Preprints.org

Article

Not peer-reviewed version

# Integrated Cross-Modal Learning for Interactive Video Conversation

Wyatt Carter , Emily Marwood , Riley Dawson *

Posted Date: 21 February 2025

doi: 10.20944/preprints202502.1772.v1

Keywords: Interactive Conversation; Multimodal Learning; Video Understanding; Cross-Modal Representation; Scene Analysis

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Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1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*

# Integrated Cross-Modal Learning for Interactive Video Conversation

**Wyatt Carter, Emily Marwood and Riley Dawson ***

Flinders University; wyatt.carter@flinders.edu.au (W.C.); marwoodemily@gmail.com (E.M.)
* Correspondence: dawson@flinders.edu.au

**Abstract:** Interactive video conversation is a complex multimodal task that requires the simultaneous interpretation of dynamic visual scenes, textual dialogue, and auditory signals when available. In recent years, significant progress has been achieved by leveraging powerful transformer-based language models, which have established new performance benchmarks. However, many of these advanced systems tend to focus excessively on textual features, resulting in an underutilization of the rich visual cues present in video data. To address this challenge, we propose a novel cross-modal framework, referred to as **CIMT**, which seamlessly integrates 3D convolutional neural networks (3D-CNNs) with transformer-based architectures into a unified visual encoder. This encoder is engineered to extract robust semantic representations by learning local temporal features and contextualizing them through self-attention mechanisms. The resulting visual features are effectively combined with text and audio representations within an end-to-end trained architecture. Experimental results on established interactive video conversation benchmarks demonstrate that CIMT significantly outperforms baseline models on both generative and retrieval tasks, highlighting the benefits of integrated visual-textual learning.

**Keywords:** interactive conversation; multimodal learning; video understanding; cross-modal representation; scene analysis

## 1. Introduction

The objective of interactive video conversation is to generate precise and contextually appropriate responses based on a short video clip and a series of dialogue turns. In practice, a model is provided with a dynamic visual scene and an ongoing conversation about that scene; for example, when queried with a follow-up question such as "Did she re-enter the room?", the system must correctly resolve the pronoun "she" by referring to prior mentions in the dialogue, while simultaneously detecting the corresponding visual action within the video. This intricate interplay between visual cues and linguistic context renders the task a formidable challenge in multimodal representation learning.

In recent years, transformer-based models have revolutionized natural language processing by delivering unprecedented performance in text understanding. Nevertheless, despite substantial advances in textual modeling, current approaches for interactive video conversation often extract visual features independently—typically via frozen 3D-CNNs trained on action recognition datasets—without sufficiently capturing the interplay between visual content and dialogue context. Liu et al. [26] have demonstrated that many of these models exhibit a pronounced bias towards text, thereby neglecting critical visual information. Such an imbalance can result in suboptimal performance, particularly when the visual modality carries decisive semantic cues.

To overcome these limitations, our work introduces **CIMT** (Cross-modal Interaction Multimodal Transformer), a comprehensive framework that tightly couples visual, textual, and auditory information through joint end-to-end training. Unlike traditional pipelines where each modality is processed in isolation, CIMT employs a two-stage visual encoding process: first, a 3D-CNN is used to capture localized temporal features, and then a transformer-based encoder refines these features by establishing long-range dependencies via self-attention. Concretely, if we denote the intermediate feature matrix

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



extracted from video segments as $F \in \mathbb{R}^{N \times d}$, the self-attention mechanism refines the representation as follows:

$$\mathrm{SA}(F) = \mathrm{softmax}\left(\frac{FW_Q(FW_K)^T}{\sqrt{d}}\right)FW_V,$$

where $W_Q$, $W_K$, and $W_V$ are learnable projection matrices. This formulation enables the network to capture nuanced relationships across temporal segments, ensuring that even subtle visual cues are effectively integrated with textual context.

The challenges inherent in this task can be summarized along four primary dimensions:

1. *Visual Feature Extraction:* Developing discriminative representations that encapsulate both spatial and temporal dynamics from video streams.
2. *Textual Representation:* Capturing the sequential and contextual semantics from dialogue history with high fidelity.
3. *Cross-Modal Fusion:* Seamlessly integrating heterogeneous signals—visual, textual, and auditory—into a coherent multimodal representation. This is particularly critical given that each modality may contribute complementary information.
4. *Natural Language Generation:* Synthesizing fluent and context-aware responses that accurately reflect the multimodal input.

In addressing these challenges, CIMT not only leverages well-established architectures but also introduces novel integration strategies that allow for mutual reinforcement between modalities. For instance, while conventional methods tend to pre-extract visual features, our approach learns visual representations in tandem with textual and audio cues, ensuring that the latent space captures cross-modal correlations more effectively. Such joint training is critical, as it enables the model to dynamically weigh the importance of each modality based on the context of the query.

Beyond technical novelty, the implications of enhanced interactive video conversation extend to a wide array of real-world applications. Autonomous vehicles may use such systems to provide real-time road assistance by interpreting complex traffic scenes [14]. Similarly, assistive technologies can empower visually impaired individuals by offering detailed scene descriptions, thereby facilitating better spatial awareness. Other applications include advanced surveillance systems, video summarization, and long-form video content navigation. Recent works [3,17,18,23,34] have largely focused on individual modalities; however, our integrated approach exemplifies the potential of cross-modal learning in tackling multifaceted dialog scenarios.

The comprehensive design of CIMT also includes a flexible architecture that supports the incorporation of additional modalities as required. For example, when audio signals are available, a dedicated audio encoder is integrated into the joint framework. The resulting multimodal feature vector is then used to generate responses or perform retrieval tasks. This unified design is governed by an overall loss function that balances the contributions of each modality, ensuring robustness even in scenarios where one or more modalities may be degraded or noisy.

The contributions of this work are summarized as follows:

- We introduce a novel unified framework, **CIMT**, for interactive video conversation that bridges the gap between visual and textual representations through joint end-to-end learning.
- Our method leverages a two-stage visual encoder that first extracts local temporal features via a 3D-CNN and then refines these features using transformer-based self-attention, thereby capturing long-range dependencies in dynamic scenes.
- We demonstrate the flexibility of our approach by seamlessly incorporating additional modalities such as audio, and we validate the effectiveness of our design through extensive experiments on both response generation and retrieval tasks.
- Comprehensive ablation studies illustrate that the integrated cross-modal learning strategy not only mitigates the text-bias observed in previous models but also leads to significant performance gains across various evaluation metrics.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



Overall, our study highlights the critical importance of balanced multimodal fusion in interactive video conversation tasks. By jointly optimizing the contributions of visual, textual, and auditory cues, CIMT sets a new benchmark for future research in cross-modal representation learning.

## 2. Related Work

The field of video and language understanding has evolved rapidly over the past decade, driven by the need for advanced human-computer interaction systems. Early work in video captioning [5,44,49] laid the groundwork for generating natural language descriptions of video content, while subsequent research in video question answering [19,20,31,48] has focused on aligning dynamic visual features with textual queries. Moreover, video dialog systems [1,13,24] have pushed the boundaries by requiring models to interpret not only static content but also the evolving conversational context over time.

A seminal contribution in this area is the Audio-Visual Scene Aware Dialog (AVSD) task introduced by Alamri et al. [1], which challenges systems to answer questions based on a short video clip accompanied by both audio signals and dialog history. This task encompasses two primary settings. In the discriminative approach, models rank a list of candidate responses [1,29] based on their relevance, whereas the generative approach trains decoders to produce responses in an auto-regressive manner [11,23]. Both paradigms require deep integration of heterogeneous modalities and have spurred extensive research into effective cross-modal representation.

The advent of transformer architectures [42] has revolutionized natural language processing by enabling the extraction of rich, context-aware linguistic representations. Models such as GPT2 [33], ELMO [32], and BERT [8] have set new standards in various NLP tasks by pre-training on vast unlabelled corpora and subsequently fine-tuning on specialized downstream applications. In our work, we employ a pre-trained BERT model to encode the input question and dialog history, ensuring that high-level semantic information is robustly captured from the textual modality.

Inspired by these successes, researchers have extended self-attention mechanisms to multimodal settings. A significant body of work now adapts transformer models to handle tasks such as image question answering [7,21,22,27,41] and image dialog [7], where the fusion of visual and textual tokens is critical. In the realm of video, models such as VideoBERT [39] and contrastive learning frameworks [38] have been proposed to capture both spatial and temporal dynamics. In parallel, video dialog systems such as VideoGPT2 [23] and PLATO [3,17] have emphasized generating coherent responses by leveraging cross-modal interactions.

Approaches in multimodal fusion generally fall into two categories: single-stream and two-stream architectures. Single-stream methods merge tokens from both video and text into a unified sequence that is processed by a single transformer network [22,29,37]. This approach allows for deep joint representation learning but may obscure modality-specific features. Conversely, two-stream methods process each modality independently using dedicated transformer encoders and later merge the representations via concatenation or cross-attention mechanisms [27,41]. For instance, Luo et al. [28] fuse visual and textual features extracted from modality-specific encoders through a cross-modal attention network, which has shown improved performance in video dialog tasks.

Despite these advances, several limitations persist. A common shortcoming in many existing methods is the reliance on pre-extracted visual features from 3D-CNNs that are not updated during training. This practice restricts the model's capacity to capture subtle temporal dynamics and adapt to the specific demands of the video dialog task. Liu et al. [26] highlighted that models overly dependent on static visual features often underutilize the rich, dynamic information available in video content. Such limitations motivate the need for frameworks that allow end-to-end optimization of visual feature extractors.

Our proposed framework, **CIMT** (Cross-modal Interaction Multimodal Transformer), builds upon and extends these earlier approaches. Unlike previous methods that maintain a frozen visual encoder, CIMT integrates a 3D-CNN whose parameters are updated jointly with the rest of the network. This

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



end-to-end training strategy ensures that the visual extractor adapts specifically to the demands of video dialog, capturing fine-grained temporal and spatial cues. In addition, CIMT incorporates a dedicated audio transformer-based encoder that processes auditory signals in parallel with visual and textual modalities. The outputs from these modality-specific encoders are then fused via a cross-modal encoder, which dynamically balances and integrates the contributions of each modality.

Other notable contributions in the literature further illustrate the trend towards enhanced multimodal fusion. Le et al. [17] introduced a multimodal transformer network augmented with query-attention, which selectively focuses on relevant regions of the video and corresponding dialog context. Zekang et al. [33] extended a pre-trained GPT2 model to jointly learn representations from audio, visual, and textual inputs through multi-task learning, demonstrating that cross-modal pre-training can significantly boost performance. Moreover, Cherian A. et al. [35] improved upon audio-visual transformers by incorporating a student-teacher learning paradigm, thus enhancing the extraction of complementary features across modalities.

It is also worth noting that many of these models are evaluated under both discriminative and generative frameworks. While discriminative models excel at ranking candidate responses by leveraging strong textual priors, generative models focus on synthesizing fluent, context-aware responses. However, the prevailing challenge remains the effective utilization of visual features, which are frequently relegated to a secondary role compared to textual information. This imbalance limits the overall potential of multimodal dialog systems, especially in scenarios where visual context plays a crucial role.

In summary, the literature on video and language understanding is rich with innovative approaches that span from early video captioning to sophisticated video dialog systems. The evolution of transformer-based architectures has enabled significant strides in capturing deep contextual relationships, and the ongoing shift towards joint multimodal training has opened new avenues for integrating visual, textual, and auditory data. Our proposed framework, CIMT, addresses the shortcomings of previous models by introducing an end-to-end trainable visual extractor and a dedicated audio encoder, thereby achieving a more balanced and effective fusion of modalities. This holistic integration paves the way for more robust and contextually sensitive video dialog systems, advancing the state-of-the-art in multimodal learning and human-computer interaction.

## 3. Proposed Methodology and Architecture

This section details our novel framework for the video-based dialog task. We describe the processing of individual modalities, the fusion mechanism, and the training and inference strategies, thereby presenting a comprehensive view of our proposed approach, **CIMT** (Cross-modal Interaction Multimodal Transformer).

### *3.1. Problem Definition and Task Setup*

Given an input video

$$V = (V_1, V_2, \ldots, V_i, \ldots, V_n),$$

where each $V_i$ represents the $i^{th}$ frame extracted from the video, a dialog history

$$DH_t = \Big(C, (Q_1, Ans_1), \cdots, (Q_{t-1}, Ans_{t-1})\Big),$$

with $C$ being the video caption and $(Q_{t-1}, Ans_{t-1})$ a past question-answer pair, along with an audio track $A$, the goal is to generate a response $R_t$ to a follow-up question $Q_t$. This is accomplished by conditioning on all available modalities—video $V$, audio $A$, dialog history $DH_{1:(t-1)}$, and the current question $Q_t$. Formally, the conditional probability is defined as:

$$P(R_t \mid V, A, DH_{1:(t-1)}, Q_t; \theta) = \prod_{j=0}^{t-1} P(R_j \mid V, A, DH_{1:j-1}, Q_t; \theta),$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



and the training objective is to minimize the cross-entropy loss:

$$\mathcal{L}(\theta) = -\log P(R_t \mid V, A, DH_{1:(t-1)}, Q_t; \theta),$$

where $\theta$ denotes the set of trainable parameters across the network.

### 3.2. Overall Model Architecture

Our framework, CIMT, is designed with a multi-stream architecture to independently process and then fuse textual, visual, and audio features. The major components of the system are as follows:

- **Text Encoder:** Processes the dialog history, captions, questions, and answers.
- **Visual Encoder:** Utilizes a 3D-CNN to extract spatiotemporal features from video frames.
- **Audio Encoder:** Extracts auditory cues using a dedicated convolutional network.
- **Cross-Modal Fusion Module:** Integrates the modal-specific embeddings using a cross-attention mechanism.
- **Decoder:** An auto-regressive module that generates responses based on the fused multi-modal representation.

In the following subsections, we describe each component in detail, including additional formulas that capture the interdependencies and refinement procedures within the model.

#### 3.2.1. Text Encoder

All textual inputs—including the dialog history $DH$, video caption $C$, current question $Q$, and previous answers $Ans$—are concatenated to form a single string. Following the approach in Devlin et al. [9], we tokenize the input using a WordPiece tokenizer [45] to produce a sequence of tokens

$$t = \{t_i \mid i \in [1, n]\},$$

where $n$ is the total number of tokens. A special *[CLS]* token is prepended, and *[SEP]* tokens are inserted between sentences. This tokenized sequence is then passed through a pre-trained BERT-based model to obtain a contextualized text embedding:

$$E_d = BERT(t), \quad E_d \in \mathbb{R}^{n \times d},$$

with $d$ representing the dimensionality of the hidden states.

#### 3.2.2. Visual Encoder

For visual processing, frames are uniformly sampled from the video $V$ at 16 frames per second and resized to $224 \times 224$ pixels. Denote the set of subsampled frames as

$$V_n = \{v_j \mid j \in [1, m]\}.$$

These frames are fed into a 3D convolutional neural network—specifically, the I3D network [4] pretrained on ImageNet—to extract temporal features:

$$f_v = I3D(V_n).$$

The raw feature maps $f_v$, extracted from multiple inception blocks (e.g., $Mixed_4$ and $Mixed_5$), have dimensions $m \times d$. A subsequent transformer-based visual encoder refines these features via a series of self-attention and feed-forward layers:

$$E_v = FFN(f_v) + MHA(f_v).$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



To further enhance temporal consistency, we introduce an auxiliary temporal smoothness loss:

$$\mathcal{L}_{temp} = \frac{1}{m-1} \sum_{j=1}^{m-1} \|f_v^{(j+1)} - f_v^{(j)}\|_2^2,$$

which encourages gradual transitions between consecutive frame representations.

#### 3.2.3. Audio Encoder

The audio modality is processed by first extracting $m$-dimensional features from the audio signal $A$ using a VGGish network [10]:

$$f_a = VGGish(A_m).$$

These initial features are then refined using a transformer-based encoder analogous to the visual branch:

$$E_a = FFN(f_a) + MHA(f_a).$$

To mitigate the effects of noise, we also compute the spectral energy of the audio features:

$$E_{audio}^{(j)} = \|f_a^{(j)}\|_2,$$

and include an audio regularization term:

$$\mathcal{L}_{audio} = \frac{1}{m} \sum_{j=1}^{m} \left(E_{audio}^{(j)} - \mu\right)^2,$$

where $\mu$ represents the expected mean energy, thus promoting a stable representation across the audio sequence.

#### 3.2.4. Cross-Modal Fusion Module

To achieve robust multimodal integration, we fuse the embeddings $E_v$, $E_a$, and $E_d$ using a cross-attention encoder. The embeddings are concatenated to form a unified sequence:

$$H_{\text{fuse}} = \left[E_v; E_a; E_d\right].$$

A series of $N = 6$ layers combining Multi-Head Attention (MHA) and Feed-Forward Networks (FFN) is then applied to generate the fused multimodal embedding:

$$h_{\text{embd}} = FFN(H_{\text{fuse}}) + MHA(H_{\text{fuse}}).$$

To further enhance cross-modal interactions, we segment $h_{\text{embd}}$ into modality-specific parts—denoted $h_v$, $h_a$, and $h_d$ for visual, audio, and textual modalities respectively—and minimize the inter-modal discrepancy via the regularization term:

$$\mathcal{L}_{inter} = \|h_v - h_a\|_2^2 + \|h_a - h_d\|_2^2 + \|h_d - h_v\|_2^2.$$

Finally, a non-linear projection refines the final embedding:

$$h_{\text{final}} = \sigma(W h_{\text{embd}} + b),$$

with $W$ and $b$ as learnable parameters and $\sigma$ representing a non-linear activation function (e.g., ReLU).

#### 3.2.5. Training and Inference Strategy

The training of CIMT is guided by two primary objectives, alongside the auxiliary losses introduced above.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



**Conditioned Masked Language Modeling (CMLM):** In this objective, 15% of the input text tokens are replaced by a special *[MASK]* token. The model is then trained to predict these masked tokens, conditioned on the fused multimodal representation $h_{\text{final}}$.

**Decoder Reconstructive Loss:** During decoding, the auto-regressive decoder generates the answer token-by-token. At each time step $i$, given the previously generated token $\hat{y}_i$ and the multimodal context $h_{\text{final}}$, the next token is predicted as:

$$\hat{y}_{i+1} = \arg\max P\left(y_{i+1} = y \mid \hat{y}_i, h_{\text{final}}\right).$$

To stabilize the training process, the total loss function incorporates the primary losses along with the auxiliary terms:

$$\mathcal{L}_{total} = \mathcal{L}_{CMLM} + \mathcal{L}_{decoder} + \lambda_{temp}\mathcal{L}_{temp} + \lambda_{audio}\mathcal{L}_{audio} + \lambda_{inter}\mathcal{L}_{inter},$$

where $\lambda_{temp}$, $\lambda_{audio}$, and $\lambda_{inter}$ are hyperparameters that control the influence of the temporal, audio, and inter-modal regularization terms, respectively.

During inference, the decoder generates the response in an auto-regressive manner, selecting at each step the token with the highest probability until an end-of-sequence token is produced. The combined effect of the end-to-end trained visual, audio, and textual encoders—along with the refined fusion mechanism—ensures that the generated responses are both contextually relevant and grounded in the rich, multimodal input data.

## 4. Experiments

In this section, we present a comprehensive evaluation of our proposed framework, **CIMT** (Cross-modal Interaction Multimodal Transformer), on the Audio-Visual Scene-Aware Dialog (AVSD) dataset [1]. We describe the dataset and evaluation metrics, detail our preprocessing and training procedures, and then provide an in-depth analysis of both generative and retrieval tasks. In addition, we examine the impact of varying the number of fine-tuned visual blocks, the number of sampled video frames, and the length of dialog history on the overall performance.

### *4.1. Dataset and Evaluation Metrics*

We evaluate our framework on the AVSD dataset [1], which comprises dialogs grounded in human-centric action videos and includes clips drawn from the Charades dataset [36]. Each dialog contains a video caption and 10 rounds of question-answer pairs related to the video events. The dataset is partitioned into 7,659 dialogs for training, 1,787 for validation, and 1,710 for testing. This diverse and challenging dataset enables the investigation of cross-modal interactions between visual, audio, and textual modalities.

For quantitative evaluation on the DSTC-test set, we adopt a suite of widely used natural language generation metrics. Specifically, we report scores for **BLEU** [30], **METEOR** [2], **ROUGE-L** [25], and **CIDEr** [43]. These metrics assess the fluency, relevance, and overall quality of the generated responses, with the test set containing a single ground truth answer per question.

### *4.2. Preprocessing*

**Dialog History:** For constructing the dialog input, we incorporate up to three turns of dialog history, restricting the total word count to 100 words. This window has proven sufficient to capture the essential context from previous interactions alongside the current question.

**Video-Audio Features:** In our experiments, visual features are extracted from the $\boldsymbol{Mixed}_{5c}$ and $\boldsymbol{Mixed}_{4c}$ layers. In order to maintain consistency with previous baselines [1], we pre-extracted visual representations using the I3D [4] network trained on the ImageNet dataset and used the 1024-dimensional output from the $\boldsymbol{Mixed}_{5c}$ layer for baseline comparisons. For the audio modality, we adopt pretrained

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



1024-dimensional features from the *VGGish* network [10]. It is important to note that the audio encoder is not fine-tuned on the AVSD dataset, thereby reflecting its generalization across domains.

### *4.3. Training*

Our training regime employs the Adam optimizer [15] with an initial learning rate of $5 \times 10^{-5}$ and a batch size of 64. The model is trained on 8 RTX-6000 GPUs, and we utilize early stopping to select the checkpoint that achieves the best performance on the validation set. In addition to the primary training loss, we also monitor auxiliary losses (e.g., temporal smoothness for video features and audio regularization) to ensure that all modalities contribute meaningfully to the final joint representation.

## 5. Results and Analysis

This section provides a detailed examination of the performance of **CIMT** across both generative and retrieval tasks. We present quantitative results, ablation studies, and insights into how different design choices—such as the inclusion of audio, the depth of fine-tuning in the visual encoder, the number of sampled frames, and the extent of dialog history—affect the overall system performance.

### *5.1. Results on the Generative Task*

We first evaluate our method in the generative setting, where the system is required to generate a free-form response given the video, audio, dialog history, and the current question. Table 1 summarizes the performance of various methods on the AVSD test set using standard metrics. In this table, methods marked with an asterisk (*) incorporate audio features, and those marked with a dagger (†) additionally use a summary input.

**Table 1.** Model performance on the AVSD test for the generative task. * includes Audio, † includes summary.

| Method | BLEU2↑ | BLEU3↑ | BLEU4↑ | METEOR↑ | ROGUE-L↑ | CIDEr↑ |
|---|---|---|---|---|---|---|
| DGR* (2021) | – | – | 0.357 | 0.267 | 0.553 | 1.004 |
| JST*† (2021) | – | – | 0.406 | 0.262 | 0.554 | 1.079 |
| VideoGPT2*† (2020) | 0.570 | 0.476 | 0.402 | 0.254 | 0.544 | 1.052 |
| MTN † (2019) | 0.242 | 0.174 | 0.135 | 0.165 | 0.365 | **1.366** |
| JMAN (2020)[6] | 0.521 | 0.413 | 0.334 | 0.239 | 0.533 | 0.941 |
| Le H. et al. (2021)[16] | 0.577 | 0.476 | 0.398 | 0.262 | 0.549 | 1.040 |
| TimeSformer*† (2022) [47] | 0.572 | 0.477 | 0.403 | 0.255 | 0.547 | 1.049 |
| Ours + Audio modality* | 0.587 | 0.483 | 0.401 | **0.271** | 0.565 | 1.155 |
| Ours | **0.592** | **0.493** | **0.415** | 0.269 | **0.569** | 1.159 |

Our approach surpasses several competitive baselines, including VideoGPT2 and JST, across most evaluation metrics. For instance, our model achieves improvements in BLEU2 (0.592 vs. 0.570), BLEU3 (0.493 vs. 0.476), and BLEU4 (0.415 vs. 0.406). Similarly, METEOR and ROUGE-L scores are higher for our method, reflecting the benefits of joint training in effectively integrating visual and textual cues. Although our CIDEr score is marginally lower than that of MTN, it is worth noting that MTN leverages additional textual information (a summary in addition to the caption and dialog history) and utilizes a larger model capacity.

These results strongly indicate that the end-to-end joint training paradigm in CIMT significantly enhances the model's ability to leverage visual features. With only a modest increase in computational cost, our framework achieves comparable or better results than deeper networks that do not explicitly model cross-modal interactions.

**Role of Audio Modality:** An interesting observation from Table 1 is that incorporating audio features (methods marked with *) does not yield a significant improvement in performance over using text alone. This outcome is likely attributable to the nature of the AVSD dataset, where the audio track primarily contains ambient sounds rather than dialog-specific cues. Nonetheless, including the audio

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



modality serves to enhance the generalization of the model and is expected to be more beneficial when applied to other datasets with richer audio content.

### *5.2. Results on the Retrieval Task*

To further assess the contribution of visual information, we evaluate CIMT in a retrieval setting. In this configuration, the model receives identical multimodal inputs as in the generative task but is required to retrieve the correct answer from a candidate pool by producing a ranked list. This evaluation directly measures the quality of the encoded representations without the influence of the decoder.

For this purpose, we designed a streamlined retrieval framework in which video embeddings (extracted via I3D) and text embeddings (derived from the BERT model) are concatenated and optimized using several objectives: the Masked Language Model loss ($L_{\text{mlm}}$), the Next Sentence Prediction loss ($L_{\text{nsp}}$), and a text-video alignment loss ($L_{\text{vta}}$). The retrieval model is trained in an end-to-end manner to capture the interplay between visual and textual features.

**Table 2.** Model performance on the AVSD dataset. $\text{XXX}_{\text{ft}}$ refers to fine-tuned models, $\text{XXX}_{\text{no-ft}}$ to non-fine-tuned models. ↑ indicates that higher scores are better, while ↓ indicates lower scores are preferable.

| Input | Text Encoder | Vid Encoder | ↑ MRR | ↑ R@1 | ↑ R@5 | ↑ R@10 | ↓ MR |
|---|---|---|---|---|---|---|---|
| DH | LSTM | - | 50.40 | 32.76 | 73.27 | 88.60 | 4.72 |
| | BERT | - | 69.71 | 56.93 | 86.18 | 92.93 | 5.07 |
| DH + V | LSTM | $\text{I3D}_{\text{no-ft}}$ | 53.41 | 36.22 | 75.86 | 89.79 | 4.41 |
| | LSTM | $\text{S3D}_{\text{no-ft}}$ | 53.57 | 36.49 | 75.64 | 89.82 | 4.45 |
| | LSTM | $\text{I3D}_{\text{ft}}$ | 54.28 | 37.12 | 76.62 | 90.23 | 4.33 |
| **DH + V (Ours)** | $\text{BERT}_{\text{ft}}$ | $\text{S3D}_{\text{no-ft}}$ | 71.32 | 59.51 | 86.92 | **95.22** | 4.89 |
| | $\text{BERT}_{\text{ft}}$ | $\text{S3D}_{\text{ft}}$ | **77.28** | **67.28** | **90.39** | 94.87 | **4.18** |

**Evaluation Metrics:** For the retrieval task, we report several standard metrics including R@1, R@5, R@10, Mean Rank (MR), and Mean Reciprocal Rank (MRR). The ideal outcome is to rank the correct answer at the top of the candidate list.

**Table 3.** Performance by fine-tuning different inception blocks from the visual encoder.

| | Retrieval Mode | | | | | Generative Mode | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Trained Inception Blocks** | ↑ MRR | ↑ R@1 | ↑ R@5 | ↑ R@10 | ↓ MR | ↑ BLEU2 | ↑ BLEU3 | ↑ BLEU4 | ↑ METEOR | ↑ ROUGE-L | ↑ CIDEr |
| $\text{S3D}_{\text{no-ft}}$ | 53.41 | 36.22 | 75.86 | 89.79 | 4.41 | 0.58 | 0.488 | 0.407 | 0.268 | 0.561 | 1.115 |
| $\text{S3D}_{\text{Mixed-5}}$ | **77.21** | **67.20** | 90.22 | **95.06** | **4.15** | **0.592** | 0.492 | 0.413 | 0.267 | 0.563 | 1.134 |
| $\text{S3D}_{\text{Mixed-4,Mixed-5}}$ | 76.88 | 66.72 | **90.39** | 94.77 | 4.48 | **0.592** | **0.493** | **0.415** | **0.269** | **0.569** | **1.159** |

**Performance of Language Encoders:** Table 2 shows that using BERT as the text encoder yields a substantial improvement over traditional LSTM-based encoders; for instance, BERT achieves an MRR of 69.71, which represents a 19% absolute gain compared to the LSTM encoder. This improvement underscores the advantage of transformer-based pre-training on large text corpora, which helps in generating rich contextualized representations. Notably, even with simpler language models such as LSTM, the joint training with a visual encoder leads to meaningful performance enhancements.

**Effect of Different Visual Encoders:** We experimented with various 3D-CNN architectures for visual feature extraction. Our results indicate that both I3D and S3D yield comparable MRR scores (53.41 vs. 53.57, respectively) when not fine-tuned, suggesting that the benefits arise primarily from the joint training procedure rather than the specific choice of visual backbone.

**Joint Training Efficacy:** When the fine-tuned BERT encoder is combined with a fine-tuned S3D visual encoder ($\text{BERT}_{\text{ft}} + \text{S3D}_{\text{ft}}$), the model achieves the highest performance across all metrics, with a notable

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



6% increase in MRR compared to models that only utilize textual information. This confirms that end-to-end joint training is effective in capturing the complex interactions between modalities.

### 5.3. Effect of the Number of Fine-Tuned Blocks

To assess the influence of fine-tuning depth within the visual encoder, we conducted experiments where different numbers of inception blocks in the S3D network were fine-tuned. The S3D architecture consists of three convolutional layers followed by five inception blocks. Table 3 reports the performance under various fine-tuning configurations in both the retrieval and generative settings. Our results indicate that fine-tuning additional inception blocks (e.g., both Mixed-4 and Mixed-5) leads to a consistent improvement in performance, suggesting that deeper adaptation of the visual features is beneficial for the task.

### 5.4. Impact of Video Frame Sampling

We further investigated the effect of varying the number of sampled video frames on model performance. The key question is: *How many frames are necessary to capture sufficient visual context for accurate response generation?* We experimented with sampling rates of 6, 16, 32, and 40 frames per video. As summarized in Table 4, the model shows significant performance gains when increasing the number of frames from 6 to 32. However, a slight performance drop is observed at 40 frames, which we attribute to redundancy since the pre-trained models are typically optimized for around 30 frames per sequence.

**Table 4.** Evaluation results on the AVSD test set as a function of the number of sampled video frames.

| Number of Sampled Frames | ↑ MRR | ↑ R@1 | ↑ R@5 | ↑ R@10 | ↓ MR |
|---|---|---|---|---|---|
| 6 | 46.38 | 31.15 | 64.70 | 78.87 | 8.46 |
| 16 | 74.90 | 64.11 | 89.19 | 94.47 | 4.63 |
| 32 | **77.28** | **67.92** | **90.22** | **95.06** | **4.15** |
| 40 | 77.21 | 66.20 | 89.62 | 94.82 | 4.46 |

**Table 5.** Impact of dialog history length on performance.

| Dialog Round | MRR | R@1 | R@5 | R@10 | Mean Rank |
|---|---|---|---|---|---|
| 1 | 59.57 | 46.39 | 75.70 | 87.71 | 4.49 |
| 3 | **81.66** | **71.65** | **94.82** | **98.81** | **1.82** |
| 5 | 77.21 | 67.20 | 89.62 | 94.82 | 4.46 |
| 10 | 62.52 | 52.96 | 74.46 | 78.77 | 18.34 |

### 5.5. Effect of Dialog History Length

To explore the influence of dialog context on model performance, we conducted experiments varying the length of the dialog history. We evaluated the model in different rounds: in Round 1 (with no prior dialog turns), Round 3 (with two previous turns), Round 5, and Round 10. As reported in Table 5, performance improves substantially from Round 1 (MRR of 59.57) to Round 3 (MRR of 81.66), indicating that the additional context from earlier turns greatly assists the model. However, performance degrades beyond Round 3, likely because the dialog becomes increasingly generic and less informative, thereby diminishing its utility for generating accurate responses.

In summary, our extensive experiments demonstrate that the proposed CIMT framework consistently outperforms existing methods in both generative and retrieval tasks. The ablation studies highlight the importance of jointly fine-tuning the visual encoder, carefully selecting the number of sampled frames, and incorporating an appropriate amount of dialog history. These findings underscore the significance of balanced multimodal integration and provide promising directions for future research in video-based dialog systems.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1



## 6. Conclusions and Future Directions

In this paper, we introduced a novel framework for video-based dialog tasks, termed **CIMT** (Cross-modal Interaction Multimodal Transformer). Our approach is centered on the end-to-end joint training of the visual encoder along with other modalities such as text and audio. By optimizing the learning process from the visual input concurrently with linguistic and auditory signals, our framework is able to generate a rich, balanced multimodal representation that significantly alleviates the bias toward textual information.

Our extensive experiments on both generative and retrieval tasks have demonstrated that the CIMT framework yields consistent performance improvements across a wide range of evaluation metrics, including BLEU, METEOR, ROUGE-L, and CIDEr for generative tasks, as well as MRR and Recall scores for retrieval tasks. Although our method incurs additional computational and memory costs, the substantial gains observed in our experiments reinforce the critical importance of fine-tuning the visual encoders in concert with textual and audio modalities. The improvements in performance validate our hypothesis that the joint learning of visual and textual features is indispensable for tackling the complexities inherent in video dialog tasks.

Looking ahead, several promising directions for future research emerge. First, we aim to extend our current work by pretraining the CIMT model on vast amounts of raw, unlabeled video data using self-supervised learning techniques. For instance, by designing auxiliary tasks such as masked frame prediction or temporal order verification, we can further enhance the multimodal representations.

In addition, we plan to explore the integration of reinforcement learning methods to refine the response generation process further by incorporating interactive user feedback. Such approaches could dynamically adjust the generated responses based on real-time performance metrics and user satisfaction scores. Moreover, extending our framework to accommodate additional modalities (e.g., motion flow, depth information) may provide a more comprehensive understanding of the scene dynamics, thus paving the way for applications in more complex tasks such as video captioning and video retrieval.

Furthermore, we are interested in evaluating the robustness and generalizability of CIMT by testing it on alternative datasets and in different dialog scenarios. This will help determine its adaptability across varied domains and identify potential areas for further enhancement. Overall, our work emphasizes that a balanced, joint learning strategy across visual, textual, and auditory inputs is crucial for advancing video-based dialog systems.

In conclusion, the proposed CIMT framework establishes a new benchmark for integrating multimodal information in video dialog tasks. Our findings strongly suggest that the end-to-end joint training paradigm not only improves performance but also offers a scalable path toward more sophisticated, context-aware multimodal systems. Future work will continue to build on these insights, striving to further bridge the gap between visual perception and language understanding in complex interactive environments.

## References

1. Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.
2. Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
3. Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*, 2019.
4. Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
5. Ming Chen, Yingming Li, Zhongfei Zhang, and Siyu Huang. Tvt: Two-view transformer network for video captioning. In *Asian Conference on Machine Learning*, pages 847–862. PMLR, 2018.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1




6. Yun-Wei Chu, Kuan-Yen Lin, Chao-Chun Hsu, and Lun-Wei Ku. Multi-step joint-modality attention network for scene-aware dialogue system. *arXiv preprint arXiv:2001.06206*, 2020.
7. Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
10. Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.
11. Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE, 2019.
12. Chiori Hori, Anoop Cherian, and Tim K Marks. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc8. In *DSTC7 at AAAI, 2019 Workshop*, 2019.
13. Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. Video dialog via multi-grained convolutional self-attention context networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465–474, 2019.
14. Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.
15. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
16. Hung Le, Nancy F Chen, and Steven CH Hoi. $c^3$: Compositional counterfactual constrastive learning for video-grounded dialogues. *arXiv preprint arXiv:2106.08914*, 2021.
17. Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*, 2019.
18. Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator. *arXiv preprint arXiv:2004.08299*, 2020.
19. Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.
20. Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019.
21. Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.
22. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
23. Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, Cheng Niu, and Jie Zhou. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *arXiv preprint arXiv:2002.00163*, 2020.
24. Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2476–2483, 2021.
25. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.
26. Aishan Liu, Huiyuan Xie, Xianglong Liu, Zixin Yin, and Shunchang Liu. Revisiting audio visual scene-aware dialog. *Neurocomputing*, 2022.
27. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1




28. Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.
29. Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer, 2020.
30. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
31. Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021.
32. Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
33. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
34. Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019.
35. Ankit P Shah, Shijie Geng, Peng Gao, Anoop Cherian, Takaaki Hori, Tim K Marks, Jonathan Le Roux, and Chiori Hori. Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. *arXiv preprint arXiv:2110.06894*, 2021.
36. Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.
37. Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
38. Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 3(5), 2019.
39. Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.
40. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
41. Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
42. 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.
43. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
44. Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018.
45. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
46. Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017.
47. Yoshihiro Yamazaki, Shota Orihashi, Ryo Masumura, Mihiro Uchida, and Akihiko Takashima. Audio visual scene-aware dialog generation with transformer-based video representations. *arXiv preprint arXiv:2202.09979*, 2022.
48. Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. Videoqa: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641, 2003.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1




49. Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018.
50. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553): 436–444, may 2015. http://doi.org/10.1038/nature14539.
51. 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/.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100: 101919, 2023.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
68. 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.
69. 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.
70. 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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1




71. 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.
72. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
73. 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.
74. 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.
75. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2025 doi:10.20944/preprints202502.1772.v1




90. 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.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.


**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.