MTChat: A Multimodal Time-Aware Dataset and Framework for Conversation

Anonymous ACL submission

Abstract

Understanding temporal dynamics is critical for applications ranging from conversations and multimedia content analysis to decision-making. However, time-aware datasets, particularly for conversations, are still limited, which narrows their scope and diminishes their complexity. To overcome these limitations, we introduce MTChat, a multimodal time-aware dialogue dataset that integrates linguistic, visual, and temporal elements in dialogue and persona memory. Based on MTChat, we design two time-sensitive tasks, Temporal Next Response Prediction (TNRP) and Temporal Grounding Memory Prediction (TGMP), utilizing implicit temporal cues and dynamic aspects to challenge model's temporal awareness. Furthermore, we present an innovative framework with an adaptive temporal module to integrate these multimodal streams and build interconnections effectively. The experimental results confirm that novel challenges of MTChat and effectiveness of our framework in multimodal time-sensitive scenarios. The codes are publicly available at Anonymous Link and MTChat is submitted to ARR system.

1 Introduction

Research on temporal awareness has attracted considerable interest subsequent to (Min et al., 2020) seminal work, which illuminated the temporal dynamics inherent in answers to questions. This temporal dimension is critical across various domains, such as financial decision-making, event outcomes, multimedia content analysis and perceptions of topics. To explore the temporal awareness of large language models (LLMs), several time-sensitive datasets have been developed for research purposes. Among these, the TimeQA (Chen et al., 2021) and SituatedQA (Zhang and Choi, 2021) datasets offer time-sensitive questions accompanied by free-text contexts extracted from WikiData (Vrandečić and Krötzsch, 2014). Additionally,

Figure 1: An example in multimodal time-sensitive scenarios: different dialogue responses from the user with temporal dynamic of dialogue and his memories.

the TEMPLAMA dataset (Dhingra et al., 2022) was constructed based on the temporal knowledge base. StreamingQA (Liska et al., 2022) was compiled from collections of news articles in the English WMT challenges spanning 2007 to 2020.

Considering temporal aspects in a multimodal dialogue dataset, common in real-world applications, is challenging. However, there is limited work addressing this problem. For previous datasets, firstly, they are confined to the task setting: QA tasks, and secondly, both the questions and contexts being free-text (only linguistic information). A recently proposed time-sensitive multimodal dataset for long video understanding, termed TimeIT (Ren et al., 2023). This dataset, while innovative, presents three primary limitations: 1) its concentration on QA tasks restricts broader application scope; 2) the explicit temporal markers in the videos fail to fully challenge the model's capabilities in temporal sensitivity to implicit temporal cues; and 3) the fixed response format "<timestamp_start> to <timestamp_end> seconds: <event_description>" simplifies the task by reducing the requirement for complex temporal reasoning.

Addressing the limitations found in current time-related datasets, we introduce MTChat, an innovative multimodal time-aware dialogue dataset.

Dataset	Knowledge Corpus	Samples	Time-Sensitive	Task	has Images
TempLama (Dhingra et al., 2022)	CustomNews	50.0k	YES	Question Answering	NO
TimeQA (Chen et al., 2021)	Wikipedia	41.2k	YES	Question Answering	NO
StreamingQA (Liska et al., 2022)	WMT07-20	138.0k	YES	Question Answering	NO
TempReason-L2L3 (Tan et al., 2023)	Wikipedia	49.0k	YES	Question Answering	NO
PhotoChat (Zang et al., 2021)	OpenImage V4	12.3k	NO	Dialogue	YES
MMDialog (Feng et al., 2022)	SocialMedia	1.1M	NO	Dialogue	YES
MTChat	Reddit	28.7k	YES	Dialogue	YES

Table 1: Related datasets overview, including free-text time-sensitive datasets and multimodal dialogue datasets.

Firstly, This dataset features a comprehensive data structure that integrates linguistic, visual, and temporal elements within its dialogues and persona memories, which directly addresses the limitations of the free-text time-sensitive data formats currently available. Secondly, MTChat offers various time-sensitive tasks: Temporal Next Response Prediction (TNRP) and Temporal Grounding Memory Prediction (TGMP). These tasks with temporal dynamic aspect are designed to make models aware of the impact of time and predict varying responses and grounding memories evolve significantly over time. The variety of task settings broadens the scope of research in time-sensitive domains. Thirdly, MTChat increases the complexity of the dataset by utilizing time as implicit cues. It skillfully employs the time order of dialogues and memories to demonstrate the influence of time on human cognition processes. Fig 1 depicts an example in multimodal time-sensitive scenarios.

Moreover, based on the tasks presented in MTChat, we propose a pioneering framework featuring an adaptive temporal module. This framework is designed to augment the model's capacity for integrating linguistic, visual, and temporal elements, thereby facilitating more coherent interconnections among them. Specifically, this adaptive temporal module is used to dynamically merge features based on their relevance, enhancing the coherence and efficacy of the integration.

Finally, we conducted experiments on MTChat using SBERT (Reimers and Gurevych, 2019) and CLIP (Radford et al., 2021) models, which demonstrated that MTChat poses novel challenges to the model in multimodal time-sensitive scenarios. Furthermore, we compared our framework with other methods of feature integration, proving that our framework can effectively and markedly enhance the model's capabilities in integrating multimodal streams with temporal awareness.

The main contributions of this work are summarised as:

- We create the first multimodal time-aware dialogue dataset contains numerous instances where both dialogue responses and the grounding memories evolve markedly over time.
- We offer various time-sensitive tasks: Temporal Next Response Prediction and Temporal Grounding Memory Prediction, extending the the research landscape in time-sensitive domains.
- We propose a innovative framework with an adaptive temporal module to enhance the model's capabilities in integrating multimodal streams with temporal awareness.
- We present experimental results that demonstrate MTChat dataset poses novel challenges, and that our framework surpasses other methods in feature integration.

2 Comparison with Existing Datasets

We start with a brief comparison of existing datasets, emphasizing multi-modal and time-aware strategies (see Table 1 for an overview).

Time-Sensitive QA Datasets Time-Sensitive Question Answering (TSQA) involves interpreting and responding to questions that are dependent on specific time points or intervals. We analyse a set of TSQA datasets (Dhingra et al., 2022; Chen et al., 2021; Liska et al., 2022; Tan et al., 2023), as shown in the upper part of Table 1. Currently, TSQA datasets typically use free-text form or knowledge graphs (KGs) and are structured as QA tasks. However, our work introduces the first multimodal time-aware dataset based on conversation. Similar to TSQA, we modify the time of dialogues, which affects the responses and the related grounding memory, thereby testing the model's ability to understand time.

MultiModal Dialogue Datasets Multimodal dialogue datasets generally comprise one or more images and multi-turn textual dialogues. As depicted in the lower half of Table 1, we analyse two representative datasets (Zang et al., 2021; Feng et al., 2022). These datasets are designed for models to interpret images and utterances within a dialogue framework and generate coherent responses. Our MTChat dataset, although drawing on the conversational structure and task, distinctively emphasizes the annotation and manipulation of time information. MTChat allows the model to acknowledge the influence of temporal dynamics on dialogue interaction and memory processes, demonstrating temporal awareness.

Time-Sensitive Video-Centric Dataset TimeIT (Ren et al., 2023) is a novel dataset focused on video-based instructions, encompassing a collection of long-video datasets annotated with timestamps. It requires models to describe video content across specified time intervals. The description follows a structured format, such as "<timestamp_start> to <timestamp_end> seconds: <event_description>". Ingeniously, our dataset integrates time of dialogues and memories, making model awareness of the time order of dialogue and memory significant influence on dialogue responses and memory recall. In contrast to TimeIT's tasks that directly answer timestamp and associated content, MTChat offers a more complex challenge with implicit time factor, pushing the boundaries of temporal understanding in multimodal dialogue models.

3 MTChat Dataset

Our dataset is built on the basis of MPChat (Ahn et al., 2023), a comprehensive multimodal persona-grounded dialogue dataset that includes both linguistic and visual components derived from episodic-memory-based personas. MPChat gathered from the social media platform Reddit, consists of memory image-sentence pairs and dialogue instances grounded on the speakers' multimodal memories.

A significant challenge is the ingenious integration of time information and multimodal dialogue, aiming to establish a multimodal time-aware dataset. Based on MPChat dataset, we develop a novel methodology that involves three primary steps: 1) Time annotations, 2) Constructing time-aware conversations, and 3) Memory annotations. These efforts achieve the creation of a pioneering multimodal time-aware dialogue dataset. MTChat breaks away from the limitations of current time-sensitive datasets confined to QA tasks, free-text formats and relying on explicit time information. We believe that our work fosters the development of more diverse time-sensitive datasets and advancing research toward achieving human-level temporal understanding in models.

3.1 Time Annotations

We converted the UTC strings in MPChat dataset into date format "yyyy/mm/dd" and incorporated this feature into both the dialogue and memory components. The dialogue in our dataset is structured as a triplet consisting of (dialogue context, dialogue image, dialogue time), while each memory of the speaker is similarly organized as a triplet (memory description context, memory image, memory time).

3.2 Time-Aware Conversations

In real-world scenarios, conversations can vary significantly based on the time they occur, even with similar contexts. For instance, as a high school student asked, "What is machine learning?", you might respond with no knowledge on the subject. However, after three years of studying machine learning at university, your response to the same conversation would be more detailed, potentially including discussions about deep learning and related topics.

Inspired by how the temporal order of conversation and memories influences human responses, we constructed conversational data with temporal orders:

- Later Stage Conversations: We used the original memories and conversations from the MPChat dataset, adding time annotations as described in Section 3.1. For instance, if you are a university student with three years of study in machine learning and are asked, "What is machine learning?", your response might include topics like deep learning.

- Early Stage Conversations: To simulate conversations from earlier times, we assumed there was no prior memory of the discussion topic. We used the context of the original conversations but removed the original responses. We then add new, earlier time annotations and responses. The newly created responses differ from the original ones and contain minimal information about the discussion topic due to the lack of relevant memory. For example, if you are a high school student asked, "What is machine learning?", you might respond with little to no knowledge on the subject.

Specifically, we utilized GPT-4 (Ouyang et al., 2022) to process a combination of inputs: the dialogue context, dialogue image, newly modified dialogue time, and speaker memories predating this new dialogue time. GPT-4 generated responses under the following guidelines: 1) responses could not exceed 40 words; 2) if the provided memories' topics significantly differed from the conversation, the response should indicate the speaker's lack of familiarity with the conversations topic; 3) if the provided memories and conversation topics were only slightly different, the response should reflect the speaker's intention to engage with and explore the conversation topic.

3.3 Memory Annotations

To gain a more precise understanding of the model's capabilities in temporal awareness, we align conversations with memory. For the memory component, we add time annotations as outlined in Section 3.1. Since the memories of the speakers are sourced from real users on Reddit, we avoid creating fabricated memories to preserve data authenticity. Additionally, we incorporate a "No Memory" category into the speaker's memory set. Structured similarly to existing memory triplets (memory description context, memory image, memory time), the "No Memory" category is assigned as the description context, indicating that there is no memory to align with the response. [1] This memory category is used to align early-stage conversations. We then synchronize the memory time with the conversation's time information.

3.4 Dataset Statistics

MTChat comprises 18,973 conversations and 25,877 users. We divided MTChat into training, validation, and test sets with 15,056, 1,994, and 1,923 conversations respectively. We analyzed the proportion of later stage conversations and early stage conversations, finding a ratio of 3:1. As well as later stage conversations with grounding memories (some later stage conversations lack grounding memory) and early stage conversations with "No Memory", resulting in a ratio of 2:1. Furthermore, to gain deeper insight into the time information within MTChat, we charted the distribution of times across conversations and memories in Fig 2.

[1]We also correlate "No Memory" with a plain white image as the memory image.

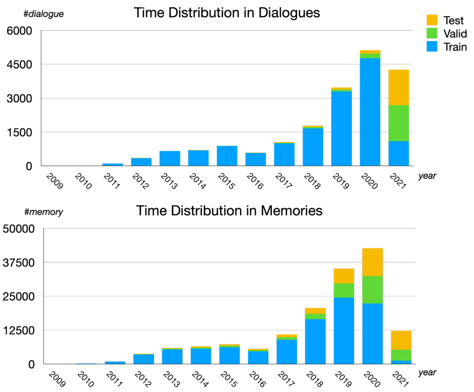

Figure 2: The distribution of times across conversations and memories in training, validation, and test set.

4 Task Definition

The MTChat datasets consist of N examples $\mathcal{D} = \{(d_n, r_n, \mathcal{M}_n)\}_{n=1}^{N}$, where $\forall n \in \{1, \ldots, N\}$ and each example contains a dialogue d_n, the speaker's response r_n to the dialogue d_n and a memory set $\mathcal{M}_n$ from the speaker. Each dialogue $d_n = (c^{d_n}, i^{d_n}, t^{d_n})$ encompasses the context c^{d_n} (context utterances), an associated image i^{d_n} and the time marking t^{d_n} (formatted as yyyy/mm/dd) when the dialogue occurred. The memory set for the speaker consists of m distinct memories $\mathcal{M}_n = \{M_{n_1}, \ldots, M_{n_m}\}$, where each memory $M_{n_m} = (c^{M_{n_m}}, i^{M_{n_m}}, t^{M_{n_m}})$ characterized by a description context $c^{M_{n_m}}$ (context utterances), an image $i^{M_{n_m}}$ and the time marking $t^{M_{n_m}}$ (formatted as yyyy/mm/dd) when the memory occurred.

4.1 Temporal Next Response Prediction

As illustrated in the Fig 3, Temporal Next Response Prediction (TNRP) is a retrieval task aimed at predicting the next response $\tilde{r}$ from a set R_c containing C response candidates based on the dialogue $d = (c^d, i^d, t^d)$ and the speaker's memories $\mathcal{M} = \{M_1 = (c^{M_1}, i^{M_1}, t^{M_1}), \ldots, M_m\}$. The response candidate set R_c comprises one ground truth and $C - 1$ distractor responses. It is essential to emphasize that, 1) Identical dialogue content and speaker memories can lead to vastly different responses depending on the time of the dialogue. 2) To intensify the task's complexity and underline the temporal factor's significance, our response candidate set includes responses from later-stage dialogue and early-stage dialogue. The remainder of the response candidates are randomly selected from other dialogues.

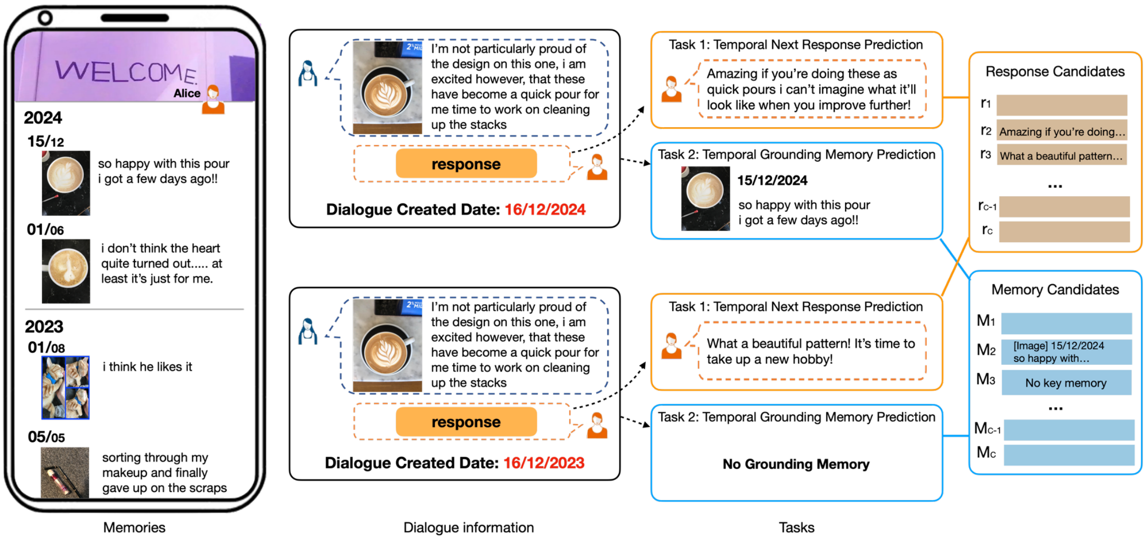

Figure 3: An overview of Temporal Next Response Prediction (TNRP) and Temporal Grounding Memory Prediction (TGMP) task. A user Alice's memories (i.e., four image-sentence-time triplet) and a same dialogue with different created date in the left part. Predicting responses and grounding memory from candidates is depended on the understanding temporal dynamic of dialogue and Alice's memories.

4.2 Temporal Grounding Memory Prediction

Temporal Grounding Memory Prediction (TGMP) task is also a retrieval task that requires predicting the most likely memory element from a set M_c containing C memory candidates based on a given dialogue $d = (c^d, i^d, t^d)$ and a remainder memory set (except grounding memory) before producing a response. The memory candidate set M_c comprises one grounding memory, one "No Memory" option and $C - 2$ distractor memories randomly selected from other speakers. As shown in Fig 3, time variations within the dialogue substantially influence the choice of the grounding memory. Specifically, when the time of the dialogue is later than the time of the grounding memory, suggesting the availability of memory related to the dialogue for supporting the speaker's response, the model is capable of predicting the grounding memory. Conversely, if the time of the dialogue is earlier than that of the grounding memory, indicating an absence of relevant dialogue memory, the model must predict a "No Memory" outcome.

In TGMP task, we deliberately exclude the speaker's response from the input. This decision is based on the consideration that potential responses of early-stage dialogue can vary significantly—ranging from disinterest in the dialogue topic to expressing a desire to learn. These different but reasonable responses could potentially confuse the model to predict grounding memory. The principal objective of the TGMP task is making model recognize the critical temporal order between dialogue and memory. By focusing on whether the model can identify the appropriate grounding memory or its absence for a given time information, we obtain a clearer measure of its temporal awareness capabilities.

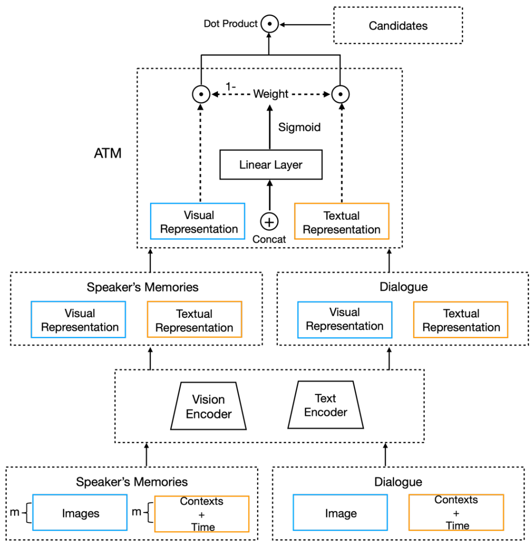

Figure 4: The architecture of our framework with Adaptive Temporal Module (ATM).

5 Framework

In this section, we present a framework to perform above retrieval tasks based on dialogue and memory. The inputs include dialogue d_n, the speaker's response r_n to the dialogue and a memory set $\mathcal{M}_n$. We define various encoders to process different modalities of data, fuse the extracted features, and achieve both the temporal next response prediction task and the temporal grounding memory prediction task. The architecture of our framework is shown in Fig 4.

Text Encoder In this study, we employ the text encoder to process textual components within tasks, specifically extracting representations of text and date information from dialogues, memories, and responses. For both dialogue and speaker memories, which may contain multiple entries, we first concatenate the text and date information for each entry. These concatenated strings are then further combined using a delimiter, forming unified representations. This method ensures comprehensive feature extraction by the text encoder, facilitating a more robust analysis of the textual data involved.

Vision Encoder Similarly, our vision encoder to extract features from images embedded in dialogues and memories. In datasets featuring speaker memories with multiple images, each image is processed by this vision encoder. The extracted features are then aggregated via mean-pool operation to create a consolidated visual representation. This methodology ensures a coherent integration of visual data, significantly enhancing the model's capacity to process multi-image features effectively.

Adaptive Temporal Module Following the extraction of textual and visual representations, it is essential to effectively integrate these features. As the inclusion of date information into textual representations can impact the correspondence between the text and vision features extracted by text encoder and vision encoder, we propose a method to dynamically balance these modalities to maintain the alignment of text and visual information within the same set of memories and dialogues. We introduce a module called the Adaptive Temporal Module (ATM), which is designed to be both simple and effective.

First, we concatenate the corresponding text and vision features and map them through a linear layer. Subsequently, a sigmoid layer is used to derive the weights for both text and vision features. These weights are then employed to merge the features based on their relevance, ensuring better alignment and integration. This approach allows for a more coherent and contextually appropriate fusion of multimodal features, enhancing the overall interpretative capability of the model.

6 Experiments

6.1 Experimental Setup

Baselines We consider the following baselines:

- SBERT+CLIP: We adopt a Transformer (Vaswani et al., 2017) initialized weights of SBERT (Reimers and Gurevych, 2019) and CLIP-ViT-B/32 vision model (Radford et al., 2021) as text encoder and vision encoder to represent text and image respectively. SBERT enhances the original BERT model (Devlin et al., 2018) to better handle similarity comparisons of dialogue and memory textual information. CLIP-ViT-B/32 vision model utilizes a Vision Transformer (ViT) (Dosovitskiy et al., 2020) with 32 attention heads, which enables it to capture more visual features.
- CLIP+CLIP: We utilize the CLIP-ViT-B/32 model (Radford et al., 2021) as text encoder (CLIP-ViT-B/32 text model) and vision encoder (CLIP-ViT-B/32 vision model). CLIP-ViT-B/32 text model employs a Transformer similar to GPT (Radford et al., 2018), designed specifically for processing textual input, making it ideally suited for textual analysis requirements.

Training We train both baselines and our framework for 5 epochs with a batch size of 8 on a NVIDIA Tesla V100 GPU. The model is optimized using Adam (Kingma and Ba, 2014) with a learning rate of $3e^{-6}$. For our framework, we incorporated the Adaptive Temporal Module (ATM) into two baselines to validate the effectiveness of framework. We set the number of speaker's memories is $m = 20$ and the number of candidates is $C = 100$.

Evaluation Metrics We assess the performance of the model on two tasks using Recall@1 and Mean Reciprocal Rank (MRR), which is the standard evaluation metrics on dialogue task (Lee et al., 2021; Feng et al., 2022; Ahn et al., 2023). Recall@1 quantifies the model's accuracy in retrieving the most relevant result as the top result for each query, effectively capturing the model's ability to return the most relevant result as the first item. MRR evaluates the average inverse ranking of the first relevant result across queries, providing insight into the model's overall retrieval quality.

6.2 Experimental Results

We conduct experiments of two baselines with and without our framework on time-sensitive tasks in MTChat. Besides, we define two input settings: one limited to dialogue, and the other encompassing both dialogue and speaker's memories. The

Model	Input Setting	TNRP		TGMP	
		R@1	MRR	R@1	MRR
SBERT+CLIP	d	58.26	69.90	49.17	63.38
	$d, \mathcal{M}$	61.32	72.55	58.90	73.53
SBERT+CLIP+ATM	d	58.70	70.26	52.04	65.35
	$d, \mathcal{M}$	61.55	72.78	60.22	74.26
CLIP+CLIP	d	66.20	76.34	56.91	70.64
	$d, \mathcal{M}$	68.75	78.66	67.25	80.50
CLIP+CLIP+ATM	d	66.97	76.96	57.35	71.04
	$d, \mathcal{M}$	**69.26**	**78.92**	**71.82**	**83.68**

Table 2: Results of the Temporal Next Response Prediction (TNRP) and Temporal Grounding Memory Prediction (TGMP) tasks. Symbols means: dialogue $d = (c^d, i^d, t^d)$ contains a context, an image and time information. A speaker's memory set $\mathcal{M} = \{M_1, \ldots, M_m\}$, where each memory $M = (c^M, i^M, t^M)$ characterized by a context, an image and time information.

Method	Temporal Grounding Memory Prediction	
	R@1	MRR
Attention Fusion	63.65	76.72
Linear Fusion	66.41	79.59
Mean-Pool Fusion	67.25	80.50
ATM (ours)	**71.82**	**83.68**

Table 3: Comparison of Adaptive Temporal Module (ATM) with other methods of feature integration on Temporal Grounding Memory Prediction task.

findings, as depicted in Table 2, reveal several insights: 1) MTChat poses challenges in terms of the multimodal temporal awareness capabilities of models. Despite TNRP and TGMP being retrieval tasks, both baselines exhibited inadequate performance on these time-sensitive challenges, achieving Recall@1 scores not surpassing 70. 2) Our framework is model-agnostic and effective, enhancing performance over both baselines. Note that in our TNRP task, where labels contain only the response text, the ATM module—which is tailored for multimodal fusion balance—yields a less pronounced improvement. 3) The temporal ordering of dialogue and memories plays a pivotal role in MTChat. In previous works with multimodal persona-grounded dialogue datasets (Zhong et al., 2020; Wen et al., 2021), the persona information serves as supplementary data to improve the accuracy of predicted dialogue responses. However, in MTChat, both persona memory and dialogue are essential components. They not only enhance the model's temporal awareness but also significantly influence performance. For instance, for CLIP+CLIP+ATM model on TGMP task, when the input lacked memory data, performance significantly dropped by 20.1% in Recall@1 and 15.1% in MRR.

In addition, to evaluate the performance of the Adaptive Temporal Module within our proposed system, we conducted a comparative analysis against other feature fusion methods:

- Attention Fusion: This method adeptly combines textual and temporal data with image features, employing an attention-based module to learn weights. This enhances the model's sensitivity to contextually significant features.
- Linear Fusion: Incorporates two linear layers optimized during training, enabling the model to directly learn the weights that most effectively combine textual and visual information.
- Mean-Pool Fusion: This approach computes the mean of the combined features, aggregating them from different modalities by simple averaging.

These methods were assessed using the CLIP+CLIP model on the Temporal Grounding Memory Prediction (TGMP) task. The findings in Table 3 indicate that the Adaptive Temporal Module surpassed other techniques, achieving improvements of 12.8%, 8.1%, and 6.4% in Recall@1, respectively. These results substantiate the superior capability of our framework to effectively enhance multimodal integration with temporal awareness.

6.3 Ablation Study

Model	Input Setting	TNRP		TGMP	
		R@1	MRR	R@1	MRR
CLIP+CLIP	$d, \mathcal{M}$(zero-shot)	39.49	52.07	54.59	61.27
	$d, \mathcal{M}$	68.75	78.66	67.25	80.50

Table 4: Ablation study of baseline CLIP+CLIP with zero-shot setting.

Zero-Shot Setting We explore the performance of the CLIP+CLIP model with a zero-shot setting on time-sensitive tasks. As shown in Table 4, the model demonstrates poor performance on MTChat time-sensitive tasks, showing the challenges inherent in MTChat and highlighting the urgent need for research into multimodal temporal awareness.

The Importance of Temporal Awareness This study highlights the critical role of temporal awareness in models. Utilizing the CLIP+CLIP model, we trained on datasets both with and without temporal data of dialogue and memories. These models were then evaluated on the Temporal Grounding Memory Prediction (TGMP) task. Our findings (see Table 5) reveal a marked difference in

Model	Input Setting	TGMP	
		R@1	MRR
CLIP+CLIP	d, $\mathcal{M}$(without time)	60.99	65.09
	d, $\mathcal{M}$	68.75	78.66

Table 5: Ablation study of baseline CLIP+CLIP without time information.

performance: models without temporal awareness demonstrated substantial difficulties in time-sensitive tasks. Conversely, models incorporating temporal awareness significantly excelled, achieving a 12.7% increase in recall@1 and a 20.8% improvement in MRR.

7 Related Work

Time-Sensitive Datasets In recent years, numerous contemporary time-sensitive datasets have been introduced, predominantly composed in the format of question answering and exclusively in textual form (Zhang and Choi, 2021; Chen et al., 2021; Tan et al., 2023; Liska et al., 2022; Wei et al., 2023). A significant contribution to this field is the SituatedQA dataset (Zhang and Choi, 2021), which emphasizes open-domain time-sensitive QA. It uniquely reannotates questions from the Natural Questions (NQ) (Kwiatkowski et al., 2019) and Wikidata (Vrandečić and Krötzsch, 2014) to reflect context dependency and variability in answers across different times and locations. Another notable dataset, TimeQA (Chen et al., 2021) comprises 20,000 questions and its hard version requiring models to infer from implicit temporal cues within text passages.In addition, the TempReason dataset (Tan et al., 2023) introduced by Tan presents a comprehensive framework for evaluating various aspects of temporal understanding. These datasets with the Open Book Question Answering (OBQA) setting, relying on external text to help language models (Izacard and Grave, 2020; Zaheer et al., 2020; Wei et al., 2021; Ouyang et al., 2022) in deducing correct answers.

There are also time-sensitive datasets structured around Closed Book Question Answering (CBQA), where the models must rely solely on the information within the question, without external text (Févry et al., 2020; Roberts et al., 2020; Dhingra et al., 2022).

Moreover, there are time-sensitive datasets based on knowledge graphs, such as TEQUILA (Jia et al., 2018), TimeQuestions (Jia et al., 2021), and CronQuestions (Saxena et al., 2021). These datasets feature more complex questions in natural language and require models to rank entities from a knowledge graph based on their temporal relevance.

Multimodal Dialogue Datasets Recently, several multimodal dialogue datasets have emerged, incorporating one or more images alongside multi-turn textual dialogues. Research in multimodal dialogue primarily aims to comprehend images and utterances within a context to either answer questions (Antol et al., 2015; Das et al., 2017; Seo et al., 2017; Kottur et al., 2019; Li et al., 2023) or generate natural responses(Meng et al., 2020; Zheng et al., 2021; Wang et al., 2021; Zang et al., 2021; Feng et al., 2022). (Mostafazadeh et al., 2017) introduced the IGC dataset, which consists of 4,000 dialogues, each featuring an image with a textual description as well as accompanying questions and responses centered around the image. (Shuster et al., 2018) released the ImageChat dataset, which significantly larger than IGC. As research into multimodal dialogue has deepened, datasets incorporating persona information have become increasingly prevalent. Datasets such as FoCusd (Jang et al., 2022), MPChat (Ahn et al., 2023), DuLeMon (Xu et al., 2022), and MSPD (Kwon et al., 2023) include dialogues paired with persona information, ranging from purely textual to multimodal personas. Correspondingly, models are designed to extract relevant personal information, which can significantly enhance the generation of dialogue responses.

8 Conclusion

In this work, we addressed the under-explored aspect of temporal awareness in multimodal scenarios by introducing the MTChat dataset and an accompanying framework with an adaptive temporal module. The MTChat dataset, with its integration of linguistic, visual, and temporal elements, offers a high-quality resource for advancing research in temporal reasoning. MTChat challenges models by requiring comprehension of temporal dynamics, thereby extending the scope of time-sensitive research beyond traditional QA formats. Our proposed adaptive temporal module has demonstrated substantial improvements in model performance, suggesting its potential applicability in various real-world scenarios.

9 Limitations

Despite its comprehensive structure and innovative tasks, the MTChat dataset and our framework present certain limitations and need attention for future development. For MTChat dataset, while the dataset significantly enhances the challenge of temporal reasoning by incorporating implicit temporal cues, it may still not fully capture the subtleties of real-world temporal dynamics, such as those influenced by cultural, historical, or personal contexts that affect human interactions. For our framework, future research should focus on refining this framework and exploring its scalability and adaptability across different domains and temporal challenges, aiming to further our understanding of time's impact on cognitive and decision-making processes.

10 Ethics Statement

In the development of the MTCHAT dataset, we have placed a high priority on privacy and adherence to ethical standards. We ensured that the images in the dataset do not contain identifiable features such as faces, license plates, or email addresses, and the text is free from offensive language. We urge users of the dataset to be aware of these inherent risks. Additionally, commercial use of our data is strictly limited to ensure compliance with the Reddit API Terms and to protect user privacy. The MTCHAT dataset is exclusively permitted for academic research purposes.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. *arXiv preprint arXiv:2305.17388*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. *arXiv preprint arXiv:2004.07202*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10803–10812.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1807–1810.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 792–802.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. 2023. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 707–719.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyun Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. *arXiv preprint arXiv:2107.08685*.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. *Advances in neural information processing systems*, 30.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. *arXiv preprint arXiv:2310.05157*.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*.

Michael JQ Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2021. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316*.

Appendix

A Detailed Prompt of GPT-4

Prompt of GPT-4 for generating response to early-stage conversation
Given the topic of a conversation, the context of the dialogue, and multiple memories of the speaker, please write a response to the conversation. It is important to note: 1. responses could not exceed 40 words. 2. If the memories are almost unrelated to the conversation, the generated response should reflect the speaker's lack of expertise in the conversation topic. If appropriate, consider incorporating the current content of the speaker's memories. 3. If the memories are related to the conversation, the response should express a willingness to try or explore it in the future. Conversation Topic: [topic] Dialogue Context: [context] Memories: [context]

Table 6: Detailed prompt of GPT-4 for generating response to early-stage conversation.

B Detailed Parameters

The parameter settings of Temporal Next Response Prediction (TNRP) and Temporal Grounding Memory Prediction (TGMP) tasks used in our paper are illustrated in Table 7.

Parameters	**TNRP**	**TGMP**
per_gpu_train_batch_size	8	8
per_gpu_eval_batch_size	1	4
num_train_epoch	5	5
max_num_candidates	100	100
max_num_image	20	20
image_size	224	224
learning_rate	$3e^{-6}$	$3e^{-6}$
weight_decay	0.05	0.05

Table 7: Detailed Parameters of Temporal Next Response Prediction (TNRP) and Temporal Grounding Memory Prediction (TGMP) tasks.