Temporal Dynamics Perception for Audio-Visual Question Answering

Mingxiang Wen[1], Yixin Wang[1], Xujian Zhao[1,†], Hongyou Chen[1], Yin Long[1], Bo Li[1], Peiquan Jin[2], Chunming Yang[1], Hui Zhang[1]
[1]Southwest University of Science and Technology, Mianyang, China
[2]University of Science and Technology of China, Hefei, China
Email: swustwmx@mails.swust.edu.cn; yixinwang2000@outlook.com; jasonzhaoxj@swust.edu.cn; chy2019@foxmail.com; yinlong@swust.edu.cn; libo@swust.edu.cn; jpq@ustc.edu.cn; yangchunming@swust.edu.cn; zhanghui@swust.edu.cn

***Abstract*—The Audio-Visual Question Answering (AVQA) task aims to mine temporal dynamic information in videos and perform comprehensive spatio-temporal reasoning to answer given questions accurately. Although pre-trained models have achieved significant success in audio-visual question-answering tasks, they still face several challenges. On the one hand, the length of long-term videos makes it difficult for image-based and audio-based pre-trained models to capture temporal dynamic information effectively. On the other hand, the textual semantic gap between the question and the original text of the text-based pre-trained models limits the prediction results. To address the above challenges, we propose a Temporal Dynamics Perception Model (TDPM) to capture temporal dynamic dependencies in videos through a visual temporal alignment module and an audio temporal alignment module. Those temporal alignment modules are constructed via a language-guided auto-regressive task and can combine historical clues of videos and language guidance to predict the future state of videos. Also, TDPM mitigates the semantic gap between the original text of the question and the pre-trained models through a textual semantic alignment module. Experimental results on the MUSIC-AVQA dataset demonstrate the effectiveness of our proposed model. It outperforms all the compared models in terms of average accuracy, with a maximum improvement of 17.02% and an average improvement of 10.86%.**

***Index Terms*—Audio-Visual Question Answering (AVQA), Temporal Dynamics Modeling, Spatio-Temporal Reasoning, Textual Semantic Alignment, MUSIC-AVQA Dataset**

I. Introduction

Audio-Visual Question Answering (AVQA) aims to model temporal dependencies in videos and perform complex spatio-temporal reasoning to answer natural language questions. With the rapid development of pre-trained models, AVQA has attracted increasing attention for building interactive AI systems that integrate language with dynamic audio-visual scenes [1]. Despite recent progress, AVQA remains challenging due to the need for comprehensive multimodal understanding and reasoning.

Recent methods [2]–[5] leverage pre-training to enhance AVQA. Li et al. [4] introduced the MUSIC-AVQA dataset and a spatio-temporal grounding model. Li et al. [3] proposed PSTP-Net, which selects question-relevant temporal segments and spatial regions. Li et al. [5] further explored fine-grained object features with an object-aware adaptive-positivity learning strategy. These approaches typically rely on pre-trained models for feature extraction followed by fine-tuning. However, applying such models to AVQA still faces key challenges.

Challenge 1: Image- and audio-based pre-trained models cannot effectively capture temporal dynamics. Although models such as CLIP [6], ViT [7], and VGGish [8] achieve strong performance, they lack the ability to model long-term temporal dependencies in videos, limiting AVQA performance (Fig. 1).

Challenge 2: A semantic gap exists between question-answer pairs and the declarative text used in pre-training. For example, CLIP uses descriptive sentences (e.g., "A photo of a dog."), whereas AVQA involves interrogative forms (e.g., "What is in the photo?"), which hinders accurate prediction.

To address these issues, we propose a Temporal Dynamics Perception Model (TDPM) to capture temporal dependencies and cross-modal semantics. For **Challenge 1**, we design Visual and Audio Temporal Alignment Modules to model temporal dynamics. These modules employ auto-regressive mechanisms to predict future representations from historical information and language guidance, optimized via future prediction loss. For **Challenge 2**, we introduce a Textual Semantic Alignment Module. A prompt template converts question-answer pairs into declarative sentences (e.g., transforming "Where is the first sounding instrument?" + "Right" into "The right instrument in the video sounding first."), reducing the semantic gap. The aligned text is then fused with audio-visual features for answer prediction. Experiments demonstrate that TDPM effectively improves temporal perception and AVQA performance.

Briefly, our contributions are as follows:

- We propose TDPM, which captures temporal dynamics via auto-regressive visual and audio temporal alignment modules, enhancing spatio-temporal reasoning.
- We design a textual semantic alignment module that reduces the gap between question-answer pairs and pre-training text while improving key information extraction.

†Corresponding Author.

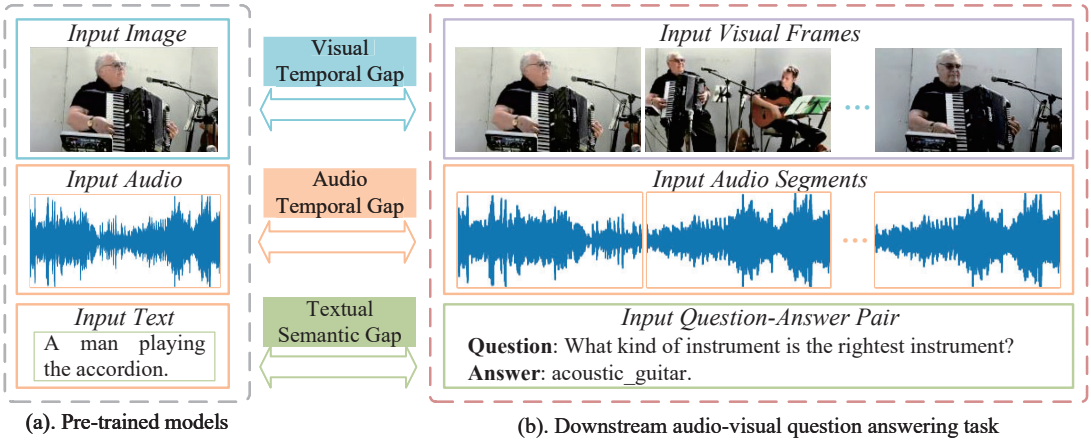

Fig. 1. Illustration of the temporal and textual semantic gaps between pre-trained models and downstream AVQA tasks.

- Extensive experiments on MUSIC-AVQA show that TDPM outperforms all baselines on audio, visual, and audio-visual tasks, achieving 76.63% average accuracy, with up to 17.02% and 10.86% improvements in maximum and average accuracy, respectively.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents TDPM. Section IV reports experiments, and Section V concludes the paper.

II. Related Works

A. Audio-Visual Representation Learning

Motivated by human multisensory perception, audio-visual scene analysis has advanced through leveraging semantic, spatial, and temporal consistency between modalities [9], [10]. Key applications include sound source localization [11], [12], audio-visual segmentation [13], [14], video parsing [15], [16], and audio-visual question answering (AVQA) [3], [4].

Sound source localization and segmentation aim to identify visual regions associated with audio. Hu et al. [11] introduced a heterogeneous analysis module with progressive training to handle scene complexity, while Liu et al. [14] proposed an encoder-cue-decoder framework with a Semantic-aware Audio Prompt for zero/few-shot localization, effectively bridging the audio-visual semantic gap.

B. Audio-Visual Question Answering

AVQA integrates audio, vision, and language for spatio-temporal reasoning [2]–[5]. Li et al. [4] introduced the MUSIC-AVQA dataset and the ST-AVQA model, establishing a spatio-temporal grounding benchmark. PSTP-Net [3] iteratively selects relevant temporal segments and regions guided by questions. Jiang and Yin [17] enhanced visual features using question topic words, while Li et al. [5] employed object-level features with an adaptive-positivity strategy to align objects with questions and audio. Lin et al. [18] further explored parameter-efficient fine-tuning of pre-trained models.

Despite progress, adapting large pre-trained models to long-form video reasoning remains challenging. We propose TPDM to capture temporal dynamics and mitigate domain gaps in AVQA.

C. Pre-Trained Knowledge Transfer

Large pre-trained models have driven state-of-the-art performance, spurring research on efficient knowledge transfer. Full-parameter fine-tuning [19] is often infeasible at scale, prompting lightweight alternatives such as partial parameter freezing [20] or adapter-based tuning within frozen backbones [18], [21]–[23].

In contrast, our method transfers prior knowledge while simultaneously alleviating temporal misalignment between pre-trained models and AVQA via visual-audio auto-regressive modeling.

III. Design of TDPM

To address the aforementioned challenges, we propose a Temporal Dynamics Perception Model (TDPM) that captures temporal dependencies and cross-modal semantics for AVQA (Fig. 2).

A. Task Definition

Given a question $\mathcal{Q}$, visual stream $\mathcal{V}$, audio stream $\mathcal{A}$, and candidate answers $\mathcal{A}_{all}$, AVQA aims to predict the correct answer:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}\in\mathcal{A}_{all}} \mathcal{F}_\theta(\mathbf{a} \mid \mathcal{V}, \mathcal{A}, \mathcal{Q}) \tag{1}$$

B. Data Representation

We divide the audio-visual stream into T non-overlapping 1-second segments $\{v_t, a_t\}_{t=1}^{T}$.

Visual. Frames sampled from each v_t are encoded by the frozen CLIP vision encoder [6], yielding $F_v \in \mathbb{R}^{T\times D}$.

Audio. Each a_t is encoded by VGGish [8], producing $F_a \in \mathbb{R}^{T\times D}$.

Text. Question-answer pairs are converted into declarative sentences via a prompt template and encoded by CLIP to obtain $F_t \in \mathbb{R}^{D}$.

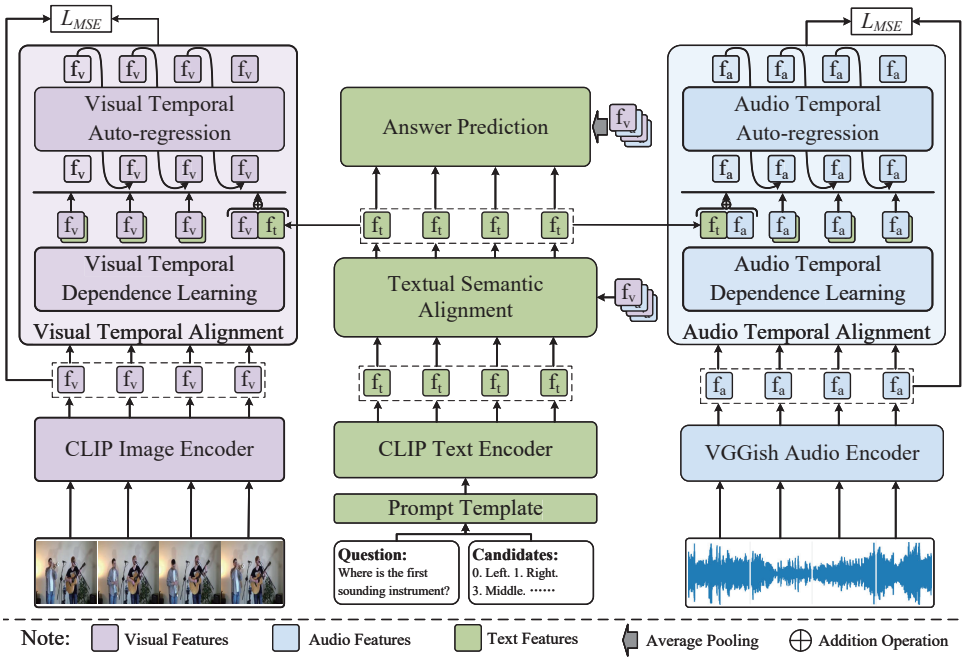

Fig. 2. Overview of TDPM, including Textual Semantic Alignment, Visual Temporal Alignment, Audio Temporal Alignment, and Answer Prediction.

C. Textual Semantic Alignment

To reduce the semantic gap, we design a Textual Semantic Alignment Module (TSAM). A prompt template converts QA pairs into declarative sentences (e.g., transforming questions into descriptive forms conditioned on answers). The text features interact with audio-visual features via multi-head attention:

$$F_{va} = \{f_v^1, f_a^1, ..., f_v^T, f_a^T\} \tag{2}$$

$$\hat{F}_t = F_t + \text{FFN}(\text{MultiHead}(F_t, F_{va}, F_{va})) \tag{3}$$

D. Visual Temporal Alignment

To model temporal dependencies, we introduce a Visual Temporal Alignment Module (VTAM):

$$\hat{F}_v = \text{FFN}(\text{MultiHead}(F_v, F_v, F_v)) \tag{4}$$

An auto-regressive mechanism predicts future frames using historical information and textual guidance:

$$F'_v = \hat{F}_v + \hat{F}_t \tag{5}$$

$$F''_v = \text{MaskMultiHead}(F_v, F_v, F_v) \tag{6}$$

$$\tilde{F}_v = \text{FFN}(\text{MultiHead}(F''_v, F'_v, F'_v)) \tag{7}$$

E. Audio Temporal Alignment

Similarly, the Audio Temporal Alignment Module (ATAM) models temporal dependencies in audio:

$$\hat{F}_a = \text{FFN}(\text{MultiHead}(F_a, F_a, F_a)) \tag{8}$$

$$F'_a = \hat{F}_a + \hat{F}_t, \quad F''_a = \text{MaskMultiHead}(F_a, F_a, F_a) \tag{9}$$

$$\tilde{F}_a = \text{FFN}(\text{MultiHead}(F''_a, F'_a, F'_a)) \tag{10}$$

F. Answer Prediction

We fuse updated features for prediction. Visual and audio features are interleaved and temporally pooled:

$$\hat{F_{va}} = \{\hat{f}_v^1, \hat{f}_a^1, ..., \hat{f}_v^T, \hat{f}_a^T\}, \quad \bar{F_{va}} = \text{AvgPool}(\hat{F_{va}}) \tag{11}$$

The answer score is computed via cosine similarity:

$$\hat{y} = \text{softmax}(\hat{F}_t \cdot \bar{F_{va}}^T) \tag{12}$$

G. Training Objective

The model is optimized using hinge loss with an auxiliary temporal prediction loss:

$$\mathcal{L}_{MSE} = \sum_{i=1}^{T} ||f_v^i - \tilde{f}_v^i||_2^2 + \sum_{i=1}^{T} ||f_a^i - \tilde{f}_a^i||_2^2 \tag{13}$$

$$\mathcal{L} = \mathcal{L}_{Hinge}(\hat{y}, y) + \gamma \cdot \mathcal{L}_{MSE} \tag{14}$$

During inference, only TSAM and the answer prediction module are used.

IV. Experiments

A. Dataset

We evaluate our method on two benchmarks: MUSIC-AVQA [4] and MUSIC-AVQA-Real.

MUSIC-AVQA [4]. This dataset targets AVQA in music performance scenarios, requiring temporal perception and spatio-temporal reasoning. It contains 9,288 videos (7,422 real and 1,867 synthetic) and 45,867 QA pairs, split into 32,087/4,595/9,185 for training/validation/testing. The QA pairs are designed with 33 templates across 9 types, including audio (*Counting, Comparative*), visual (*Counting, Localization*), and

audio-visual (*Existential, Localization, Counting, Comparative, Temporal*) questions.

MUSIC-AVQA-Real. This subset contains only real videos and corresponding QA pairs from MUSIC-AVQA, aiming to evaluate model performance in real-world (noise-free) scenarios and assess robustness by comparison with the full dataset.

For all datasets, following prior works [3]–[5], we use Accuracy (%) as the evaluation metric.

B. Implementation Details

For the visual stream, we divide videos into 1-second segments and sample frames at 1 fps. Frozen CLIP (ViT-B/32) is used as the visual encoder to extract 512-dimensional features. For the audio stream, we sample audio at 16 kHz, convert it into spectrograms, and use VGGish pre-trained on AudioSet to extract 128-dimensional features, which are then projected to 512 dimensions through a linear layer. For text, we convert question-answer pairs into declarative sentences via a prompt template and encode them with the CLIP text encoder to obtain 512-dimensional features. We train the model in PyTorch using Adam [30] with an initial learning rate of $1e$-4, decayed by 0.5 every 10 epochs. The batch size is 128, and training lasts 50 epochs on a single NVIDIA GeForce RTX 4060Ti 16G. We use the *thop* library to compute parameters and FLOPs.

C. Results

To evaluate TDPM, we compare it with 10 existing methods on MUSIC-AVQA-Real and MUSIC-AVQA, including ST-AVQA [4], PSTP-Net [3], QACL [2], and APL [5]. All comparison results are reproduced from official implementations. As shown in Table I and Table II, TDPM achieves the best average accuracy on both datasets. Compared with the strongest baseline APL, TDPM improves performance by 3.02% on MUSIC-AVQA-Real and 2.77% on MUSIC-AVQA, demonstrating stronger spatio-temporal reasoning ability and better robustness to noisy scenes.

A more detailed analysis on MUSIC-AVQA further confirms the superiority of TDPM. As shown in Table II, TDPM consistently outperforms all baselines, with clear gains on *Visual* and *Audio-Visual* questions, especially in *Localization*, *Counting*, *Comparative*, and *Temporal* tasks. Compared with APL [5], TDPM improves visual and audio-visual question accuracy by 6.03% (85.76% vs. 79.73%) and 2.36% (72.45% vs. 70.09%), respectively. In particular, for visual questions, TDPM achieves gains of 2.68% in *Counting* (79.78% vs. 77.11%) and 9.30% in *Localization* (91.59% vs. 82.29%). For audio-visual questions, it improves *Comparative* and *Temporal* tasks by 6.52% (73.04% vs. 66.52%), 1.64% (64.04% vs. 62.40%), and 11.31% (76.03% vs. 64.72%), respectively. These results verify the effectiveness of TDPM in temporal dependency modeling and complex spatio-temporal reasoning.

Table III compares TDPM with ST-AVQA [4], PSTP-Net [3], QAGL [2], and APL [5] in terms of trainable parameters and FLOPs. Compared with ST-AVQA, TDPM achieves higher accuracy with fewer parameters and lower FLOPs. Although TDPM has no advantage over PSTP-Net in parameter count, it attains better performance at lower computational cost. TDPM also surpasses QAGL and APL in both accuracy and parameter efficiency, with substantially fewer trainable parameters. Its FLOPs are higher, mainly due to additional matrix operations. Overall, TDPM achieves strong performance with relatively low cost, demonstrating both effectiveness and efficiency.

In summary, TDPM substantially improves over existing methods and provides a more effective solution for temporal dependency perception and audio-visual scene understanding.

D. Ablation Study

We first investigate the effectiveness of each component in TDPM, including PT, TSAM, VTAM, and ATAM, by removing them individually and reevaluating the model. As shown in Table IV, removing any component leads to a performance drop, validating the contribution of each module.

- **TDPM w/o. all.** To verify that the performance gain comes from the proposed modules, we remove all modules and retain only the input features with simple concatenation for fusion. As shown in Table IV, the performance drops sharply from 76.63% to 25.50%, demonstrating the necessity of the overall design.
- **TDPM w/o. PT.** The Prompt Template is designed to reduce the semantic gap between question-answer pairs and the original pre-training text. Without PT, the accuracy decreases to 75.59%, showing a 1.04% drop.
- **TDPM w/o. TSAM.** TSAM promotes cross-modal interaction and refines textual features. Removing it reduces performance by 0.83% compared with TDPM.
- **TDPM w/o. VTAM.** VTAM models temporal dependencies in the visual stream. Without it, the performance declines markedly to 72.38%, indicating the importance of visual temporal modeling.
- **TDPM w/o. ATAM.** ATAM captures temporal dependencies in segmented audio streams. Removing it lowers the accuracy from 76.63% to 74.99%, confirming its effectiveness.

Overall, removing any component consistently degrades performance, showing that all modules are necessary and that TDPM achieves the best results only when all components are included.

In addition, we also explore the impact of different hyperparameter configurations on the performance of TDPM.

- **Effects of Balance Factor γ.** The parameter γ is the balance factor used to balance the weight between Hinge loss and MSE loss. In this experiment, we explore the impact of different γ values on TDPM performance from 0.0 to 0.6. As shown in Table V, we can observe that when γ=0.1, the best performance is achieved.
- **Effects of Hyperparameters.** We tried different hyperparameters for the Textual Semantic Alignment Module, Visual Temporal Alignment Module, and Audio Temporal

TABLE I
ACCURACY (%) ON THE MUSIC-AVQA-REAL DATASET. EACH QUESTION TYPE'S BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINED.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
GRU [24]	74.55	66.80	72.89	68.32	73.46	70.83	84.25	65.75	62.52	62.12	65.83	68.43	69.75
HME [25]	77.47	64.75	74.74	66.45	67.97	67.19	81.66	66.70	56.44	63.27	60.03	66.34	67.88
MCAN [26]	80.49	50.82	74.12	71.87	71.69	71.78	82.90	69.35	55.59	57.13	60.50	65.71	68.73
PASC [27]	77.58	63.93	74.65	68.41	69.25	68.82	78.85	68.66	55.87	61.91	60.82	65.66	67.94
AVSD [28]	74.10	64.75	72.10	65.79	75.51	70.54	82.11	66.07	62.80	63.27	65.83	68.27	69.50
Pano-AVQA [29]	76.91	65.57	74.47	69.25	77.18	73.12	82.56	67.02	62.09	64.93	67.55	69.12	71.08
ST-AVQA [4]	80.49	63.11	76.76	72.71	76.69	74.65	82.45	65.77	71.16	64.72	70.38	71.05	72.95
PSTP-Net [3]	75.32	**76.67**	75.42	77.29	62.04	75.03	73.77	72.61	75.96	**75.05**	72.91	73.86	74.07
QAGL [2]	84.75	67.62	81.07	**77.38**	78.55	77.95	**85.83**	75.19	69.02	64.20	68.97	72.91	75.60
APL [5]	**84.75**	67.62	**81.07**	74.77	81.39	**78.00**	84.03	76.99	69.73	65.04	68.03	73.10	75.72
TDPM	84.53	67.21	80.81	77.29	**91.77**	**84.36**	84.36	**77.20**	**78.64**	61.39	**77.27**	**75.33**	**78.74**

TABLE II
ACCURACY (%) ON THE MUSIC-AVQA DATASET.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
GRU [24]	72.21	66.89	70.24	67.72	70.11	68.93	81.71	62.64	59.44	61.88	60.07	65.18	67.07
HME [25]	74.76	63.56	70.61	67.97	69.46	68.76	80.30	63.19	53.18	62.69	59.83	64.05	66.45
MCAN [26]	77.50	55.24	69.25	71.56	70.93	71.24	80.40	64.91	54.48	57.22	47.57	61.58	65.49
PASC [27]	75.64	66.06	72.09	68.64	69.79	69.22	77.59	63.42	55.02	61.17	59.47	63.52	66.54
AVSD [28]	72.41	61.90	68.52	67.39	74.19	70.83	81.61	63.89	58.79	61.52	61.41	65.49	67.44
Pano-AVQA [29]	74.36	64.56	70.73	69.39	75.65	72.56	81.21	64.91	59.33	64.22	63.23	66.64	68.93
ST-AVQA [4]	76.89	67.00	73.25	72.60	75.76	74.19	81.98	68.54	64.35	63.94	65.57	68.92	71.08
PSTP-Net [3]	80.24	58.92	72.38	75.19	79.27	77.25	**83.40**	70.83	64.02	61.67	67.40	69.51	72.07
QAGL [2]	82.99	**71.04**	**78.58**	**80.12**	77.88	78.89	82.29	72.73	62.83	63.40	64.36	69.43	73.58
APL [5]	81.15	69.19	76.97	77.11	82.29	79.73	81.78	**73.75**	66.52	62.40	64.72	70.09	73.86
TDPM	82.10	65.99	76.16	79.78	**91.59**	**85.76**	80.87	73.04	**73.04**	**64.04**	**76.03**	**72.45**	**76.63**

TABLE III
TRAINING PARAMETERS AND FLOPS OF TDMP.

Method	Training Param (M)	FLOPs (G)	Accuracy (%)
ST-AVQA [4]	18.48	3.19	71.08
PSTP-Net [3]	4.30	1.22	72.07
QAGL [2]	22.63	0.66	73.58
APL [5]	13.62	0.33	73.86
TPDM	8.90	1.11	76.63

TABLE IV
ABLATION STUDY ON THE DIFFERENT MODULES OF TDPM.

	Method	Average Accuracy (%)			
		Audio	Visual	Audio-Visual	All
1	TDPM w/o. all	19.06	24.07	28.22	25.50
2	TDPM w/o. PT	72.32	85.05	72.14	75.59
3	TDPM w/o. TSAM	74.61	84.43	72.08	75.80
4	TDPM w/o. VTAM	72.50	78.03	69.66	72.38
5	TDPM w/o. ATAM	73.12	84.68	70.98	74.99
6	TDPM	76.16	85.76	72.45	76.63

Alignment Module, including different layers of the Textual Semantic Alignment Module, different layers of the temporal dependence learning module, and temporal auto-regression module in the Temporal Alignment Modules (Visual & Audio). The results on the MUSIC-AVQA dataset are shown in Table VI. We adjusted the number of

TABLE V
IMPACT OF THE BALANCE FACTOR γ IN THE LOSS FUNCTION.

	Balancing Factor γ	Average Accuracy (%)			
		Audio	Visual	Audio-Visual	All
1	γ = 0.0	74.80	85.14	72.10	76.03
2	γ = **0.1**	**76.16**	**85.76**	**72.45**	**76.63**
3	γ = 0.2	74.60	84.89	72.41	76.12
4	γ = 0.3	74.74	84.93	72.43	76.15
5	γ = 0.4	74.12	85.38	72.23	76.05
6	γ = 0.5	74.98	84.85	72.37	76.14
7	γ = 0.6	74.30	84.81	72.47	76.07

layers of the Textual Semantic Alignment Module from 1 to 4, and TDPM showed a similar performance to the best results. In addition, we also adjusted the number of layers in the Temporal Alignment Modules and obtained comparable results. The stable performance shows that TDPM is sufficiently robust to different hyper-parameters.

V. CONCLUSIONS

In this study, we proposed a novel Temporal Dynamics Perception Model (TDPM) for audio-visual question-answering tasks. The proposed model overcomes the challenges that image-based, and audio-based pre-trained models can not capture temporal dynamics in videos and that question-answer pairs

TABLE VI
PARAMETER ANALYSIS: **LN** DENOTES THE LAYER NUMBER OF THE TEXTUAL SEMANTIC ALIGNMENT MODULE; **ELN** AND **DLN** DENOTE THE LAYER NUMBERS OF THE TEMPORAL DEPENDENCE LEARNING MODULE AND TEMPORAL AUTO-REGRESSION MODULE.

	LN	ELN	DLN	Average Accuracy (%)			
				Audio	Visual	Audio-Visual	All
1	1	1	1	74.61	85.22	72.55	76.27
2	2	1	1	73.62	84.60	72.49	75.90
3	3	1	1	73.74	85.05	72.27	75.92
4	4	1	1	74.24	84.89	72.14	75.89
5	1	2	1	75.05	84.31	72.53	76.10
6	1	3	1	**76.16**	**85.76**	72.45	**76.63**
7	1	4	1	74.05	85.76	**72.86**	76.49
8	1	1	2	75.42	85.43	72.15	76.25
9	1	1	3	73.93	84.89	72.68	76.14
10	1	1	4	73.12	84.64	72.19	75.66

have semantic gaps with the original text of text-based pre-trained models. It captures temporal dynamic dependencies in videos through visual and audio temporal alignment modules, thereby improving the model's spatio-temporal reasoning ability. Meanwhile, the textual semantic alignment module is used to reduce the textual semantic gap and mine the required information from the video, improving the accuracy of the answer.

We conducted extensive experiments on the MUSIC-AVQA dataset to compare TDPM with 10 existing models. The experimental results showed that our model outperformed all the competitors and achieved an accuracy of 76.63%. Compared with all baseline models, TDPM achieved a maximum performance improvement of 17.02% and an average improvement of 10.86% in terms of average accuracy. In addition, the ablation experiments proved that each component contributed to the performance improvement in TDPM.

ACKNOWLEDGMENT

This paper is supported by the Major Project of Sichuan Provincial Natural Science Foundation (2025ZNSFSC0005), the Sichuan Science and Technology Program (2024YFFK0120), the Central Government Guides Local Projects of China (2025ZYDF105) and the National Science Foundation of China (62072419).

REFERENCES

[1] G. Chen, X. Liu, G. Wang, K. Zhang, P. H. S. Torr, X. Zhang, and Y. Tang, "Tem-adapter: Adapting image-text pretraining for video question answer," in *ICCV*, 2023, pp. 13 899–13 909.

[2] Z. Chen, L. Wang, P. Wang, and P. Gao, "Question-aware global-local video understanding network for audio-visual question answering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 4109–4119, 2024.

[3] G. Li, W. Hou, and D. Hu, "Progressive spatio-temporal perception for audio-visual question answering," in *ACM MM*, 2023, pp. 7808–7816.

[4] G. Li, Y. Wei, Y. Tian, C. Xu, J. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *CVPR*, 2022, pp. 19 086–19 096.

[5] Z. Li, D. Guo, J. Zhou, J. Zhang, and M. Wang, "Object-aware adaptive-positivity learning for audio-visual question answering," in *AAAI*, 2024, pp. 3306–3314.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, vol. 139, 2021, pp. 8748–8763.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[8] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[9] Y. Wei, D. Hu, Y. Tian, and X. Li, "Learning in audio-visual context: A review, analysis, and new perspective," *CoRR*, vol. abs/2208.09579, 2022.

[10] G. Li, H. Du, and D. Hu, "Boosting audio visual question answering via key semantic-aware cues," in *ACM MM*, 2024, pp. 5997–6005.

[11] D. Hu, Z. Wang, F. Nie, R. Wang, and X. Li, "Self-supervised learning for heterogeneous audiovisual scene analysis," *IEEE Trans. Multim.*, vol. 25, pp. 3534–3545, 2023.

[12] A. Senocak, T. Oh, J. Kim, M. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *CVPR*, 2018, pp. 4358–4366.

[13] K. Li, Z. Yang, L. Chen, Y. Yang, and J. Xiao, "CATR: combinatorial-dependence audio-queried transformer for audio-visual video segmentation," in *ACM MM*, 2023, pp. 1485–1494.

[14] J. Liu, Y. Wang, C. Ju, C. Ma, Y. Zhang, and W. Xie, "Annotation-free audio-visual segmentation," in *WACV*, 2024, pp. 5592–5602.

[15] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," in *NeurIPS*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.

[16] K. K. Rachavarapu and A. N. Rajagopalan, "Boosting positive segments for weakly-supervised audio-visual video parsing," in *ICCV*, 2023, pp. 10 158–10 168.

[17] Y. Jiang and J. Yin, "Target-aware spatio-temporal reasoning via answering questions in dynamic audio-visual scenarios," in *EMNLP*, 2023, pp. 9399–9409.

[18] Y. Lin, Y. Sung, J. Lei, M. Bansal, and G. Bertasius, "Vision transformers are parameter-efficient audio-visual learners," in *CVPR*, 2023, pp. 2299–2309.

[19] M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? adapting pretrained representations to diverse tasks," in *RepL4NLP@ACL*. Association for Computational Linguistics, 2019, pp. 7–14.

[20] A. Nakamura and T. Harada, "Revisiting fine-tuning for few-shot learning," *CoRR*, vol. abs/1910.00216, 2019.

[21] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, 2024.

[22] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *CoRR*, vol. abs/2111.03930, 2021.

[23] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *ICCV*, 2021, pp. 1708–1718.

[24] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "VQA: visual question answering," *IJCV*, vol. 123, pp. 4–31, 2017.

[25] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *CVPR*, 2019, pp. 1999–2007.

[26] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019, pp. 6281–6290.

[27] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *AAAI*, 2019, pp. 8658–8665.

[28] I. Schwartz, A. G. Schwing, and T. Hazan, "A simple baseline for audio-visual scene-aware dialog," in *CVPR*, 2019, pp. 12 548–12 558.

[29] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360° videos," in *ICCV*, 2021, pp. 2011–2021.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2015.