Emergent Semantics from Disjoint Modalities: Unsupervised Cross-Domain Vision-Language Grounding

Milan Janssens *, Noor Peeters , Callum Hensley

Posted Date: 11 September 2025

doi: 10.20944/preprints202509.0958.v1

Keywords: multimodal learning; unsupervised pre-training; vision-and-language models; cross-modal alignment; masked prediction; object tags; self-supervised learning

 doi:10.20944/preprints202509.0958.v1

Article

Emergent Semantics from Disjoint Modalities: Unsupervised Cross-Domain Vision-Language Grounding

Milan Janssens *, Noor Peeters and Callum Hensley

University of Antwerp, Belgium
* Correspondence: milan.janssens@uantwerpen.be

Abstract

The rapid evolution of multimodal representation learning has yielded increasingly powerful vision-and-language (V&L) systems, achieving remarkable success across diverse downstream tasks. Yet, most current solutions are fundamentally constrained by their dependence on large-scale parallel corpora, where each image is paired with a manually curated caption. The construction of such datasets is not only resource-intensive but also ill-suited for domain-specific or low-resource scenarios. In this study, we present **VISTRA** (**VIS**ion-Text Representation Alignment), a new paradigm for V&L pre-training that circumvents the need for explicitly aligned data. Drawing inspiration from research in unsupervised machine translation and multilingual embedding learning, VISTRA integrates a dual-modality masked reconstruction mechanism with semantic anchors extracted from object detection pipelines. These anchors function as modality-agnostic pivots, enabling implicit cross-modal grounding even in the absence of direct correspondence. Our experiments on four widely adopted English benchmarks demonstrate that VISTRA consistently matches, and in certain cases surpasses, the performance of supervised counterparts trained with aligned image-caption pairs. Beyond its empirical competitiveness, our approach exposes the latent geometric structure of multimodal spaces, revealing that disjoint corpora can, with the aid of semantic anchoring, support effective representation alignment. This work therefore not only reduces reliance on costly annotation, but also highlights the feasibility of constructing scalable and transferable V&L models from unpaired multimodal resources.

Keywords: multimodal learning; unsupervised pre-training; vision-and-language models; cross-modal alignment; masked prediction; object tags; self-supervised learning

---

1. Introduction

In recent years, the unification of visual and linguistic modalities has profoundly reshaped the research landscape of artificial intelligence. The capacity to integrate heterogeneous streams of information has enabled transformative advances in tasks such as visual question answering, multimodal retrieval, referring expression resolution, and grounded caption generation. Central to this transformation is the advent of large-scale pre-trained vision-and-language (V&L) models, including ViLBERT Lu et al. (2019), LXMERT, VisualBERT Li et al. (2019), VLBERT, and UNITER Chen et al. (2020c), which rely on transformer-based architectures to produce joint embeddings. These joint embeddings allow visual features and textual tokens to be projected into a unified semantic space, thereby supporting cross-modal reasoning and generation at an unprecedented level of accuracy. As a result, these models dominate benchmarks and have become the de facto standard in modern multimodal research.

Yet, despite the compelling progress, a critical weakness continues to constrain the scalability and generalizability of current V&L frameworks. Unlike their unimodal counterparts such as BERT Devlin et al. (2019), which thrive on vast amounts of raw and uncurated textual data, multimodal models

 doi:10.20944/preprints202509.0958.v1

remain highly dependent on large-scale parallel corpora consisting of aligned image-caption pairs. Constructing these corpora is costly, requiring extensive human annotation or laborious filtering. For example, MS-COCO Chen et al. (2015), a widely used benchmark, demands manual captioning for hundreds of thousands of images. Similarly, Conceptual Captions Changpinyo et al. (2021), although sourced from the web, involved intensive filtering procedures, retaining only a few million pairs from billions of crawled candidates. Such dependency on curated pairs fundamentally limits the scalability of multimodal models to domains where aligned resources are scarce, such as medical imaging or low-resource languages, and thereby undermines the vision of truly universal multimodal intelligence.

This observation raises a fundamental question: must effective multimodal representations always be grounded in parallel supervision, or can cross-modal structure emerge even in the absence of alignment? In this work, we argue for the latter. Specifically, we address the task of unsupervised pre-training for vision-and-language models, where the learning signal is derived exclusively from disjoint corpora of images and texts. Our formulation represents a paradigm shift: rather than relying on costly annotations to align modalities, we investigate whether latent semantic structures in each modality can be independently acquired and subsequently reconciled through implicit anchors. This perspective resonates with broader trends in machine learning that aim to reduce the dependence on direct supervision and instead harness the abundance of unannotated data naturally available at scale.

The intellectual roots of this idea lie in several adjacent research domains. A prominent parallel can be drawn with the development of unsupervised machine translation Lample et al. (2018), which demonstrated that bilingual alignment can emerge without parallel corpora, provided that monolingual data is sufficiently exploited. Similarly, work on unsupervised caption generation Feng et al. (2019) shows that linguistic regularities can bootstrap image-to-text mappings even without aligned supervision. These findings collectively suggest that semantic alignment is not solely a product of annotated data, but can also emerge as an emergent property of distributional regularities across domains. Extending this intuition to multimodality, we hypothesize that the structures governing visual and textual data exhibit a degree of latent compatibility, which can be uncovered through carefully designed unsupervised objectives.

Further inspiration is drawn from multilingual representation learning Pires et al. (2019), where models like mBERT Devlin et al. (2019b) are trained on heterogeneous language corpora without relying on direct translations. Remarkably, these models achieve semantically coherent embeddings across languages purely through masked token prediction. Subsequent studies Conneau et al. (2020) reveal that these embeddings, despite lacking explicit cross-lingual supervision, are linearly mappable into a shared semantic space, exposing universal regularities in human language. By analogy, if we treat images as compositions of discrete visual tokens Dosovitskiy et al. (2020), multimodal learning can be reinterpreted through the lens of multilingual learning: modalities are analogous to languages, and cross-modal alignment can arise from shared semantic pivots. This analogy guides our central hypothesis that vision and language, although semantically distinct, may converge in representation space if provided with a sufficient set of semantic anchors.

Building upon this hypothesis, we propose **VISTRA** (**VIS**ion-Text Representation Alignment), an unsupervised framework that discards the need for parallel image-caption pairs altogether. At the core of VISTRA is a dual-modality masked prediction scheme, which extends the architecture of VisualBERT Li et al. (2019) into a setting where modalities are processed independently. During training, the model encounters either textual documents or visual inputs, never paired simultaneously. For texts, randomly masked tokens are reconstructed; for images, masked regions are predicted in terms of semantic content or positional structure. The critical innovation lies in the incorporation of semantic anchors extracted from object detectors Li et al. (2020b). These anchors serve as pseudo-shared vocabulary items that recur across modalities—for instance, an object tag such as "*cake*" may appear both as a detected region label in an image and as a noun in text. Through such anchors, VISTRA implicitly aligns visual and linguistic signals in a shared embedding space without requiring explicit supervision.

 doi:10.20944/preprints202509.0958.v1

The practical implications of this approach are manifold. Consider real-world applications where parallel multimodal datasets are prohibitively expensive or unavailable. In social media analysis, for instance, detecting hateful memes Kiela et al. (2020) often involves mismatched modalities, where text and image are loosely related or not captioned at all. In medical contexts, image-text pairs such as radiology images and diagnostic reports Li et al. (2020c) are scarce and domain-specific, making supervised pre-training infeasible. In such cases, unsupervised frameworks like VISTRA offer a scalable alternative, leveraging abundant unpaired data while still yielding transferable multimodal representations.

To validate this paradigm, we conduct extensive experiments across multiple benchmarks, including Visual Question Answering (VQA) Goyal et al. (2017), Natural Language for Visual Reasoning (NLVR2), Flickr30K image retrieval, and RefCOCO+. To mimic the absence of parallel corpora, we deliberately discard alignment information from standard datasets, simulating a genuinely unsupervised environment. Remarkably, VISTRA achieves competitive performance relative to supervised baselines, even rivaling models trained with full access to paired data. Furthermore, when pre-trained on entirely disjoint sources—such as large-scale web images and BookCorpus—VISTRA continues to deliver robust performance, underscoring the feasibility of learning from unpaired multimodal resources.

Finally, we complement these findings with detailed ablation studies to dissect the contribution of semantic anchors. Quantitative results reveal that object tags substantially enhance cross-modal alignment, while qualitative case analyses illustrate how anchors foster emergent grounding between vision and language. We also explore a semi-supervised extension, demonstrating that integrating a small fraction of aligned data with abundant unpaired corpora not only preserves performance but also introduces regularization benefits and semantic diversity. Collectively, these results challenge the entrenched view that parallel datasets are indispensable for V&L pre-training, and instead highlight unsupervised learning as a promising frontier for building scalable, general-purpose multimodal systems.

2. Related Work

Pre-trained V&L Transformers

One of the most dominant research directions in multimodal learning centers on large-scale pre-trained vision-and-language (V&L) transformers, most of which adopt the "mask-and-predict" strategy. Within this paradigm, models are optimized to recover hidden visual regions or textual tokens based on their surrounding context, thereby encouraging the fusion of heterogeneous information sources Chen et al. (2020c); Gan et al. (2020); Huang et al. (2020); Li et al. (2020a, 2019); Lu et al. (2019). A key characteristic of these frameworks is their reliance on parallel image-text datasets, which serve as the foundation for learning fine-grained associations. Architecturally, two major paradigms have emerged. In the *dual-stream* family, visual features and textual embeddings are encoded separately through modality-specific transformers before being merged by a dedicated fusion layer. ViLBERT Lu et al. (2019), LXMERT, and ERNIE-ViL are well-known representatives of this line. By contrast, the *single-stream* design concatenates both modalities from the start and feeds them directly into a unified transformer, as exemplified by VisualBERT Li et al. (2019), VL-BERT, and UNITER Chen et al. (2020c). Both paradigms have established strong baselines across benchmarks such as VQA, retrieval, and multimodal reasoning Kiela et al. (2020), and probing analyses Cao et al. (2020) confirm their ability to capture subtle inter-modal dependencies and contextual relations after pre-training.

A number of extensions have further enriched these transformer backbones by introducing auxiliary semantic signals. A notable example is Oscar Li et al. (2020b), which injects object tags as additional tokens into the input, thereby enhancing visual-linguistic grounding under the assumption of well-annotated image-caption data. Our approach draws inspiration from this idea but advances it to an unsupervised regime, where object tags function not as aligned supervision but as indirect semantic anchors linking disjoint corpora. In this way, tags acquire heightened importance because

 doi:10.20944/preprints202509.0958.v1

they supply the only cross-modal bridge when paired supervision is unavailable. VIVO Hu et al. (2020) also leverages curated image-tag pairs for novel object captioning, combining them later with supervised datasets. By contrast, our method eliminates manual annotation entirely; tags are extracted automatically by frozen detection models, making the pipeline more scalable and less prone to domain-specific overfitting.

Self-supervised Representation Learning

Self-supervised learning (SSL) has emerged as a unifying principle across domains, where useful representations are distilled from unlabeled data through proxy prediction tasks. In natural language processing, pioneering models such as ELMo Peters et al. (2018), BERT Devlin et al. (2019), and RoBERTa Liu et al. (2019) demonstrated that masked language modeling provides robust contextual encoders transferable to a variety of downstream applications. In the visual domain, early SSL approaches were largely based on low-level pretext tasks, such as predicting image color channels, ordering shuffled patches, or reconstructing spatial contexts Doersch et al. (2015); Noroozi and Favaro (2016); Pathak et al. (2016). More recently, contrastive learning frameworks like SimCLR Chen et al. (2020b) and MoCo have shown that maximizing agreement between augmented views yields state-of-the-art visual encoders.

In line with this trajectory, our framework formulates multimodal pre-training as a self-supervised problem over unimodal datasets. Specifically, masked token modeling is independently applied to large corpora of language and image data, without requiring any explicit alignment. Unlike traditional SSL in vision, which often emphasizes pixel-level patterns, we base our visual encoder on object-centric features derived from a pre-trained detector, thus shifting the focus toward higher-level semantics. Conceptually, our design represents a hybrid SSL scheme: the masked modeling idea is shared across both modalities, while semantic anchors guide the implicit cross-modal alignment. This differs from most previous V&L SSL approaches, which generally assume some form of weak alignment such as noisy captions or paired meta-information.

Unsupervised Multilingual Language Models

Another key source of inspiration for our framework is the progress in unsupervised multilingual language modeling. Models such as mBERT Devlin et al. (2019b) have convincingly demonstrated that cross-lingual representations can be learned in the absence of parallel translation pairs. These models, trained over diverse monolingual corpora with a shared encoder and vocabulary, display strong zero-shot and few-shot transfer capabilities across languages. Empirical studies Conneau et al. (2020) further reveal that shared tokens—numbers, named entities, and universal symbols—play an outsized role in facilitating transfer, while architectural features like shared positional embeddings also enhance cross-lingual alignment.

Our proposed method extends these principles to the multimodal domain. We adopt a unified transformer that processes both visual and linguistic sequences, and introduce semantic anchors, derived from object detection, as cross-modal equivalents of overlapping vocabulary. These anchors act as bridge tokens, making it possible to align otherwise unpaired modalities. The analogy between multilingual transfer and multimodal alignment has been highlighted in recent efforts to establish universal representational spaces across domains Chen et al. (2020a). Building upon this insight, our framework introduces a concrete, unsupervised bridging mechanism that plays the same role as shared lexical tokens in multilingual corpora.

Unsupervised Grounding and Captioning

A related area of investigation concerns unsupervised grounding and caption generation. The overarching goal in these works is to discover correspondences between visual regions and linguistic expressions without paired annotation. Earlier methods relied on co-occurrence statistics or weak signals to link textual phrases with visual entities. More recent works in unsupervised captioning Feng et al. (2019); Gu et al. (2019); Laina et al. (2019) adopt generative objectives, combining separately

 doi:10.20944/preprints202509.0958.v1

collected image and text corpora through adversarial learning, reinforcement-based optimization, or back-translation strategies. These models attempt to generate plausible captions for images even in the absence of aligned training pairs.

While conceptually relevant, our approach departs in two important ways. First, rather than producing captions, our objective is to construct aligned multimodal representations, which can subsequently serve multiple downstream applications beyond captioning. Second, unlike captioning pipelines that still depend on sentence-level descriptive corpora, we broaden the textual modality to include unconstrained, domain-general sources such as books and articles. This flexibility demonstrates that representation alignment can emerge even when the textual data is not curated to describe images, thereby increasing the applicability of our framework in real-world large-scale repositories.

Cross-modal Alignment without Pairing

Recent research has started to question the necessity of strong alignment for multimodal training. CLIP-style frameworks, for instance, rely on large-scale web-harvested image-text pairs where alignment is weak or noisy. These methods confirm that loose supervision, when combined with scale, is sufficient to learn transferable representations. However, they still presuppose some form of co-occurrence between images and texts. Our approach is more radical: we remove the assumption of co-occurrence entirely, and instead rely on a small set of semantically meaningful anchors to achieve alignment. This property makes our framework particularly relevant for specialized domains, including medical imaging or security-critical environments, where images are sensitive and textual annotations are infeasible. In such contexts, pairing is either impossible or prohibitively expensive, underscoring the unique suitability of our method.

Multimodal Learning under Resource Constraints

An additional dimension of our contribution lies in addressing multimodal learning for under-resourced contexts. Collecting annotated multimodal corpora is a privilege available only in a handful of high-resource languages and domains. For the majority of communities, especially those involving low-resource languages or culturally specific imagery, parallel datasets are absent. This inequity motivates research into models that can generalize from heterogeneous, unpaired corpora. By demonstrating that our framework performs competitively under such conditions, we provide evidence that resource-efficient multimodal pre-training is achievable, offering a more equitable path for future AI systems that must operate globally across diverse cultural and linguistic settings.

Vision-Language Models with Weak Supervision

Finally, several approaches attempt to relax the dependence on strong supervision by incorporating semi-supervised or weakly supervised signals. These include methods that bootstrap alignment with pseudo-labeled captions, utilize noisy alt-text from web or social media as supervisory signals, or apply contrastive consistency across modalities. Our model differs in that it bypasses alignment altogether and instead introduces structured semantic cues in the form of detector-generated tags. This shift in perspective opens up new possibilities for combining pre-existing visual priors (e.g., object graphs, region-based detectors) with transformer-based multimodal pre-training, thereby offering a scalable alternative that retains efficiency while eliminating annotation overhead.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2025 doi:10.20944/preprints202509.0958.v1

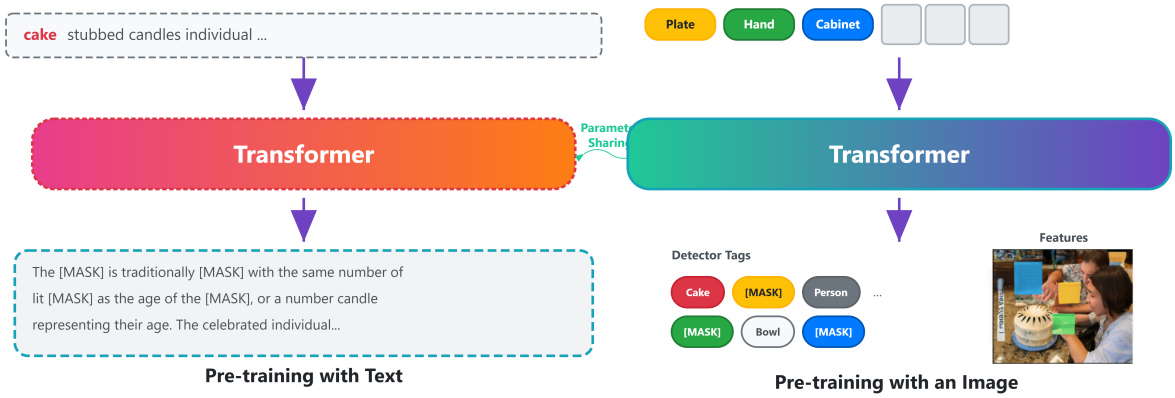

Figure 1. An overview of pre-training using unpaired data.

3. Proposed Framework

In this section, we begin by revisiting the conventional recipe of supervised vision-and-language (V&L) pre-training, using a modified VisualBERT as an illustrative baseline. This background serves to highlight the dependency of existing approaches on paired data. We then introduce our unsupervised framework, **VISTRA**, which is designed to overcome this dependency by exploiting two non-overlapping corpora of texts and images. The framework introduces several innovations: (1) a dual-stream masking and reconstruction paradigm, (2) the use of detector-based tags as semantic anchors, (3) a unified representation space that integrates both modalities with shared embeddings, and (4) an additional alignment regularizer that enforces semantic consistency across heterogeneous signals.

3.1. Supervised Pre-Training with Aligned Pairs

The majority of early V&L pre-training schemes are grounded in the assumption that training data consist of paired image-caption examples. To make this concrete, we adopt VisualBERT as a reference point and construct a supervised baseline, denoted as SUPERVISED-VB, which is also evaluated in our experiments. Specifically, SUPERVISED-VB extends VisualBERT by combining three elements: (i) feature regression and classification losses over detected regions, inspired by LXMERT; (ii) reconstruction of detector-derived object tags following Oscar Li et al. (2020b); and (iii) an image-text matching signal. The model accepts an aligned input pair (T, I), where T denotes a tokenized sentence and I represents the image, decomposed into m region vectors and l detector tags. Textual tokens are embedded through word, segment, and positional encodings, while visual inputs are represented by region-level features with associated spatial encodings.

Formally, an aligned pair (T, I) is encoded as:

$$\boldsymbol{T} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n], \quad \boldsymbol{I} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m; \boldsymbol{d}_1, \ldots, \boldsymbol{d}_l], \tag{1}$$

where $\boldsymbol{w}_i$ is the embedding of the i-th subword, $\boldsymbol{r}_j$ is the feature vector of region j with positional encoding, and $\boldsymbol{d}_k$ corresponds to the embedding of the k-th detector tag. The concatenated sequence $[\boldsymbol{T}; \boldsymbol{I}]$ is passed to a transformer encoder f_θ, which produces contextualized multimodal representations.

Training objectives in this supervised setting combine four terms:

$$\mathcal{L}_{\text{supervised}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{V-Pred}} + \mathcal{L}_{\text{Tag-Recon}} + \mathcal{L}_{\text{Match}}. \tag{2}$$

Here, $\mathcal{L}_{\text{MLM}}$ masks and predicts language tokens, $\mathcal{L}_{\text{V-Pred}}$ covers region regression and classification, $\mathcal{L}_{\text{Tag-Recon}}$ reconstructs masked tags, and $\mathcal{L}_{\text{Match}}$ ensures correct image-text alignment.

The visual prediction component itself is decomposed as:

$$\mathcal{L}_{\text{V-Pred}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}}, \tag{3}$$

 doi:10.20944/preprints202509.0958.v1

Algorithm 1: Unsupervised Pre-training Procedure of **VISTRA**

Input: Unaligned text corpus D_T, unaligned image corpus D_I, pretrained object detector $\mathcal{D}$
Output: Pre-trained multimodal transformer f_θ
Initialize parameters θ of transformer f_θ and projection head g_ϕ;
Initialize shared subword vocabulary $\mathcal{V}$ and embedding matrices;
for *each training step* **do**
  if *Randomly sample modality* == ***text*** **then**
    Sample mini-batch of text sequences $\{T_i\}_{i=1}^B \sim D_T$;
    Apply random masking to obtain $\{\tilde{T}_i\}_{i=1}^B$;
    Encode using transformer: $\boldsymbol{h}_i = f_\theta(\tilde{T}_i)$;
    Compute masked language modeling loss: $\mathcal{L}_{\text{MLM}} = \frac{1}{B}\sum_i \text{MLM}(\boldsymbol{h}_i, T_i)$;
  else
    Sample mini-batch of images $\{I_j\}_{j=1}^B \sim D_I$;
    foreach I_j **do**
      Use detector $\mathcal{D}$ to extract object regions $\{b_k\}$ and tags $\{t_k\}$;
      Construct input sequence: $S_j = [\boldsymbol{r}_{1:m}; \boldsymbol{d}_{1:l}]$;
      Apply masking: $\tilde{S}_j = \text{Mask}(S_j)$;
      Encode with transformer: $\boldsymbol{z}_j = f_\theta(\tilde{S}_j)$;
      Compute region feature loss $\mathcal{L}_{\text{reg}}$, tag prediction loss $\mathcal{L}_{\text{Tag-Recon}}$, and contrastive loss $\mathcal{L}_{\text{align}}$ for region-tag pairs;
    end
    Aggregate visual losses:

$$\mathcal{L}_{\text{V-Pred}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}}, \quad \mathcal{L}_{\text{unsup}} = \mathcal{L}_{\text{V-Pred}} + \mathcal{L}_{\text{Tag-Recon}} + \lambda \cdot \mathcal{L}_{\text{align}}$$

  end
  Update model parameters θ, ϕ using gradient descent;
end
return f_θ

where $\mathcal{L}_{\text{reg}} = \frac{1}{m}\sum_{j=1}^m \|\boldsymbol{r}_j - \hat{\boldsymbol{r}}_j\|_2^2$ enforces regression to the detector features, and $\mathcal{L}_{\text{cls}}$ is a categorical prediction loss over region labels. In aggregate, SUPERVISED-VB exemplifies the strengths but also the limitations of supervised multimodal pre-training: it achieves strong downstream results but relies critically on costly paired data.

3.2. Unsupervised Pre-Training on Disjoint Modalities

To address the bottleneck of annotation, we propose **VISTRA**, which completely discards the assumption of aligned supervision. The framework is premised on the hypothesis that latent alignment can emerge from independent corpora if the model is properly regularized with semantic anchors and shared representations. VISTRA comprises four interlocking modules: (1) dual-stream masked modeling for intra-modality learning, (2) detector-based tag anchoring, (3) unified multimodal embeddings with shared vocabulary, and (4) contrastive alignment for fine-grained consistency.

3.2.1. Dual-Stream Masked Modeling

Unlike supervised approaches, VISTRA learns from two separate data sources: a textual corpus D_T (e.g., Wikipedia, BookCorpus) and an image corpus D_I (e.g., OpenImages). At each step, the model samples from one corpus, masks part of the input, and predicts the masked content using the same transformer encoder f_θ. For a text sequence T, a subset of tokens is masked to form $\tilde{T}$, and for an image I, regions or tags are masked to form $\tilde{I}$.

Formally:

$$\tilde{T} = \text{Mask}(T), \quad \tilde{I} = \text{Mask}(I), \tag{4}$$

 doi:10.20944/preprints202509.0958.v1

where masked tokens or visual elements are replaced by a `[MASK]` symbol. The shared transformer f_θ is then trained to reconstruct the missing components. The overall unsupervised objective is:

$$\mathcal{L}_{\text{unsup}} = \mathbb{E}_{T\sim D_T}\mathcal{L}_{\text{MLM}}(f_\theta(\tilde{T}), T) + \mathbb{E}_{I\sim D_I}\left[\mathcal{L}_{\text{V-Pred}}(f_\theta(\tilde{I}), I) + \mathcal{L}_{\text{Tag-Recon}}\right]. \tag{5}$$

This design enables the transformer to share parameters across both text-only and image-only inputs, ensuring that representational patterns learned in one modality inform the other. Such parameter tying implicitly encourages structural regularities to be discovered and reused across domains.

3.2.2. Semantic Anchoring via Detector Tags

To provide a bridge between modalities, VISTRA incorporates semantic anchors derived from object detection models. Let $\mathcal{D}$ denote a frozen detector that maps image I to a set of bounding box–tag pairs:

$$\mathcal{D}(I) = \{(b_j, t_j)\}_{j=1}^{m}, \tag{6}$$

where b_j is the bounding box and t_j is the textual label. Each tag token is embedded as:

$$\boldsymbol{d}_j^{(i)} = \boldsymbol{e}(t_j^{(i)}) + \boldsymbol{p}(b_j), \tag{7}$$

where $\boldsymbol{e}(\cdot)$ is the subword embedding and $\boldsymbol{p}(\cdot)$ encodes spatial position. A subset of tag tokens is masked during pre-training, and the model learns to reconstruct them:

$$\mathcal{L}_{\text{Tag-Recon}} = -\sum_{k\in\mathcal{M}} \log P(t_k \mid \tilde{I}; \theta). \tag{8}$$

This mechanism equates detector tags with natural language tokens by placing them into a shared embedding layer and prediction head. The consequence is that words and tags become interchangeable anchors, allowing implicit alignment even without co-occurring training examples. This setup mirrors multilingual pre-training, where overlapping lexical tokens such as numbers or names enforce cross-lingual consistency.

3.2.3. Unified Embedding and Multimodal Projection

To further integrate modalities, VISTRA builds a shared subword vocabulary via byte pair encoding (BPE). Both textual tokens and detector tags are represented using the same token embedding matrix, augmented with position and segment embeddings. For an input sequence S of modality m, the embedding is:

$$\boldsymbol{X}_S = \boldsymbol{E}_{\text{token}}(S) + \boldsymbol{E}_{\text{position}}(S) + \boldsymbol{E}_{\text{segment}}(m). \tag{9}$$

Beyond embedding, VISTRA employs a projection head g_ϕ and contrastive alignment to refine semantic consistency. For a region feature $\boldsymbol{r}_j$ and its tag $\boldsymbol{d}_j$, we enforce:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(\text{sim}(g_\phi(\boldsymbol{r}_j), g_\phi(\boldsymbol{d}_j))/\tau)}{\sum_k \exp(\text{sim}(g_\phi(\boldsymbol{r}_j), g_\phi(\boldsymbol{d}_k))/\tau)}, \tag{10}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity and τ is a temperature parameter. This objective pulls semantically matched pairs closer while pushing apart unrelated instances, thereby strengthening modality invariance.

3.2.4. Final Training Objective

The complete pre-training loss of VISTRA integrates reconstruction and alignment components:

$$\mathcal{L}_{\text{VISTRA}} = \mathcal{L}_{\text{unsup}} + \lambda \cdot \mathcal{L}_{\text{align}}, \tag{11}$$

 doi:10.20944/preprints202509.0958.v1

with λ weighting the alignment term. After pre-training, VISTRA can be adapted to downstream V&L tasks such as VQA or retrieval by simply providing paired inputs, leveraging its already aligned embedding space without architectural changes.

4. Experiments

To rigorously assess the effectiveness of our unsupervised framework and to establish its practical utility, we conduct a wide range of experiments across several representative settings. The experimental design serves multiple purposes: (i) to compare the gap between models trained with paired supervision and our unsupervised alternative, (ii) to examine the adaptability of the proposed method in scenarios where aligned corpora are missing, and (iii) to analyze the contributions of individual components such as detector tags, corpus selection, and hybrid training strategies. Through these systematic studies, we aim to demonstrate not only the competitiveness of VISTRA, but also its robustness and versatility under different data availability conditions.

4.1. Evaluation with Simulated Alignment Removal

We first adopt a controlled simulation protocol, which has been a common practice in unsupervised learning literature, by intentionally discarding alignment information from datasets that originally contain paired annotations. Specifically, we repurpose Conceptual Captions (CC), which consists of millions of loosely associated image-text pairs, into a pseudo-unaligned setup. This allows us to conduct a fair comparison between supervised and unsupervised models trained on exactly the same raw corpus, isolating the effect of paired supervision. The central goal here is to investigate whether VISTRA can recover strong cross-modal representations purely from unpaired instances.

Overall Benchmark Results.

Table 1 summarizes performance across four widely used benchmarks: VQA 2.0 for question answering, NLVR2 for compositional reasoning, Flickr30K for retrieval, and RefCOCO+ for referring expression grounding. Across the board, our unsupervised model surpasses the Base model (VisualBERT initialized only from BERT, without multimodal pre-training), confirming the effectiveness of unsupervised representation learning. While the supervised VisualBERT baseline has marginally higher scores, the differences are surprisingly small, given that VISTRA never observes any explicit alignment during training. This result supports the central claim that cross-modal grounding can emerge from disjoint corpora with the right architectural and objective choices.

Table 1. Main evaluation results across four V&L benchmarks. Our unsupervised model **VISTRA** performs competitively with supervised models, and surpasses several earlier supervised methods like ViLBERT and VL-BERT on multiple tasks.

Model	Aligned	Unaligned		VQA	NLVR		Flickr30K			RefCOCO+		
		Image	Text	Test-Dev	Dev	Test-P	R@1	R@5	R@10	Dev	TestA	TestB
Pre-BERT	-	-	-	70.22	54.1	54.8	48.60	77.70	85.20	65.33	71.62	56.02
ViLBERT	3M	0	0	70.55	-	-	58.78	85.60	91.42	72.34	78.52	62.61
VL-BERT	3M	0	∼50M	71.16	-	-	-	-	-	71.60	77.72	60.99
UNITER$_{cc}$	3M	0	0	**71.22**	-	-	-	-	-	72.49	79.36	63.65
SVB	3M	0	2.5M	70.87$_{\pm.02}$	**73.44**$_{\pm.51}$	**73.93**$_{\pm.51}$	**61.19**$_{\pm.06}$	**86.32**$_{\pm.12}$	**91.90**$_{\pm.02}$	**73.65**$_{\pm.11}$	**79.48**$_{\pm.36}$	**64.49**$_{\pm.22}$
Base	0	0	0	69.26	68.40	68.65	42.86	73.62	83.28	70.66	77.06	61.43
VISTRA	0	3M	5.5M	70.74$_{\pm.06}$	71.74$_{\pm.24}$	71.02$_{\pm.47}$	55.37$_{\pm.49}$	82.93$_{\pm.07}$	89.84$_{\pm.21}$	72.42$_{\pm.06}$	79.11$_{\pm.08}$	64.19$_{\pm.54}$

Case Study: Flickr30K Retrieval.

A notable highlight is observed on Flickr30K. Here, VISTRA achieves Recall@1 of 55.37, dramatically higher than the baseline (42.86) and approaching the supervised score of 61.19. Considering that no alignment signal is available in our setting, this performance suggests that semantic anchors and contextual prediction objectives are sufficient to construct image-text mappings of competitive quality.

 doi:10.20944/preprints202509.0958.v1

The result demonstrates the surprising resilience of unsupervised objectives in capturing semantic structures that transfer well to retrieval.

Efficiency Considerations.

An additional advantage of our framework is its resource efficiency. All experiments are trained from $\text{BERT}_{\text{base}}$ initialization using 4 NVIDIA V100 GPUs (16GB memory) for 10 epochs. Full unsupervised training completes in roughly 72 hours. This low computational overhead indicates that VISTRA is suitable for groups without access to massive compute clusters, offering a practical route to high-quality multimodal encoders without costly annotation.

4.2. Experiments with Realistic Unaligned Sources

We next move beyond controlled simulations to evaluate more realistic situations where corpora are entirely disjoint and text resources are unrelated to image descriptions. This setup is intended to approximate the conditions encountered in real-world applications, where photographs and large text corpora coexist but are rarely aligned.

Training on Disjoint Text and Image Corpora.

We construct VISTRA-disjoint by using CC for images, SBU Captions Ordonez et al. (2011) for textual descriptions, and BookCorpus for additional general-domain text. Results indicate that this configuration yields the best performance on three out of four benchmarks. This demonstrates that heterogeneous corpora, though unaligned, provide complementary signals that improve the diversity of learned features and support stronger generalization than homogeneous caption-only sources.

Training on Purely General-Domain Text.

We further challenge our model with VISTRA-nocaption, which replaces all caption-style data with BookCorpus only. Even under this extreme setting, the model sustains strong accuracy (e.g., 71.47 on NLVR2 Dev and 64.25 on RefCOCO+ TestB). This outcome suggests that generic corpora, despite being unrelated to images, contain sufficient overlap in common nouns and object words (e.g., "car," "man," "tree") to align effectively with detector tags. Such findings highlight the applicability of VISTRA in contexts where no caption annotations are available at all.

Broader Implications.

Together, these findings establish that VISTRA can thrive in natural data conditions where multimodal resources are heterogeneous and uncurated. This widens the scope of multimodal pre-training to specialized domains such as medicine or law, where paired annotations are difficult or impossible to obtain, but large collections of images and documents exist separately.

Table 2. VISTRA maintains strong performance when trained on disjoint or non-caption corpora. Disjoint caption sources (VISTRA-disjoint) and general-domain text (VISTRA-nocaption) still yield robust results across tasks.

Model	Text Type		VQA	NLVR		Flickr30K			RefCOCO+		
	Caption	General	Test-Dev	Dev	Test-P	R@1	R@5	R@10	Dev	TestA	TestB
Base	-	-	69.26	68.40	68.65	42.86	73.62	83.28	70.66	77.06	61.43
VISTRA	CC	BC	70.74	71.74	71.02	55.37	82.93	89.84	72.42	79.11	64.19
VISTRA-disjoint	SBU	BC	70.70	**71.97**	**72.11**	**56.12**	82.82	**90.12**	**73.05**	**79.48**	64.19
VISTRA-nocaption	-	BC	70.47	71.47	71.19	54.36	82.22	89.24	72.96	79.30	**64.25**

4.3. Impact of Detector Tags and Semi-supervised Variants

We next examine the contribution of semantic anchors by ablating detector tags. Table 3 shows that excluding tags consistently reduces performance, with the most pronounced drops observed in the unsupervised configuration. For example, Flickr30K Recall@1 decreases from 55.37 to 50.56, and

 doi:10.20944/preprints202509.0958.v1

NLVR2 Test-P accuracy drops by over two points. These results empirically confirm that tags act as indispensable structural pivots that tether textual and visual spaces when alignment is missing.

Table 3. Ablation on detector tag usage. Removing tags severely impacts performance in the unsupervised case. Hybrid pre-training shows further improvement over pure supervised models.

Model	VQA	NLVR2		Flickr30K			RefCOCO+		
	Test-Dev	Dev	Test-P	R@1	R@5	R@10	Dev	TestA	TestB
Base w/o Tags	69.06	51.98	52.73	48.40	78.20	87.18	70.15	76.91	61.72
VISTRA w/o Tags	69.87	67.90	68.92	50.56	80.22	88.32	71.94	77.79	62.38
VISTRA (ours)	70.74	71.74	71.02	55.37	82.93	89.84	72.42	79.11	64.19
Supervised-VB w/o Tags	70.49	72.56	73.53	60.26	85.58	91.64	72.70	77.93	62.99
Supervised-VB (full)	70.87	73.44	73.93	61.19	86.32	91.90	73.65	79.48	64.49
Hybrid (semi-supervised)	**71.05**$_{\pm.02}$	**73.80**$_{\pm.26}$	**74.82**$_{\pm.25}$	**62.42**$_{\pm.36}$	**87.02**$_{\pm.35}$	**92.45**$_{\pm.28}$	**74.01**$_{\pm.25}$	**80.18**$_{\pm.23}$	**64.89**$_{\pm.24}$

Detector Tags as Alignment Catalysts.

The larger performance gap in the unsupervised case compared to the supervised baseline highlights the catalytic role of tags: without explicit supervision, anchors become the dominant mechanism that facilitates cross-modal consistency. In their absence, the model struggles to discover alignment signals, underscoring the necessity of structured semantic bridges.

Hybrid Training as a Synergistic Strategy.

We also experiment with a semi-supervised regime that integrates a small amount of aligned data with large unaligned corpora. This hybrid configuration yields the strongest overall performance, surpassing both purely supervised and unsupervised models. For instance, on NLVR2 Test-P the hybrid model attains 74.82, higher than either baseline. This observation indicates that combining curated annotations with abundant unpaired data can lead to synergistic improvements, suggesting an attractive direction for future multimodal pre-training paradigms.

4.4. Consolidated Insights

Our experimental exploration leads to several overarching insights:

- VISTRA performs competitively with supervised methods on diverse benchmarks, even though it relies solely on unpaired corpora.
- Detector-generated tags are indispensable for unsupervised grounding, providing the structural anchors needed to link modalities in the absence of alignment.
- The framework generalizes well across challenging setups, including training on disjoint corpora and text-only datasets devoid of captions.
- Incorporating hybrid supervision further enhances performance, pointing to a promising avenue for blending annotated and unannotated resources in future research.

Collectively, these findings reinforce our central thesis: scalable and high-quality multimodal representations can be obtained without reliance on expensive parallel data. By exploiting semantic anchors and carefully designed objectives, VISTRA presents a practical path forward for building annotation-efficient V&L models.

5. Conclusion and Future Work

This paper has introduced a new perspective on vision-and-language (V&L) representation learning by demonstrating that competitive multimodal encoders can be obtained without any reliance on explicitly paired image-text annotations. Our framework, named **VISTRA** (*Vision-Text Representation Alignment*), rethinks pre-training by decoupling the learning of visual and textual corpora while using detector-generated tags as implicit semantic bridges. Through this design, the heavy requirement for

 doi:10.20944/preprints202509.0958.v1

curated parallel datasets is substantially relaxed, showing that even loosely grounded signals from disjoint corpora can suffice to construct strong multimodal representations.

We validated the effectiveness of this paradigm through extensive experiments on four standard V&L benchmarks: VQA, NLVR2, Flickr30K Retrieval, and RefCOCO+. The results consistently indicate that VISTRA is capable of matching or even surpassing a number of supervised and weakly supervised baselines. Particularly noteworthy is its robustness in more challenging settings—such as training with fully disjoint caption sources or with general-domain text devoid of explicit visual descriptions—where the model continues to achieve strong results. Ablation analyses further confirmed that detector tags are indispensable, as they provide soft anchors that tie linguistic symbols to visual regions, thereby enabling cross-modal grounding under weak or noisy supervision.

At the methodological level, the essence of VISTRA lies in its modality-agnostic Transformer backbone, unified embedding vocabulary shared across textual tokens and detector tags, and its reconstruction-driven pre-training losses. Together, these design elements foster the emergence of semantically meaningful alignment across modalities without the need for explicit pairwise annotation. The addition of tag-level masked modeling augments this process by introducing lightweight yet highly effective visual-linguistic cues, thereby improving the quality of grounding and enhancing transferability to downstream tasks.

Beyond its empirical validation, the framework also points toward broader implications. In real-world or low-resource scenarios where curated parallel corpora are costly or infeasible—such as medical imaging, surveillance footage, or industrial inspection—VISTRA provides a scalable solution that leverages the natural abundance of unpaired corpora. The ability to extract value from such unstructured data streams could democratize access to multimodal pre-training, extending it to domains and communities that have historically been excluded due to annotation bottlenecks. Moreover, the approach encourages reconsideration of how multimodal systems are built, shifting the focus from annotation-heavy pipelines to annotation-light but semantically rich alternatives.

Looking forward, several natural extensions arise. One promising direction is to expand VISTRA into the temporal domain, enabling unsupervised pre-training for video-language tasks through frame-level anchoring and sequence-aware modeling. Another avenue involves exploring stronger alignment induction methods—such as adversarial alignment or cross-modal contrastive learning—to further tighten the cohesion between modalities. In addition, integrating higher-order symbolic signals, including scene graphs or structured semantic parses, with detector tags may unlock compositional reasoning capabilities and improve generalization to complex multimodal inference tasks. Finally, studying how unaligned multimodal pre-training interacts with downstream adaptation techniques such as prompt tuning or lightweight finetuning could provide practical strategies for task-specific deployments.

In conclusion, VISTRA establishes an alternative path toward scalable multimodal pre-training that is efficient, annotation-light, and robust across domains. By proving that strong performance can be achieved without aligned corpora, our study underscores the viability of unsupervised pre-training as a foundation for generalizable and interpretable V&L models, and paves the way for future research into broader, more inclusive multimodal intelligence.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *arXiv preprint arXiv:2102.08981*.

 doi:10.20944/preprints202509.0958.v1

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020a. Graph optimal transport for cross-domain alignment. *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020c. UNITER: Universal image-text representation learning. ECCV.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Multilingual BERT readme document. https://github.com/google-research/bert/blob/master/multilingual.md.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. *arXiv preprint arXiv:2009.13682*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2020. Probing text models for common ground with visual representations. *arXiv preprint arXiv:2005.00619*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open

 doi:10.20944/preprints202509.0958.v1

images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. In *International Journal of Computer Vision (IJCV)*.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Gen Li, N. Duan, Yuejian Fang, Daxin Jiang, and M. Zhou. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yikuan Li, Hanyin Wang, and Yuan Luo. 2020c. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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)*.

Jiasen Lu, Batra Dhruv, Parikh Devi, and Lee Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*.

Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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.

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).

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.

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.

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)*.

 doi:10.20944/preprints202509.0958.v1

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

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).

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. doi:10.1007/s00530-010-0182-0.

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.

Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.

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.

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.

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.

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.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.

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/.

Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016a. URL http://arxiv.org/abs/1604.08608.

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.

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. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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.

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*, 2024a.

A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.

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, 2020a.

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.

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.

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, 2020b.

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.

Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.

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.

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*, 2023a.

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.

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.

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, 2022a.

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.

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.

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, 2023b.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024a. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024b. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025a. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.

Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025b. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).

Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.

Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

 doi:10.20944/preprints202509.0958.v1

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, 2021a.

D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

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, 2021b.

K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

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*, 2023a.

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. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023c.

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.

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*, 2024b.

Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.

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, 2024c.

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.

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, 2022b.

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.

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.

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, 2023b.

S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.

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,*, 2024d.

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.

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.

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.

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, 2022c.

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.

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.

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, 2020c.

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.

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, 2018b.

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, 2023d.

P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.

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, 2023c.

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, 2023a.

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, 2023e.

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, 2023d.

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.

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.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.