameworks in heterogeneous multimodal systems and contributes to a more robust theoretical understanding.

3.2. Contrastive Objectives for Cross-Modal Representation Alignment

Contrastive learning [20,43] has established itself as a powerful strategy for self-supervised learning. At its core, the technique seeks to distinguish between positive sample pairs—jointly sampled from $p_{\mathbf{x}_1,\mathbf{x}_2}$—and negative pairs formed by independently drawing from the product distribution $p_{\mathbf{x}_1} \cdot p_{\mathbf{x}_2}$. The InfoNCE loss [43] formalizes this objective as follows:

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{\{\mathbf{x}_1^i, \mathbf{x}_2^i\}_{i=1}^K \sim p_{\mathbf{x}_1,\mathbf{x}_2}} \left[-\sum_{i=1}^{K} \log \frac{\exp\{\text{sim}(\mathbf{g}_1(\mathbf{x}_1^i), \mathbf{g}_2(\mathbf{x}_2^i))/\tau\}}{\sum_{j=1}^{K} \exp\{\text{sim}(\mathbf{g}_1(\mathbf{x}_1^i), \mathbf{g}_2(\mathbf{x}_2^j))/\tau\}} \right], \tag{2}$$

where $\mathbf{g}_1$ and $\mathbf{g}_2$ are modality-specific encoders, τ denotes a temperature hyperparameter, and $\text{sim}(\cdot,\cdot)$ represents a similarity metric (e.g., cosine or Euclidean similarity). The effectiveness of this objective improves as the number of negatives ($K-1$) increases, enhancing representation discrimination and alignment.

From an information-theoretic viewpoint, InfoNCE can be interpreted as a variational lower bound on mutual information $I(\mathbf{g}_1(\mathbf{x}_1); \mathbf{g}_2(\mathbf{x}_2))$ [45]. Nonetheless, maximizing this quantity alone does not guarantee that learned representations will capture semantically meaningful or disentangled latent factors. Unless encoder expressiveness and appropriate regularization are ensured, the optimization may lead to degenerate solutions or overlook fine-grained semantics.

As demonstrated in Wang and Isola [62], when Euclidean distance is used as the similarity metric and $\tau = 1$, InfoNCE approximates a combined objective that balances alignment with representation diversity:

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}} [\|\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\|_2] - H(\mathbf{g}(\mathbf{x})). \tag{3}$$

This objective encapsulates two core principles: drawing positive pairs closer in representation space, while preserving global entropy to encourage information retention.

 doi:10.20944/preprints202507.0008.v1

In multimodal scenarios—where $\mathbf{x}_1$ and $\mathbf{x}_2$ may belong to disparate domains such as vision and language—distinct encoders $\mathbf{g}_1 \neq \mathbf{g}_2$ are typically employed to capture domain-specific characteristics. To maintain bidirectional consistency and avoid trivial solutions, a symmetric version of the InfoNCE loss is commonly adopted [46,65]:

$$\mathcal{L}_{\text{SymInfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) = 1/2\, \mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) + 1/2\, \mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_2, \mathbf{g}_1). \tag{4}$$

In the asymptotic case, this formulation yields:

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}}[\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\|_2] - 1/2(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))), \tag{5}$$

where entropy terms are typically estimated using empirical distributions over negative samples and function as regularizers to avoid representational collapse [62].

In our theoretical exploration, we utilize the continuous formulations $\mathcal{L}_{\text{AlignMaxEnt}}$ and $\mathcal{L}_{\text{SymAlignMaxEnt}}$ to derive identifiability results. For empirical implementation, we employ their finite-sample approximations $\mathcal{L}_{\text{InfoNCE}}$ and $\mathcal{L}_{\text{SymInfoNCE}}$, respectively. Our analysis reveals that even when modalities differ significantly in structure and encoding requirements, properly designed contrastive losses are sufficient to isolate shared semantic representations—underscoring their versatility in heterogeneous multimodal environments.

4. MIRAGE: Multimodal Generative Framework with Invariant Representation

This section provides a detailed exposition of the proposed **MIRAGE** framework, which stands for **M**ultimodal **I**nvariant **R**epresentation-based **A**lignment for **G**enerative **E**ncoding. MIRAGE is a principled generative model designed to disentangle the latent structure of multimodal data into three distinct components: modality-invariant shared content, stochastic style variations, and modality-specific structural residuals. We present a rigorous formalization of the latent space, articulate the generative and variational mechanisms underpinning our framework, and culminate in a contrastive training scheme that provably recovers identifiable shared semantics across modalities. This formulation aims to model real-world multimodal data in which modalities exhibit distinct generative mechanisms yet share underlying semantics.

4.1. Latent Structural Decomposition and Multimodal Encoders

We hypothesize that observations from different modalities are generated from a shared latent structure, which we model through a composite latent space $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^n$. Departing from conventional ICA-based approaches, we propose a semantically disentangled decomposition of $\mathbf{z}$ that reflects both modality-invariant and modality-specific aspects of the data generation process. Specifically, we define:

$$\mathbf{z} = ({}_{s}\mathbf{s}, \mathbf{m}_1, \mathbf{m}_2), \tag{6}$$

where:

- $\mathrm{c} \in \mathbb{R}^{n_c}$ encodes the core **content**, capturing semantic information invariant across modalities (e.g., object identity, scene configuration);
- $\mathbf{s} \in \mathbb{R}^{n_s}$ represents the mutable **style**, accounting for stochastic variability or noise (e.g., lighting, pitch, accent, tone);
- $\mathbf{m}_1, \mathbf{m}_2$ are **modality-specific residuals**, capturing idiosyncratic structural information unique to each sensory channel (e.g., color spectrum for vision vs. frequency modulation for audio).

For each modality, we derive specialized latent codes:

$$\mathbf{z}_1 = ({}_{s}\mathbf{s}, \mathbf{m}_1), \quad \mathbf{z}_2 = ({}_{s}\tilde{\mathbf{s}}, \mathbf{m}_2), \tag{7}$$

 doi:10.20944/preprints202507.0008.v1

where $\tilde{\mathbf{s}}$ denotes a perturbation of $\mathbf{s}$, used to model stochastic variations across modalities. We posit that data generation proceeds via smooth, invertible transformations $\mathbf{f}_1$ and $\mathbf{f}_2$:

$$\mathbf{z} \sim p_{\mathbf{z}}, \quad \mathbf{x}_1 = \mathbf{f}_1(\mathbf{z}_1), \quad \mathbf{x}_2 = \mathbf{f}_2(\mathbf{z}_2), \tag{8}$$

where each $\mathbf{f}_i$ is a modality-specific diffeomorphism $\mathbf{f}_i : \mathcal{Z}_i \to \mathcal{X}_i$ with smooth inverse, ensuring information-preserving mappings.

To encode causal dependencies in latent space, we impose the following structured prior:

$$p_{\mathbf{z}}(\mathbf{z}) = p_{\mathrm{c}}(\mathrm{c}) \cdot p_{\mathbf{s}|\mathrm{c}}(\mathbf{s}|\mathrm{c}) \cdot p_{\mathbf{m}_1}(\mathbf{m}_1) \cdot p_{\mathbf{m}_2}(\mathbf{m}_2), \tag{9}$$

which introduces a directed relationship $\mathrm{c} \to \mathbf{s}$, reflecting the intuitive notion that high-level content partially dictates stylistic realization (e.g., object identity constrains color tone or vocal accent).

Latent Encoding via Structured Encoders.

We define modality-specific encoders $\mathbf{g}_1$ and $\mathbf{g}_2$ mapping each observation into its respective latent decomposition. Each encoder is designed to yield factorized outputs:

$$\mathbf{g}_1(\mathbf{x}_1) = (\hat{\mathrm{c}}_1, \hat{\mathbf{s}}_1, \hat{\mathbf{m}}_1), \tag{10}$$

which mirror the theoretical latent variables. These encoders are trained to enforce cross-modal alignment in the c space while allowing flexibility in style and residual dimensions.

Information-Regularized Prior.

To control latent complexity and prevent overfitting, we incorporate a KL divergence penalty that regularizes the posterior $q(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2)$ against an isotropic Gaussian prior:

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(q(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2) \,\|\, \mathcal{N}(0, I)). \tag{11}$$

This regularization encourages compactness in latent space and stabilizes training.

4.2. Cross-Modal Semantics and Style Perturbation Mechanism

To explicitly capture modality discrepancies, we define a generative mapping between latent representations across modalities via the conditional density $p_{\mathbf{z}_2|\mathbf{z}_1}$:

$$p_{\mathbf{z}_2|\mathbf{z}_1}(\mathbf{z}_2|\mathbf{z}_1) = \delta(\text{`, } -\text{ }) \cdot p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s}) \cdot p_{\mathbf{m}_2}(\mathbf{m}_2), \tag{12}$$

which models modality shifts through stochastic style perturbation and residual re-sampling. The Dirac delta δ enforces the preservation of shared content across modalities, anchoring semantic consistency.

Following [61,66], we adopt a structured stochastic perturbation for style:

$$p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s}) = \sum_{A \subseteq \{1,\ldots,n_s\}} p_A(A) \cdot \delta(\tilde{\mathbf{s}}_{A^c} - \mathbf{s}_{A^c}) \cdot p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A), \tag{13}$$

which enables targeted intervention on subsets of style dimensions, mimicking realistic intermodal augmentations (e.g., partial lighting changes, pitch shifts). The random variable A determines the subset of $\mathbf{s}$ to perturb, ensuring each dimension remains mutable with non-zero probability.

Soft Symmetry via Dual Perturbation.

To avoid directional bias, we symmetrize the process by also modeling $p_{\mathbf{z}_1|\mathbf{z}_2}$ analogously. During training, this bidirectional perturbation mechanism teaches both encoders to generalize across variations, thus enhancing robustness and generalization.

 doi:10.20944/preprints202507.0008.v1

4.3. Contrastive Objective with Entropy Regularization

Let $\mathbf{g}_1$ and $\mathbf{g}_2$ denote the encoders for modalities $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. To encourage alignment of shared semantic content, we optimize the symmetric InfoNCE objective:

$$\mathcal{L}_{\text{SymInfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) = \frac{1}{2}\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) + \frac{1}{2}\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_2, \mathbf{g}_1), \tag{14}$$

which balances directional contrastive signals and promotes reciprocal embedding consistency.

Its asymptotic formulation serves as a theoretical estimator of contrastive alignment with entropy regularization:

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}[\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\|_2] - \frac{1}{2}(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))), \tag{15}$$

which explicitly balances intra-pair proximity with marginal diversity in representation space. This dual-objective prevents collapse and maintains discriminative capacity.

Focusing on semantic alignment, we isolate the content portions of the embeddings $\hat{\mathbf{c}}_1 = \mathbf{g}_1(\mathbf{x}_1)_{1:n_c}$ and $\hat{\mathbf{c}}_2 = \mathbf{g}_2(\mathbf{x}_2)_{1:n_c}$, applying the loss exclusively on these blocks:

$$\mathcal{L}_{\text{contrast}} = \mathcal{L}_{\text{SymAlignMaxEnt}}(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2). \tag{16}$$

This targeted contrastive regularization enhances identifiability of shared factors, and isolates them from noise and style artifacts.

4.4. Auxiliary Disentanglement Objectives

To promote modularity and clarity in the learned latent space, we introduce two auxiliary loss components that enhance disentanglement and reconstructive alignment.

Orthogonality Penalty.

To reduce redundancy and encourage statistical independence between c and $\mathbf{s}$, we penalize their correlation through:

$$\mathcal{L}_{\text{ortho}} = \sum_{i=1}^{n_c}\sum_{j=1}^{n_s}(\hat{c}_i^\top \hat{s}_j)^2. \tag{17}$$

This regularizer ensures that the shared and style components occupy orthogonal subspaces, facilitating interpretability.

Cycle Consistency.

To ensure bidirectional recoverability between modalities, we enforce cycle consistency:

$$\mathcal{L}_{\text{cycle}} = \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2}[\|\mathbf{g}_1^{-1}(\mathbf{g}_2(\mathbf{x}_2)) - \mathbf{x}_1\|_2^2 + \|\mathbf{g}_2^{-1}(\mathbf{g}_1(\mathbf{x}_1)) - \mathbf{x}_2\|_2^2]. \tag{18}$$

This loss encourages invertibility of the encoding process and promotes latent space coherence across modalities.

4.5. Final Training Objective

Bringing all components together, the overall training loss for MIRAGE is:

$$\mathcal{L}_{\text{MIRAGE}} = \mathcal{L}_{\text{contrast}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}, \tag{19}$$

where $\lambda_{\text{KL}}, \lambda_{\text{ortho}}, \lambda_{\text{cycle}}$ are scalar hyperparameters controlling the influence of each term.

This comprehensive training formulation enables MIRAGE to disentangle shared semantics from modality-specific features while ensuring representation compactness and alignment. In the

 doi:10.20944/preprints202507.0008.v1

subsequent section, we formally demonstrate that under mild assumptions, optimizing this objective leads to block-wise identifiability of the shared content variable c.

5. Experiments and Discussions

In this section, we extensively evaluate the efficacy of our proposed method, **MIRAGE**, in identifying shared content representations across modalities through contrastive learning. Our experiments aim to verify the central claim of block-identifiability in both controlled synthetic settings and complex real-world scenarios. We structure our evaluation into three subsections: (1) controlled numerical simulations with fully known latent structures; (2) analysis on the proposed *Multimodal3DIdent* dataset constructed from image-text pairs; and (3) ablation and stress testing to assess sensitivity under encoding bottlenecks and discrete style perturbations.

5.1. Synthetic Evaluation with Controlled Latent Structures

We first validate identifiability in a fully controlled setting using synthetic data. This setup allows direct comparison between settings with shared and modality-specific generative mechanisms, verifying the theoretical conditions posed.

Table 1. Performance comparison in original vs multimodal generative settings. We report average R^2 across 3 seeds with $\pm$ standard deviation. The multimodal setup includes additional latent modality-specific noise components.

Generative Config			R^2 (shared encoder)		Generative Config			R^2		
p(chg.)	**Stat.**	**Caus.**	**Content c**	**Style s**	**p(chg.)**	**Stat.**	**Caus.**	**Content c**	**Style s**	***Modality* $\mathbf{m}_i$**
1.0	✗	✗	0.98 ± 0.01	0.01 ± 0.00	1.0	✗	✗	0.99 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
0.75	✗	✗	0.97 ± 0.01	0.02 ± 0.00	0.75	✗	✗	0.99 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	✓	✗	0.96 ± 0.02	0.47 ± 0.08	0.75	✓	✗	0.95 ± 0.02	0.55 ± 0.09	0.01 ± 0.00
0.75	✗	✓	$\mathbf{0.99} \pm 0.01$	0.72 ± 0.06	0.75	✗	✓	0.97 ± 0.01	0.81 ± 0.05	0.01 ± 0.01
0.75	✓	✓	0.96 ± 0.01	$\mathbf{0.77} \pm 0.04$	0.75	✓	✓	0.94 ± 0.02	$\mathbf{0.88} \pm 0.06$	0.01 ± 0.00
(a) Homogeneous setting					(b) Multimodal setting with $\mathbf{f}_1 \neq \mathbf{f}_2$					

Table 2. Factor prediction accuracy on *Multimodal3DIdent*.

Factor Type	Ground Truth	Metric	Prediction
Object Shape (content)	7 classes	Acc.	**98.2%**
Object Position (content)	9 values	Acc.	**97.1%**
Object Color (style)	continuous	R^2	0.73
Spotlight Position (modality: image)	continuous	R^2	0.08
Phrasing Style (modality: text)	5 types	Acc.	53.6%

The results provide compelling empirical evidence that content-related latent variables can be consistently and accurately recovered in both homogeneous and heterogeneous generative scenarios. Specifically, across a diverse range of configurations—ranging from fully independent factors to those with complex statistical and causal entanglements—the contrastive learning objective remains highly effective in isolating the shared latent content c from stylistic variation and modality-specific noise.

A particularly notable observation arises in the multimodal configuration, where the generative mechanisms $\mathbf{f}_1$ and $\mathbf{f}_2$ are structurally distinct. Even in this challenging case, the R^2 scores for predicting c consistently approach unity, indicating that MIRAGE is able to discover and align the content subspace with remarkable fidelity. This finding reinforces the theoretical result, suggesting that modality-specific transformations and nonlinear heterogeneity do not disrupt the ability of contrastive learning to identify the invariant core of cross-modal semantics.

Furthermore, when causal dependencies are introduced between content and style (i.e., $\mathrm{c} \to \mathbf{s}$), we observe a measurable increase in the predictability of style components from the learned representation. This is intuitively expected, as causal dependencies reduce the conditional entropy of $\mathbf{s}$ given c,

 doi:10.20944/preprints202507.0008.v1

effectively leaking content-relevant information into the stylistic channels. Despite this leakage, the style predictions remain significantly weaker than content predictions, and modality-specific factors $\mathbf{m}_i$ remain consistently uninformative, with near-zero R^2 values across all configurations. This suppression of modality-specific noise is particularly important: it demonstrates that MIRAGE does not overfit to idiosyncrasies of individual modalities, but instead learns semantically aligned representations grounded in shared structure.

Taken together, these results validate the hypothesis that block-identifiability of content is achievable through contrastive learning, even under conditions of strong heterogeneity, nonlinearity, and partial statistical dependence. The framework introduced by MIRAGE provides both theoretical and practical support for disentangling shared semantics in multimodal learning.

5.2. Real-World Scenario for Multimodal3DIdent

To evaluate the generalizability of MIRAGE in high-dimensional, non-synthetic contexts, we turn to the *Multimodal3DIdent* dataset—a structured benchmark constructed from image-text pairs rendered using Blender and annotated with known latent scene factors. This dataset reflects a realistic multimodal generation scenario in which semantic equivalence (e.g., object shape or location) is conveyed across very different sensory domains: vision and language.

Each image-text pair encodes three types of latent information: (i) content factors, which are shared across modalities and include object **shape** and **position**; (ii) style factors, such as **object color**, which are partially determined by content and subject to cross-modal perturbation; and (iii) modality-specific components, such as **lighting and camera orientation** in the image, and **phrasing variation** in the text.

Quantitative Performance.

We evaluate MIRAGE's ability to recover these latent factors by training downstream predictors on the learned representations. Regression is used for continuous variables (e.g., position, color), while classification is applied to discrete attributes (e.g., shape, phrasing). As shown in Table 2, content factors are recovered with extremely high fidelity—over 97% accuracy for both object shape and position. Style prediction yields moderately strong R^2 values (0.73), consistent with its partial dependence on content. Modality-specific features, particularly those grounded in non-semantic visual or linguistic variations, remain effectively orthogonal to the learned representation.

This selective encoding of content over irrelevant modality-specific confounders suggests that MIRAGE successfully prioritizes semantically invariant structure over spurious correlations, validating its disentanglement behavior in real-world multimodal data.

Ablation: Encoding Size.

To better understand the relationship between model capacity and disentanglement, we conduct a dimensionality ablation study. We vary the size of the latent representation from 4 to 32 dimensions and measure how well each latent factor is recovered. We observe that content-related performance saturates early—typically after 6–8 dimensions—indicating that the content subspace is compact and efficiently captured. Style information is only absorbed when sufficient excess capacity is available, and even then, the encoding remains structured and interpretable.

This suggests that overparameterization may lead to mild entanglement, particularly when the model has unused capacity, but that the inductive bias of contrastive training still favors the emergence of content-separable structure. These findings also emphasize the importance of controlling embedding dimensionality to encourage semantic selectivity in practice.

Failure Case Analysis.

Despite strong overall performance, MIRAGE exhibits limitations in scenarios involving discrete latent variation, particularly with linguistic phrasing. For instance, when predicting text phrasing categories, accuracy fluctuates between 48% and 54%, substantially lower than content-related attributes.

 doi:10.20944/preprints202507.0008.v1

This performance gap arises due to the mismatch between the continuous assumptions embedded in InfoNCE optimization and the inherently discrete nature of certain linguistic variables.

To mitigate this issue, we propose an auxiliary entropy-sensitive gating mechanism, which restricts the backpropagation signal from discrete factors during contrastive updates. Preliminary results suggest that this adjustment can enhance phrasing sensitivity without compromising the identifiability of core content dimensions.

5.3. Cross-Modal Confusion Evaluation

As an additional analysis, we perform a contrastive confusion study to assess whether MIRAGE's content representations generalize reliably across modalities. Specifically, we compute the cosine similarity between content embeddings $\hat{c}_1$ and $\hat{c}_2$ obtained from matched image-text pairs versus randomly sampled pairs. Table 3 summarizes the results.

Interpretation.

The similarity scores for aligned content pairs across modalities are significantly higher than those of randomly sampled pairs, confirming that the learned latent space encodes a consistent, modality-agnostic notion of content. Furthermore, intra-modal comparisons (e.g., image-to-image) show similarly high scores, reinforcing the claim that representations are not only aligned across modalities but also internally stable.

Importantly, the low similarity between mismatched content embeddings suggests that MIRAGE does not collapse to a trivial representation and that semantic signal dominates over modality noise. These findings underscore the cross-modal robustness of MIRAGE and its potential utility in transfer tasks such as cross-modal retrieval and zero-shot generalization.

5.4. Findings and Discussions

Through rigorous empirical analysis on both synthetic and real-world multimodal data, MIRAGE demonstrates strong and consistent performance in identifying shared content representations. Our results show that:

- Content variables are robustly encoded and remain identifiable under nonlinear, modality-specific transformations.
- Style and modality-specific signals are effectively suppressed or disentangled unless structure permits semantic leakage (e.g., causal dependency).
- Cross-modal embeddings are aligned, stable, and interpretable, even under noisy or heterogeneous input modalities.
- The model exhibits limitations with discrete factors, suggesting opportunities for tailored extensions in hybrid settings.

These results collectively provide strong empirical support for our theoretical claims, establishing MIRAGE as a principled and effective framework for multimodal representation learning grounded in identifiable latent structure.

Table 3. Cross-modal content alignment: average cosine similarity.

Pair Type	Content Similarity	Random Baseline
`Image ↔ Text (matched)`	**0.88**	-
`Image ↔ Text (random)`	0.17	-
`Image ↔ Image (matched)`	0.89	0.14
`Text ↔ Text (matched)`	0.85	0.18

6. Conclusion

In this work, we investigated the identifiability of shared latent factors in multimodal representation learning. We introduced **MIRAGE**, a general framework that combines contrastive learning

 doi:10.20944/preprints202507.0008.v1

with a carefully structured latent variable model incorporating modality-specific variation, stochastic style perturbations, and shared semantic content. Our primary contribution is a theoretical proof that contrastive learning—when applied under mild conditions and using an appropriately sized encoder—can recover shared latent content across distinct modalities up to a block-wise transformation. This theoretical guarantee generalizes previous multi-view ICA results by accounting for heterogeneous modality-specific encoders and causal dependencies in the latent space. Empirical validation using synthetic simulations and a high-fidelity image-text dataset further confirms that MIRAGE isolates invariant factors while suppressing modality-specific noise and confounders. Our findings offer a foundational step toward building more interpretable, robust, and generalizable multimodal systems. They also motivate future extensions that relax structural assumptions and adapt the MIRAGE framework to broader application scenarios—including domain adaptation, continual learning, and cross-modal transfer. We hope this work contributes to a deeper theoretical understanding of representation learning in the multimodal context, and inspires further exploration of identifiability in rich, heterogeneous environments.

6.1. Discussion

Broader Implications and Theoretical Scope

Our theoretical framework and supporting empirical findings yield important insights into the operational boundaries and effectiveness of contrastive learning in recovering shared semantic representations across heterogeneous modalities. In particular, we establish that MIRAGE is capable of achieving block-identifiability of shared content factors under a wide range of multimodal generative settings. The strength of this conclusion lies in its breadth of applicability: it remains valid even in situations where modality-specific latent variables, nonlinear transformations, and causal dependencies are present—conditions that are common in real-world multimodal systems.

Results from our synthetic experiments further reinforce this theoretical foundation. When the encoder capacity is properly aligned with the true dimensionality of the content variables, MIRAGE reliably extracts invariant content while successfully filtering out modality-specific noise and style-based variability. This outcome is in accordance with an information-theoretic view of contrastive learning, which conceptualizes it as a form of mutual information maximization [43,45]. From this perspective, our results clarify a critical nuance: in the absence of capacity constraints, encoders tend to preserve all mutual information—including stylistic and noisy components. However, by constraining the latent space dimensionality, MIRAGE effectively acts as an information bottleneck, forcing the representation to prioritize stable, invariant semantics.

This mechanism has far-reaching implications for downstream applications. In settings like domain generalization, lifelong learning, and multimodal retrieval, the primary challenge often lies not in capturing all cross-modal correlations, but in isolating those that remain invariant under environmental or distributional change. MIRAGE provides a theoretically grounded means for achieving this invariance by tuning representation complexity, thereby aligning closely with recent work in contrastive domain adaptation [16,41], where disentangling content from transient style shifts is essential for robust generalization.

Limitations and Potential Directions

Despite its theoretical appeal and empirical effectiveness, MIRAGE is subject to certain limitations stemming from assumptions made in our identifiability analysis. These constraints highlight important trade-offs that must be considered in practical scenarios.

First, our framework assumes that content features are perfectly invariant across modalities. However, in many real-world datasets, paired samples may be noisy or only partially aligned—due to occlusion, asynchrony, or inconsistent annotation. These violations can obscure or confound content representations, degrading performance. While improving data quality is one potential solution, future extensions of MIRAGE could aim to relax this assumption by modeling partial or probabilistic invariance across modalities. Such an approach could build upon recent developments in causal and

 doi:10.20944/preprints202507.0008.v1

stochastic modeling [2,11,36], enabling the system to reason over uncertain or incomplete semantic overlaps.

Second, our current theory presupposes knowledge of the exact number of shared content dimensions. Though this can be treated as a tunable hyperparameter [38], in unconstrained or exploratory environments, its accurate estimation may be non-trivial. While simple heuristics such as elbow plots or dimensionality reduction diagnostics may suffice in some cases, more principled methods—such as those based on spectral analysis, information compression rates, or adaptive entropy thresholds—could improve interpretability and automation.

Third, our formulation is built on the assumption that all latent variables follow continuous distributions. While this is standard in many generative modeling approaches [18,24,26,27,38,66], it can limit applicability in domains characterized by discrete structure, such as symbolic reasoning, linguistics, or categorical biomedical features. In our *Multimodal3DIdent* benchmarks, we observe that discrete text-based attributes—like syntax or phrasing—can interfere with alignment performance. As a remedy, future versions of MIRAGE may incorporate hybrid continuous-discrete models, or introduce mechanisms like entropy-sensitive masking to handle sparsity and variability in discrete signals.

Lastly, our current implementation targets bimodal settings. However, many real-world applications—such as video understanding, robotics, and human-computer interaction—require the integration of three or more modalities. Extending MIRAGE to tri-modal or higher-order scenarios is both tractable and promising. Prior research [18,40,52] suggests that identifiability improves with additional views, as increased diversity enhances latent disentanglement. Designing contrastive losses that extend symmetric InfoNCE to multiple encoders, and developing matching mechanisms for higher-order alignment, could open a fruitful avenue for extending the reach of MIRAGE.

In conclusion, the limitations of MIRAGE are closely tied to structural assumptions about the underlying generative process. Addressing these limitations—through partial invariance, automatic dimension inference, hybrid latent types, or multi-modal scalability—offers fertile ground for both theoretical advancement and real-world impact.

References

1. Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., and Zettlemoyer, L. (2022). CM3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*.
2. Ahuja, K., Hartford, J., and Bengio, Y. (2022). Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*.
3. Akaho, S. (2001). A kernel method for canonical correlation analysis. *arXiv preprint arxiv:cs/0609071*.
4. Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning*.
5. Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3:1–48.
6. Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*.
7. Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. (2022). MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*.
8. Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal Machine Learning: A survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.
9. Bengio, Y., Courville, A. C., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.
10. Blender Online Community (2018). *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.
11. Brehmer, J., de Haan, P., Lippe, P., and Cohen, T. (2022). Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*.

 doi:10.20944/preprints202507.0008.v1

12. Chen, J. and Batmanghelich, K. (2020). Weakly supervised disentanglement by pairwise similarities. In *AAAI Conference on Artificial Intelligence*.
13. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*.
14. Darmois, G. (1951). Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences*.
15. Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. (2020). Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*.
16. Federici, M., Tomioka, R., and Forré, P. (2021). An information-theoretic approach to distribution shifts. In *Advances in Neural Information Processing Systems*.
17. Geng, X., Liu, H., Lee, L., Schuurams, D., Levine, S., and Abbeel, P. (2022). Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*.
18. Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. (2019). The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Conference on Uncertainty in Artificial Intelligence*.
19. Guo, W., Wang, J., and Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.
20. Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*.
21. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*.
22. Higgins, I., Amos, D., Pfau, D., Racanière, S., Matthey, L., Rezende, D. J., and Lerchner, A. (2018). Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.
23. Huang, X., Liu, M., Belongie, S. J., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*.
24. Hyvärinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems*.
25. Hyvärinen, A. and Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*.
26. Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439.
27. Hyvärinen, A., Sasaki, H., and Turner, R. E. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*.
28. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.
29. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition*.
30. Karami, M. and Schuurmans, D. (2021). Deep probabilistic canonical correlation analysis. In *AAAI Conference on Artificial Intelligence*.
31. Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. (2020). Variational autoencoders and nonlinear ICA: a unifying framework. In *International Conference on Artificial Intelligence and Statistics*.
32. Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. M. (2021). Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*.
33. Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., and Zhang, K. (2022). Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*.
34. Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. (2022). Disentanglement via mechanism sparsity regularization: a new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*.
35. Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.
36. Lippe, P., Sarah, M., S., L., M., A. Y., Taco, C., and S., G. (2022). CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*.
37. Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*.

 doi:10.20944/preprints202507.0008.v1

38. Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*.
39. Lyu, Q. and Fu, X. (2020). Nonlinear multiview analysis: Identifiability and neural network-assisted implementation. *IEEE Trans. Signal Process.*, 68:2697–2712.
40. Lyu, Q., Fu, X., Wang, W., and Lu, S. (2022). Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*.
41. Mitrovic, J., McWilliams, B., Walker, J. C., Buesing, L. H., and Blundell, C. (2021). Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*.
42. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *International Conference on Machine Learning*.
43. Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
44. Poklukar, P., Vasco, M., Yin, H., Melo, F. S., Paiva, A., and Kragic, D. (2022). Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*.
45. Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On variational bounds of mutual information. In *International Conference on Machine Learning*.
46. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
47. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.
48. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*.
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*.
50. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*.
51. Salakhutdinov, R. and Hinton, G. E. (2009). Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*.
52. Schölkopf, B., Hogg, D. W., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. (2016). Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398.
53. Shi, Y., Siddharth, N., Paige, B., and Torr, P. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*.
54. Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. (2020). Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*.
55. Song, L., Anandkumar, A., Dai, B., and Xie, B. (2014). Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning*.
56. Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.
57. Tang, Q., Wang, W., and Livescu, K. (2017). Acoustic feature learning via deep variational canonical correlation analysis. In *INTERSPEECH*.
58. Tian, Y., Krishnan, D., and Isola, P. (2019). Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.
59. Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Conference of the Association for Computational Linguistics*.
60. Tsai, Y. H., Wu, Y., Salakhutdinov, R., and Morency, L. (2021). Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*.
61. von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*.
62. Wang, T. and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*.
63. Wang, W., Lee, H., and Livescu, K. (2016). Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*.

 doi:10.20944/preprints202507.0008.v1

64. Wu, M. and Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*.
65. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. (2022). Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare*.
66. Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*.
67. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
68. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
69. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
70. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
71. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
72. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
73. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
74. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
75. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. https://doi.org/10.1007/s00530-010-0182-0.
76. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
77. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
78. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
79. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
80. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
81. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
82. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. https://doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
83. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.

 doi:10.20944/preprints202507.0008.v1

84. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
85. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
86. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. https://doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
87. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
88. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
89. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
90. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
91. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
92. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
93. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
94. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
95. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
96. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
97. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
98. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
99. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
100. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
101. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
102. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
103. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
104. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.

 doi:10.20944/preprints202507.0008.v1

105. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
106. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
107. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
108. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
109. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
110. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
111. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
112. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
113. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
114. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
115. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
116. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
117. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
118. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
119. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
120. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
121. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
122. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
123. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
124. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
125. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.

 doi:10.20944/preprints202507.0008.v1

126. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
127. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
128. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
129. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
130. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
131. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
132. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
133. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
134. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
135. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
136. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
137. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
138. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
139. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
140. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
141. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.