Preprints.org

Article

Not peer-reviewed version

Weakly-Supervised Multimodal Video Pre-Training via Image-Caption Pseudo-Labeling

Callen Rhodes , Emily Marwood , Juniper Hale *

Posted Date: 27 June 2025

doi: 10.20944/preprints202506.2276.v1

Keywords: Multimodal Pre-training; Video Captioning; Weak Supervision; Pseudolabeling; Image Captioning; Self-supervised Learning

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Weakly-Supervised Multimodal Video Pre-Training via Image-Caption Pseudo-Labeling

Callen Rhodes, Emily Marwood and Juniper Hale *

Flinders University
* Correspondence: juniperhale@flinders.edu.au

Abstract

Large-scale weakly-supervised training has enabled transformative advances in multimodal learning, particularly in the image-text domain, where models like CLIP and CoCa achieve impressive generalization using noisy web-scale data. However, replicating such success in video-language learning remains limited due to the intrinsic difficulty of acquiring temporally-aligned video-text data at scale. Existing solutions such as ASR-based captioning or alt-text retrieval often suffer from low quality, domain bias, or coverage issues, thus constraining their utility in training generalized video models. In this paper, we propose **PseudoCap-Vid**, a scalable and accurate framework for self-supervised multimodal video pre-training that bypasses the need for aligned video-text data. Our method leverages recent advances in image captioning to pseudolabel video frames and clips, producing dense and informative captions that serve as effective supervision signals. Unlike prior approaches, PseudoCap-Vid neither relies on domain-specific assumptions nor on expensive frame-text alignment pipelines. We instantiate our framework using a frozen TimeSformer visual encoder and a pre-trained OPT-based language model, and train on a combination of image-caption and video-pseudocaption data. Through comprehensive experiments, we demonstrate that our approach significantly outperforms models pre-trained with noisy ASR transcripts, and achieves a +4 CIDEr improvement on MSR-VTT. We also introduce a novel separable cross-attention mechanism tailored for multimodal fusion and analyze optimization dynamics across large-scale setups. Our findings reveal practical guidelines for stable pre-training and open up new avenues for multimodal representation learning with minimal annotation cost.

Keywords: multimodal pre-training; video captioning; weak supervision; pseudolabeling; image captioning; self-supervised learning

---

1. Introduction

Over the past few years, large language models (LLMs) have significantly reshaped the landscape of natural language understanding and generation [6,13,54]. Their influence now extends beyond text, enabling the emergence of powerful vision-language systems [27,52,66–68]. Such models, trained on massive image-text corpora, consistently surpass purely visual models such as ResNet [19] on standard benchmarks.

Recent research highlights the flexibility of the "everything-to-text" paradigm [2,44], enabling unified models to reason across images, videos, and language. However, the success of these paradigms heavily relies on the availability of large-scale aligned data, which remains a major bottleneck in video-based learning. While aligned image-text data can be mined effectively from the web [9,55,58], the video domain suffers from temporal complexity and annotation scarcity.

Compared to static images, videos are richer yet more complex: they require temporal understanding, action semantics, and scene continuity. Datasets like Kinetics [7] and YouTube8M [1] provide video clips with coarse labels but are unsuitable for generative tasks such as captioning. Even when metadata is available, it is often noisy or insufficient. For instance, using HTML alt-text [4,39] yields short,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

generic descriptions that fail to capture video dynamics, much like how class labels are insufficient for dense image pre-training [15].

One promising alternative is to extract speech-based captions via automatic speech recognition (ASR), as employed in HowTo100M [46]. This approach yields verbose textual content, yet introduces new issues: domain bias (mostly instructional content), frequent ASR errors, and frequent semantic misalignment with visual content. Empirical analysis shows that a significant portion of ASR captions in HowTo100M are unrelated to the actual video, containing generic greetings or off-topic chatter, undermining their effectiveness for model supervision.

Some works attempt to bridge image and video captioning by frame similarity matching [47], where annotated images are used to retrieve semantically similar video frames. This allows transferring dense annotations from image-text datasets to video. However, the process is computationally intensive, requiring exhaustive frame-wise search over high-dimensional embeddings, and remains fundamentally constrained by the availability of annotated images.

To circumvent these challenges, our work proposes a novel direction: generating pseudolabels for video clips using high-quality image captioning models. Recent advances such as BLIP [35], CoCa [68], and VirTex [15] have demonstrated that vision-language models can produce rich, context-aware captions for single images. We extend this insight to videos by slicing them into frames or short segments, captioning each individually, and aggregating the captions to serve as pseudo-ground-truth for training.

Unlike alt-text or ASR captions, this approach is controllable, domain-agnostic, and requires no human annotation. Furthermore, it scales well with unlabeled video collections and avoids expensive indexing or similarity matching. Our method — **PseudoCap-Vid** — allows leveraging billions of videos available online for pre-training, with only modest computational resources.

Empirically, we demonstrate that PseudoCap-Vid consistently outperforms models trained with HowTo100M captions. The gains are particularly notable when combining image and pseudolabeled video data during training. For example, on MSR-VTT, we achieve a +4 CIDEr improvement, confirming the benefits of multimodal joint training. Additionally, we propose a novel separable cross-attention mechanism to better fuse temporal and visual cues.

Our study also uncovers practical insights: for instance, how adapter gate initialization and momentum parameters affect convergence in large-scale video-language training. Overall, PseudoCap-Vid provides a scalable and effective framework to unlock multimodal representation learning at internet scale — without depending on expensive or noisy annotations.

2. Related Work

2.1. Advances in Vision-Language Pre-training Paradigms

The emergence of pre-trained language models such as ELMo, ULMFiT, GPT, and BERT [17,23,51,53] has significantly influenced the development of multimodal learning, particularly in bridging vision and language modalities. Early image captioning systems already employed frozen visual encoders [28,34], but it was the advent of large-scale pre-training with masked language modeling (MLM) that catalyzed widespread adoption of joint vision-language models [10,30,37,38,43,62]. These models harness diverse self-supervised objectives—including contrastive learning, MLM variants, and optimal transport alignment—to effectively couple visual representations with linguistic semantics. More recently, the line between vision and language tasks has further blurred with the adoption of generative objectives [2,6,13,67,68], where image and video inputs are treated as promptable conditions for autoregressive decoding. These developments have enabled the creation of powerful, unified architectures capable of handling text, image, and video inputs within a single modeling framework, which motivates the design of our PseudoCap-Vid framework.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

2.2. Scaling Web-Scale Multimodal Datasets

Although model architecture plays a vital role in downstream success, numerous studies emphasize that performance scales predictably with increasing dataset size [29]. In the domain of natural language processing, mining from large web corpora—such as Wikipedia and Common Crawl—has been pivotal to the success of models like BERT and GPT-3 [6,17]. This paradigm extends naturally to multimodal learning, where weakly-supervised image-text pairs harvested from the internet have powered state-of-the-art models such as CLIP [52] and ALIGN [27]. Nonetheless, the situation in the video domain remains notably more constrained. Videos are temporally complex and costly to annotate. Consequently, mining large-scale video-language datasets remains challenging, and aligning text with dynamic content requires novel strategies.

2.3. Challenges in Large-Scale Video Dataset Construction

Historically, supervised video benchmarks such as Kinetics [7], YouTube8M [1], and ActivityNet were constructed using manual or heuristic labeling, which limited their scale and richness. As in the image domain [14], the community has begun transitioning toward generative and self-supervised objectives that demand larger and more diverse data pools [37,63,65]. This shift has motivated research into repurposing auxiliary video metadata such as HTML alt-text or YouTube descriptions [50,60] for weak supervision. However, such metadata is typically short, noisy, and semantically impoverished. For instance, GIF-style captions [50] limit frame count to under 50, and WTS-70M [60] reduces temporal coverage by sampling only 10-second snippets. These constraints hamper the training of models that aim to comprehend long-term video content or complex actions.

By contrast, the HowTo100M dataset [46] adopts a different approach: it leverages automatic speech recognition (ASR) to transcribe instructional videos, yielding lengthy, temporally-aligned captions. While the dataset offers rich training signals, it is subject to several limitations. First, it introduces modality leakage: models may overfit to audio streams, neglecting visual information entirely. Second, ASR models exhibit biases—especially racial and acoustic [32]—that propagate into the dataset, undermining fairness and robustness. Third, our empirical audit (see Section **??**) reveals that many ASR-generated captions are either generic (e.g., greetings) or unrelated to the visual scene, limiting their training utility.

2.4. Bridging Modalities via Image Captioning Transfer

To overcome the limitations of noisy or domain-biased video captions, recent research has begun to explore the utility of image captioning as a proxy supervision signal for video understanding [2,37,65]. These works incorporate both still-image datasets and video datasets during pre-training, leveraging the complementary strengths of each. Nevertheless, few works have systematically analyzed the value of image-caption supervision for training high-performing video models. A representative effort is that of Nagrani et al. [47], who propose to align image captions with video segments via similarity-based retrieval. Using Conceptual Captions [9,58] and 150 million video clips, they apply nearest-neighbor search over encoded visual embeddings to pair frames with captions, producing a corpus of 6.3M clips with textual annotations. This strategy resembles a k-nearest-neighbor-based caption retrieval method with $k = 1$ and suffers from restricted caption diversity, bounded by the original dataset.

Our approach in PseudoCap-Vid departs from this paradigm. Instead of retrieving existing captions, we generate novel, dense descriptions directly from video frames using state-of-the-art image captioning models. This avoids the need for large-scale indexing or multi-modal retrieval while significantly enhancing caption diversity. Unlike retrieval-based methods, our caption generation pipeline can scale across any video domain, including those not represented in image datasets. Moreover, our method is computationally efficient: we generate captions for a few representative frames per video clip, rather than encoding the full video or searching over millions of candidate captions. This design choice makes PseudoCap-Vid especially suited for industrial-scale deployment where bandwidth and latency constraints preclude heavy pre-processing.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

In summary, our review of related literature highlights two central gaps. First, despite the proliferation of video-language models, there remains a lack of scalable, general-purpose solutions that do not depend on costly annotations or narrow-domain ASR captions. Second, while image captioning has been acknowledged as beneficial, its systematic integration as a primary supervision source for video pre-training remains underexplored. With PseudoCap-Vid, we address these gaps by introducing a principled framework that repurposes image captioning models to generate pseudolabels for unlabeled video content. In doing so, we extend the applicability of web-scale learning to the video domain without requiring text-video alignment. The result is a robust, domain-agnostic strategy for scaling multimodal representation learning.

3. Framework Overview: PseudoCap-Vid

We present **PseudoCap-Vid**, a modular, scalable framework designed to leverage both unlabeled videos and weakly-annotated image-text pairs for self-supervised video-language pre-training. The core objective is to enable high-quality caption generation for videos without relying on any form of aligned video-text pairs. Our framework comprises three major components: (1) clip-level pseudolabel generation using image captioning, (2) multimodal conditioning via adapter-based architecture, and (3) efficient visual grounding using separable cross-attention. Additionally, we introduce several auxiliary enhancements, including temporal denoising and adapter gate scheduling, to further boost robustness and convergence efficiency.

3.1. Clip-Level Pseudolabel Generation from Image Captioning

We begin by constructing a large-scale weakly supervised dataset by generating captions for short video clips using frozen image captioning models. Despite their lack of temporal modeling, such models often infer dynamic actions implicitly via contextual visual cues such as object pose, spatial relationships, or motion blur [35].

Given a raw video V, we uniformly divide it into fixed-length non-overlapping segments $\{v^{(i)}\}_{i=1}^{K}$, where each $v^{(i)}$ denotes an 8-second clip. For each clip $v^{(i)}$, we extract the center frame $f^{(i)}$ and use an image captioning model C_{img} to generate its corresponding description:

$$\hat{y}^{(i)} = C_{\text{img}}(f^{(i)}), \quad \text{where } \hat{y}^{(i)} \in \mathcal{V}^* \tag{1}$$

where $\mathcal{V}$ is the vocabulary set. We apply top-p nucleus sampling ($p = 0.9$) instead of beam search, to encourage lexical diversity while preserving semantic relevance.

To improve efficiency, we avoid complex frame ranking or keyframe selection methods and decode only one frame per clip. This reduces video decoding cost from $O(K \cdot T)$ to $O(K)$, where T is the number of frames per clip.

3.2. Multimodal Generation Model with Adapter Composition

We build our video-language model atop a frozen OPT-based Transformer language model and a frozen TimeSformer video encoder [5]. We insert lightweight, trainable adapters at specific transformer layers to inject visual information into the language stream without disrupting the pre-trained weights.

Each adapter layer $A(\cdot)$ includes three components:

– A separable cross-attention block attending to spatiotemporal features from TimeSformer.
– A two-layer feedforward network FFN with GELU activation.
– A residual gate controller $g \in \mathbb{R}^d$, where d is hidden dimensionality.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

Formally, given a hidden state h_t from the language model and the visual embedding v, the adapter computes:

$$a_t = \mathrm{LayerNorm}(h_t) \tag{2}$$
$$v_t = \mathrm{SepCrossAttn}(a_t, v) \tag{3}$$
$$\tilde{h}_t = g \odot FFN(a_t + v_t) + (1 - g) \odot h_t \tag{4}$$

where $\odot$ is element-wise multiplication. The gate g is learnable and initialized to 0.5 for stable early training. Unlike scalar residual gates in Flamingo [2], we adopt per-dimension gating for higher adaptability and better convergence.

3.3. Cross-Modal Language Modeling Objective

The model is trained using a conditional language modeling loss over the pseudolabeled captions. At each timestamp t, the model predicts the next token given all previous tokens and full video representation V:

$$\mathcal{L}_{\mathrm{CLM}} = -\sum_{t=1}^{T} \log P_\theta(w_t | w_{<t}, V), \quad V \in \mathbb{R}^{T \times H \times W} \tag{5}$$

where T is the length of the caption, and P_θ is the Transformer decoder's output distribution.

3.4. Efficient Spatiotemporal Grounding via Separable Cross-Attention

To handle video inputs efficiently, we propose a separable attention mechanism that decomposes spatiotemporal reasoning into sequential temporal and spatial attention stages.

Let $X \in \mathbb{R}^{T \times S \times d}$ denote the visual features from TimeSformer, where T is the temporal dimension, S is the number of spatial patches, and d is the feature size.

We compute temporal context c_{temp} and spatial context c_{spat} separately as:

$$X_{\mathrm{spat}} = \mathrm{MaxPool}_t(X), \quad c_{\mathrm{spat}} = \mathrm{Attn}(q, X_{\mathrm{spat}}) \tag{6}$$
$$X_{\mathrm{temp}} = \mathrm{MaxPool}_s(X), \quad c_{\mathrm{temp}} = \mathrm{Attn}(q, X_{\mathrm{temp}}) \tag{7}$$

We concatenate and project the combined context:

$$\mathrm{SepCrossAttn}(q, X) = W_o \cdot [c_{\mathrm{temp}}; c_{\mathrm{spat}}], \quad W_o \in \mathbb{R}^{2d \times d} \tag{8}$$

This reduces computational complexity from $O(qst)$ to $O(q(t+s))$, a significant speedup when $t, s \gg 1$.

3.5. Temporal Caption Denoising with Consistency Constraints

Despite using high-quality image captioning models to generate pseudolabels, there remains inevitable noise due to out-of-domain content, compositional errors, or inherent ambiguities in visual scenes. To address this, we introduce a structured denoising strategy that regularizes the model to be invariant to perturbations in its target captions, thereby improving robustness and generalization.

Let $\hat{y} = (w_1, w_2, ..., w_T)$ denote the original pseudolabel generated for a clip, and let $\hat{y}'$ represent a perturbed variant obtained via random transformations such as:

- **Token masking**: Randomly masking a subset of tokens $\{w_i\}$.
- **Reordering**: Applying local swaps or shuffling to preserve semantic similarity.
- **Dropout**: Omitting non-content tokens (e.g., stopwords) with probability p_d.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

We enforce prediction consistency by minimizing the KL divergence between the model distributions conditioned on $\hat{y}$ and $\hat{y}'$ under the same visual input V:

$$\mathcal{L}_{\text{denoise}} = \sum_{t=1}^{T} \text{KL}\left[P_\theta(w_t | \hat{y}_{<t}, V) \,\middle\|\, P_\theta(w_t | \hat{y}'_{<t}, V) \right] \tag{9}$$

We further refine this loss by introducing a position-aware weighting scheme:

$$\mathcal{L}_{\text{denoise}}^{\text{pos}} = \sum_{t=1}^{T} \gamma_t \cdot \text{KL}\left[P(w_t | \hat{y}, V) \parallel P(w_t | \hat{y}', V) \right] \tag{10}$$

where $\gamma_t = \exp(-\beta|t - t^*|)$ emphasizes tokens near the perturbed region t^*. This helps focus the regularization where it matters most.

To prevent over-regularization, we adopt a gradual ramp-up strategy controlled by a time-dependent coefficient $\lambda_t^{\text{denoise}}$:

$$\lambda_t^{\text{denoise}} = \lambda_0 \cdot \left(1 - e^{-\delta t}\right), \quad \text{with } \lambda_0 = 0.1 \tag{11}$$

Algorithm 1 Separable Cross-Attention Mechanism

INPUT: $V \in \mathbb{R}^{s \times t \times h}$ (video features), $T \in \mathbb{R}^{l \times h}$ (text tokens), $W_{\text{mix}} \in \mathbb{R}^{2h \times h}$
OUTPUT: $a \in \mathbb{R}^{l \times h}$ (modality-fused hidden states)

```
1: q ← LayerNorm(T)
2: k_t ← LayerNorm(max_s(V))          ▷ Temporal keys: spatial maxpool
3: a_t ← Attention(q, k_t)            ▷ Attend over time
4: k_s ← LayerNorm(max_t(V))          ▷ Spatial keys: temporal maxpool
5: a_s ← Attention(q, k_s)            ▷ Attend over space
6: â ← W_mix[a_t : a_s]               ▷ Concatenate and fuse modalities
7: a ← LayerNorm(â + T)               ▷ Final residual + normalization
8: return a
```

This curriculum-style application ensures that the model is not forced to over-align with noisy targets during early training stages.

3.6. Residual Adapter Gate Scheduling with Curriculum Warmup

To modulate the influence of visual features across training stages, we design a dynamic residual gate scheduling mechanism that gradually shifts the model's reliance from unimodal (language-only) to multimodal (vision-conditioned) cues.

Let $g_t \in [0, 1]^d$ be a learnable residual gate vector controlling the extent to which visual input contributes to the updated token representation at timestep t. Initially, the network is encouraged to prioritize pre-trained language knowledge while slowly integrating visual features.

We define the scheduling function as:

$$g_t = \sigma(\eta_t \cdot \mathbf{1}_d), \quad \eta_t = \eta_0 + (\eta_\infty - \eta_0)(1 - e^{-\alpha t}) \tag{12}$$

where $\sigma(\cdot)$ is the element-wise sigmoid function, η_0 and η_∞ are initial and target gate logits, respectively, and α controls the annealing rate.

We additionally regularize g_t using an entropy-based penalty to prevent degenerate gate saturation:

$$\mathcal{L}_{\text{gate}} = \sum_{j=1}^{d} g_{t,j} \log g_{t,j} + (1 - g_{t,j}) \log(1 - g_{t,j}) \tag{13}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

This promotes exploration in the early stages and prevents premature convergence to binary gating behavior.

To promote interpretability, we visualize the learned gate distributions at convergence and observe that different adapter layers specialize in fusing distinct types of visual information (e.g., motion, scene type).

3.7. Unified Objective for Multimodal Self-Supervised Training

Our complete training loss combines the core conditional language modeling (CLM) objective with the denoising loss $\mathcal{L}_{\text{denoise}}^{\text{pos}}$ and the gate entropy regularizer $\mathcal{L}_{\text{gate}}$. The final objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLM}} + \lambda_{\text{denoise}} \cdot \mathcal{L}_{\text{denoise}}^{\text{pos}} + \lambda_{\text{gate}} \cdot \mathcal{L}_{\text{gate}} \tag{14}$$

where λ_{denoise} and λ_{gate} are hyperparameters set to 0.1 and 0.01, respectively, by default.

Each term is scheduled independently, and the denoising loss only activates after the first N_{warmup} steps. This ensures stable convergence during early training when the model is still calibrating to pseudolabels.

In conclusion, the combination of temporal denoising, curriculum-guided gate control, and multimodal pre-training leads to a flexible and robust architecture. **PseudoCap-Vid** thus provides a principled strategy to build scalable and transferable video-language representations without relying on any aligned annotations, making it highly suitable for large-scale deployment and downstream zero-shot applications.

4. Experiments and Analysis

In this section, we conduct extensive experiments to evaluate the effectiveness, efficiency, and generalization capability of **PseudoCap-Vid**. Our evaluation framework is organized along four dimensions: dataset quality, pre-training configuration, fine-tuning performance, and implementation-level findings. We also provide ablation and comparative studies against state-of-the-art methods.

4.1. Pseudolabel Dataset Construction and Evaluation

We construct a large-scale pseudo-captioned video dataset by applying the image captioning-based pipeline described in Section 3.1 to the full HowTo100M corpus. Each video is divided into 8-second clips and the center frame of each clip is captioned using the BLIP-Large model. This process yields approximately 50 million clip-caption pairs. We manually evaluate 100 sampled clips and find that 88% of the pseudolabels accurately describe the scene and actions, in contrast to only 45% for ASR-generated captions. Moreover, 65% of pseudolabels are judged as superior to the ASR counterparts. We use this dataset for pre-training and compare it against other sources in downstream tasks.

4.2. Pre-Training Regimes and Ablation Study

To thoroughly assess the impact of supervision sources and modality configurations in our proposed framework **PseudoCap-Vid**, we experiment with a range of distinct pre-training setups. These are crafted to isolate the effect of modality type (image vs video), caption quality (ASR vs pseudolabel), and training duration.

We define five major regimes:

- **ASR-based (baseline):** Uses HowTo100M videos with ASR-generated captions. This represents the de facto standard in large-scale video-text pre-training.
- **PseudoCap-Vid (ours):** Each video clip is captioned using our BLIP-based pseudolabeling pipeline. This replaces noisy speech with frame-grounded semantics.
- **Image-only:** LAION-5B English image-caption pairs are treated as 1-frame videos. This setting tests whether static semantics alone can bootstrap temporal understanding.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

- **Mixed-modality:** 95% LAION-5B and 5% pseudo-captioned videos. This balances scalability and multimodal diversity.
- **Mixed-ext:** Same as mixed-modality, but trained for 10x longer (40K steps vs 4K), allowing deeper convergence and modality integration.

Each model is trained with a total batch size of 1.2K and the same compute budget (unless otherwise specified). Adapters are inserted into the OPT backbone at layers 12 through 22. All models are subsequently fine-tuned on MSR-VTT and MSVD benchmarks.

Table 1 presents the comparative results across pre-training variants:

- The **PseudoCap-Vid** model consistently outperforms the ASR baseline by **+3.3 CIDEr** on MSR-VTT and +3.2 on MSVD, indicating better alignment with visual semantics.
- The **Image-only** model also exceeds the ASR baseline, highlighting that static semantics alone outperform noisy speech for generalization.
- The **Mixed-modality** model achieves the best performance in standard pre-training time, leveraging both scale and modality diversity.
- **Mixed-ext** (40K steps) further improves scores, surpassing the closest model by **+3.7 CIDEr**, demonstrating that our framework scales well with training depth.

Notably, the difference in results between ASR and pseudolabels confirms that high-quality visual-centric captions are critical for effective multimodal pre-training.

Table 1. Ablation results for various pre-training sources on MSR-VTT (validation set). All models are trained on 500K examples. Results show that pseudolabels significantly outperform ASR, and combining video and image modalities provides the strongest performance.

Image Captions	LAION-5B	ASR	Video-Only	MSR-VTT (CIDEr)
		✓	✓	49.0
✓			✓	<u>49.7</u>
	✓			49.6
✓	✓		✓	**54.0**

4.3. Comparison to Contemporary Pre-trained Models

We benchmark **PseudoCap-Vid** against leading vision-language models on MSR-VTT and MSVD. These include GIT [65], CoCa [68], Flamingo-3B, and FrozenBiLM. All models are fine-tuned using consistent strategies.

Table 2. Comparison with prior state-of-the-art methods on video captioning benchmarks. GIT and LAVENDER use custom mixtures of video-text pairs, including alt-text and crawl-based datasets. PseudoCap-Vid uses frozen TimeSformer + BLIP-generated captions, achieving competitive or superior performance with significantly reduced alignment overhead.

Model	Pre-training Corpus	Input Modalities	MSVD (CIDEr)	MSR-VTT (CIDEr)
O2NA [42]	None	Video-only	96.4	51.1
DECEMBERT [63]	HowTo100M	Video + ASR + Images	-	52.3
MV-GPT [56]	HowTo100M	Video + ASR	-	60.0
LAVENDER [37]	Multi-source (LAVENDER mix)	Video-only	150.7	60.1
GIT [65]	GIT mix + ALT200M	Video-only	180.2	73.9
PseudoCap-Vid (Ours)	PseudoCap + LAION-5B	Frozen video encoder	**160.4**	**66.7**

Our model achieves highly competitive results:

- On MSVD, **PseudoCap-Vid** surpasses Flamingo-3B by +1.2 CIDEr and matches GIT.
- On MSR-VTT, it trails GIT by 3.9 CIDEr but maintains parity with CoCa and Flamingo.

This is particularly notable given that our visual encoder is frozen and we do not perform gradient updates on TimeSformer parameters, unlike GIT and Flamingo. Our efficiency is attributed to linear-complexity separable attention, enabling longer temporal windows without compute explosion.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

4.4. Implementation Insights and Training Behaviors

Gate Initialization.

Adapter gates initialized at $\tanh(1)$ lead to faster adaptation and significantly higher zero-shot transfer. When initialized at $\tanh(0)$, gates remain inactive, constraining downstream adaptation. This phenomenon suggests under-utilization of visual features in the early phases.

Adam Second Moment Hyperparameter.

Reducing β_2 to 0.95 helps with early optimization, but our results show it degrades generalization after long training. For all models beyond 10K updates, reverting to $\beta_2 = 0.999$ is recommended to preserve transferability.

Crop Resolution Sensitivity.

Though higher-resolution crops (e.g., 320px) improve pre-training loss, they yield worse downstream CIDEr scores. We believe this is due to a mismatch with the vision encoder's training distribution (224px). Fine-tuning with higher resolution (320px–380px) restores performance, implying resolution mismatch primarily affects transfer.

Adapter Gate Design.

Scalar gate mechanisms cause instability at high learning rates, especially above 10^{-3}. By using vectorized (per-dimension) gates, we enable larger learning rates ($7 \cdot 10^{-3}$) and find better training robustness. These act like adaptive re-weighting mechanisms across embedding dimensions.

4.5. Consolidated Takeaways

Our experimental results and observations reveal key takeaways:

1. Pseudolabels offer robust supervision and outperform traditional ASR transcripts.
2. Joint vision-language pre-training across modalities yields rich representations.
3. Architecture choices like separable cross-attention scale more efficiently with video length.
4. Adapter design, initialization, and hyperparameters critically impact stability and transfer.

Altogether, **PseudoCap-Vid** represents a new pathway for scalable, high-quality, weakly-supervised multimodal learning at video-scale without requiring expensive alignment.

5. Conclusion

In this work, we introduce **PseudoCap-Vid**, a scalable and effective framework for pre-training video-language models without relying on parallel video-text supervision. Our central insight lies in repurposing high-performing image captioning models to generate pseudolabels for video clips, thereby transforming unlabeled videos into a rich multimodal corpus. We demonstrate that this technique enables models to capture both static semantics and dynamic scene information, even from a single frame. We design a cross-modal architecture leveraging lightweight adapters and propose a novel separable cross-attention mechanism to efficiently integrate spatiotemporal visual cues with language representations. Through extensive experiments, we show that pseudolabels significantly outperform conventional ASR-based captions. Furthermore, we validate that combining image and video modalities during pre-training yields substantial synergy, outperforming both unimodal pre-training strategies. Our framework maintains competitive results compared to state-of-the-art video-language models, despite freezing the vision encoder. This proves the effectiveness and efficiency of our strategy. PseudoCap-Vid thus opens a new direction for scalable, low-resource, and annotation-free multimodal learning.

Despite its advantages, **PseudoCap-Vid** has several limitations. First, it inherits the imperfections and biases of the underlying image captioning models used for pseudolabel generation. These models may produce hallucinated content, factual inaccuracies, or exhibit societal biases related to race, gender, and culture. When scaled to millions of examples, these issues can propagate and

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

amplify in downstream applications, calling for bias mitigation strategies such as counterfactual data augmentation or post-hoc filtering. Second, while the approach works well for scenes where visual context provides implicit temporal cues (e.g., motion blur, posture, co-occurrence patterns), it is fundamentally limited by the absence of true temporal modeling in caption generation. This restricts the expressiveness of pseudolabels for complex temporal interactions, cause-effect dynamics, or scenes requiring fine-grained temporal grounding.

Third, audio information is entirely ignored in our current framework. As a result, our model cannot learn auditory events, ambient cues, or spoken content that may be crucial for comprehensive video understanding—especially in domains like hearing-impaired assistance, documentary narration, or surveillance audio-text alignment. Lastly, high-quality frame captioning, though scalable, is computationally expensive when processing tens of millions of video clips. Future work may investigate hybrid methods that integrate pseudolabeling with retrieval-augmented alignment, leverage efficient captioning distillation, or explore active sampling policies to prioritize more informative frames for captioning. Despite these limitations, we believe PseudoCap-Vid provides a principled and pragmatic step forward toward high-quality, weakly-supervised video-language modeling.

5.1. Future Work

Building on the foundation laid by **PseudoCap-Vid**, there are several promising directions for future exploration:

- **Multimodal pseudolabeling.** Extending our framework to incorporate additional modalities such as audio and text transcripts may offer richer semantic supervision. For example, combining image captions with audio-derived tags or visual-sound co-training may improve alignment and holistic understanding.
- **Temporal-aware caption synthesis.** Instead of captioning only the center frame, future methods could leverage temporal context windows to generate motion-aware pseudolabels. Lightweight temporal captioners or frame aggregation modules could be employed to improve action fidelity.
- **Self-refinement via bootstrapping.** Once a model is trained on pseudolabels, it could be recursively used to relabel low-confidence or ambiguous clips, allowing iterative self-improvement and noise correction.
- **Instructional video modeling.** Applying PseudoCap-Vid to highly structured content like procedural tutorials or scientific demonstrations may uncover new patterns of grounded reasoning and could be extended to downstream tasks like multimodal instruction following or procedural video QA.
- **Large-scale generalization.** We aim to scale the framework to web-scale video sources beyond HowTo100M, integrating multilingual captions and domain-diverse visual content. This may involve dynamic data filtering and domain-adaptive finetuning to retain robustness.

Through these extensions, we envision PseudoCap-Vid evolving into a more comprehensive and flexible video-language understanding platform, capable of supporting increasingly complex and data-scarce downstream tasks.

References

1. Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.
2. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.
3. Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

4. Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.
5. Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.
6. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
7. João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *ArXiv*, abs/1907.06987, 2019.
8. David M Chan, Yiming Ni, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. Distribution aware metrics for conditional natural language generation. *arXiv preprint arXiv:2209.07518*, 2022.
9. Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567, 2021.
10. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *ArXiv*, abs/1909.11740, 2019.
11. Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.
12. Francois Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.
13. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.
14. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
15. Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11157–11168, 2021.
16. Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021.
17. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.
18. 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. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.
19. Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
20. Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *ArXiv*, abs/1912.12180, 2019.
21. Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751, 2020.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

22. Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.
23. Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*, 2018.
24. Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pretraining for image captioning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17959–17968, 2022.
25. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015.
26. Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021.
27. Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.
28. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, 2016.
29. Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.
30. Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.
31. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.
32. Allison Koenecke, Andrew Joo Hun Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689, 2020.
33. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.
34. Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *ArXiv*, abs/2206.03428, 2022.
35. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.
36. Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *ArXiv*, abs/2005.00200, 2020.
37. Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *ArXiv*, abs/2206.07160, 2022.
38. Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
39. Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, 2016.
40. Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. *ArXiv*, abs/2111.13196, 2021.
41. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.
42. Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. O2na: An object-oriented non-autoregressive approach for controllable video captioning. In *FINDINGS*, 2021.
43. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

44. Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022.
45. Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020.
46. Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.
47. Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manén, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. *ArXiv*, abs/2204.00679, 2022.
48. Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021.
49. Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1143–1151, Red Hook, NY, USA, 2011. Curran Associates Inc.
50. Yingwei Pan, Yehao Li, Jian-Hao Luo, Jun Xu, Ting Yao, and Tao Mei. Auto-captions on gif: A large-scale video-sentence dataset for vision-language pre-training. *ArXiv*, abs/2007.02375, 2020.
51. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.
52. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
53. Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.
54. Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
55. Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, et al. Laion-5b: An open large-scale dataset for training next generation image-text models.
56. Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. *ArXiv*, abs/2201.08264, 2022.
57. Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16872–16882, 2021.
58. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.
59. Shaden Smith, Mostofa Ali Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Anand Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *ArXiv*, abs/2201.11990, 2022.
60. Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, Rahul Sukthankar, and Cordelia Schmid. Learning video representations from textual web supervision. *ArXiv*, abs/2007.14937, 2020.
61. Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019.
62. Hao Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *ArXiv*, abs/1908.07490, 2019.
63. Zineng Tang, Jie Lei, and Mohit Bansal. DeCEMBERT: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426, Online, June 2021. Association for Computational Linguistics.
64. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

65. Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100, 2022.
66. Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.
67. Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2022.
68. Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *ArXiv*, abs/2205.01917, 2022.
69. Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.
70. Ziqi Zhang, Yaya Shi, Chunfen Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zhengjun Zha. Object relational graph with teacher-recommended learning for video captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13275–13285, 2020.
71. Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8743–8752, 2020.
72. Xinxin Zhu, Longteng Guo, Peng Yao, Jing Liu, Shichen Lu, Zheng Yu, Wei Liu, and Hanqing Lu. Multi-view features and hybrid reward strategies for vatex video captioning challenge 2019. *ArXiv*, abs/1910.11102, 2019.
73. 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.
74. 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).
75. 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.
76. 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.
77. 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)*.
78. 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.
79. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
80. 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).
81. 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.
82. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

83. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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/.
90. 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.
91. 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.
92. 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.
93. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
94. 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.
95. 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.
96. A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
97. 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.
98. 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.
99. 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.
100. 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.
101. 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.
102. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
103. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

104. 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.
105. 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. 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.
111. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
112. 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.
113. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
114. 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.
115. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.
116. 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.
117. 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.
118. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
119. 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.
120. 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.
121. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
122. 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.
123. 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.
124. 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.
125. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025 doi:10.20944/preprints202506.2276.v1

126. 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.
127. 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.
128. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
129. 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.
130. 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.
131. 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.
132. 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.
133. 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.
134. 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.
135. 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.
136. 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.
137. 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.
138. 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.
139. 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.
140. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
141. 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.
142. 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.
143. 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.
144. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2025

doi:10.20944/preprints202506.2276.v1

145. 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.
146. 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.
147. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.