Prompt Image to Watch and Hear: Multimodal Prompting for Parameter-Efficient Audio-Visual Learning

Kai Wang[1], *Shentong Mo*[2], *Yapeng Tian*[3], *Dimitrios Hatzinakos*[1]
[1]*University of Toronto* [2]*Carnegie Mellon University* [3]*University of Texas at Dallas*

BMVC 2025

Introduction

Motivations:

1. Pre-training audio-visual encoders on audio-visual pairs requires massive data and expensive computing resources.
2. Retraining pre-trained audio-visual models on downstream tasks increases the extra training budget and causes overfitting problems.
3. Existing prompting learning methods lose modality-specific information due to the compressed token fusion

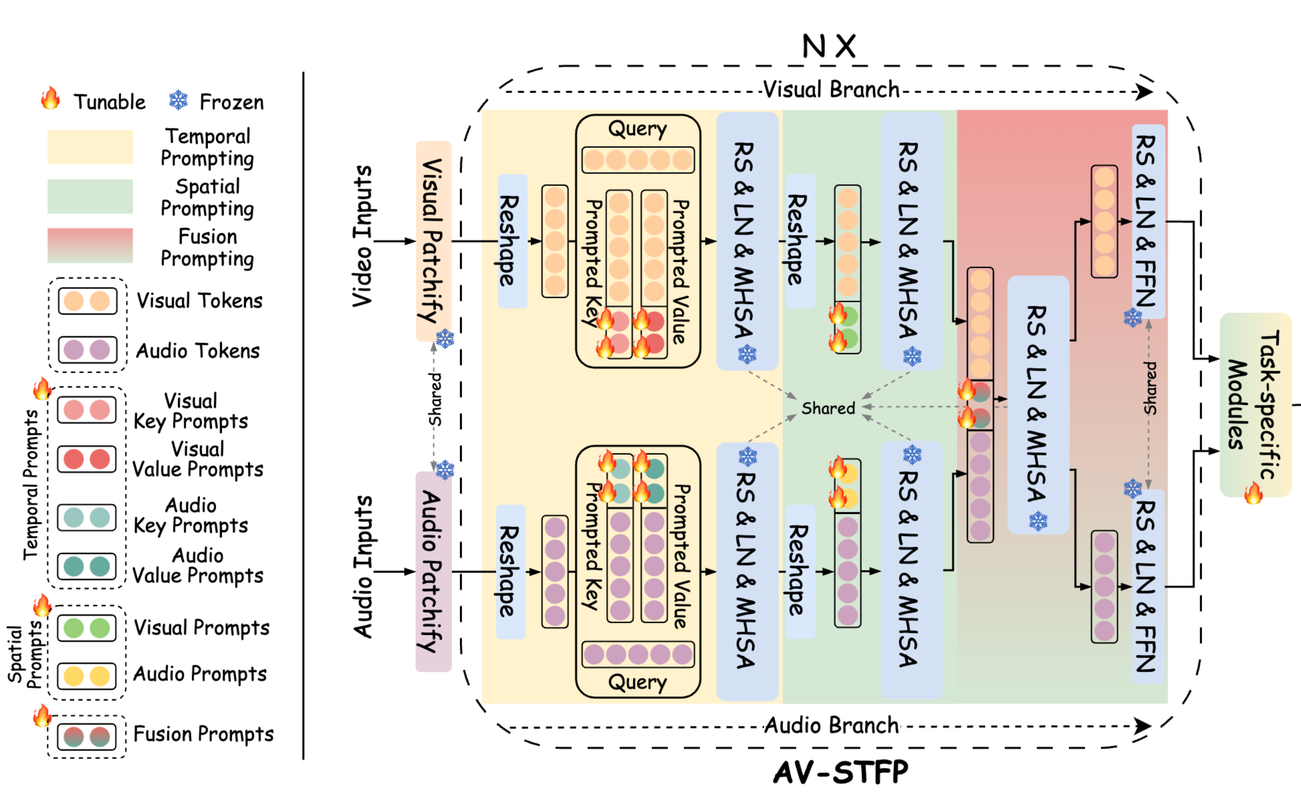
Raised Question:

Is it possible to prompt frozen image models to capture semantic-temporal patterns of audio-video modality and fuse them without losing modality-specific information? The answer is ***Yes****.*

Contributions:

1. Propose an AV-STFP, a novel prompt learning method to prompt the pre-trained image-only models to efficiently learn spatial-temporal information of audio-visual modality and fuse them for feature interactions.
2. AV-STFP adopts temporal-aware prompting, spatial-aware prompting, and multimodal fusion prompting to enable cross-modal adaptation from image to video, audio, and multi-modal fusion
3. Experiments on audio-visual tasks demonstrate that AV-STFP achieves competitive or even better performance than SOTA methods while only introducing a small amount (0.6% of model parameters) of trainable parameters

Overview of Proposed AV-STFP



Temporal-Aware Prompting:

1. Unlike static images, audio and video contain rich dynamic information with temporal patterns across frames.
2. Temporal-aware prompting is introduced to adapt the static image models to capture temporal dependencies of audio and video modalities.
3. Insert a set of key-value prompts into the audio-video embeddings along the temporal dimension.

Spatial-Aware Prompting:

1. Audio spectrogram has different semantics due to the acoustic characteristics.
2. Spatial-aware prompting aims to enable the pre-trained image priors to learn the spatial contexts of audio and further strengthen the spatial modelling of video by prepending a set of spatial prompts into the tokens.
3. Insert a set of prompts into the audio-video embeddings along the spatial dimension.

Multimodal Fusion Prompting:

1. It's crucial to integrate audio and video modalities for enhancing the downstream prediction.
2. Multimodal fusion prompting is proposed to enforce image models to integrate audio and visual modalities without losing modality-specific information
3. Incorporate a set of fusion prompts between audio and visual tokens

Main Experimental Results

Table1: Comparison with state-of-the-art AVE methods

Method	Encoder		Pretrain Dataset		Trainable Param. ↓	Total Param. ↓	Acc. ↑
	Visual	Audio	Visual	Audio			
PSP [54]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	1.7M	217.4M	77.8%
AVT [25]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	15.8M	231.5M	76.8%
RFJC [7]	VGG-19 ❄	VGGish ❄	ImageNet	AudioSet	22.8M	238.5M	76.2%
AVEL [39]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	3.7M	136.0M	74.0%
CMRAN [49]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	15.9M	148.2M	78.3%
MM-Pyramid [51]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	44.0M	176.3M	77.8%
CMBS [46]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	14.4M	216.7M	79.7%
AVSDN [26]	ResNet-152 ❄	VGGish ❄	ImageNet	AudioSet	8.0M	140.3M	75.4%
MBT* [32]	ViT-B-16 🔥	AST 🔥	ImageNet	AudioSet	172.0M	172.0M	77.8%
DG-SCT [8]	Swin-L ❄	HTS-AT ❄	ImageNet	AudioSet	43.6M	461.3M	82.2%
LAVISH [27]	Swin-L ❄, shared		ImageNet	✘	10.1M	238.8M	81.1%
CoPL [53]	Swin-L ❄, shared		ImageNet	✘	1.7M	232.3M	81.9%
LAVISH [27]	ViT-B-16 ❄, shared		ImageNet	✘	4.7M	107.2M	75.3%
	ViT-L-16 ❄, shared		ImageNet	✘	14.5M	340.1M	78.1%
STG-CMA (TINY) [45]	ViT-B-16 ❄, shared		CLIP	✘	3.5M	89.6M	76.3%
STG-CMA (BASE) [45]	ViT-L-14 ❄, shared		CLIP	✘	20.1M	323.6M	**83.3%**
AV-STFP (ours)	ViT-B-16 ❄, shared		CLIP	✘	**1.1M**	**87.1M**	76.7%
	ViT-L-14 ❄, shared		CLIP	✘	1.8M	303.4M	**83.0%**

Table2: Comparison with state-of-the-art AVQA methods

Method	Encoder		Trainable Param. ↓	Total Param. ↓	Question ↑			
	Visual	Audio			AQ	VQ	AVQ	Avg.
AVSD [37]	Non-shared 🔥		N/A	N/A	68.5%	70.8%	65.5%	67.4%
Pano-AVQA [52]	Non-shared 🔥		N/A	N/A	70.7%	72.6%	66.6%	68.9%
ST-AVQA [20]	Non-shared ❄		10.6M	94.4M	74.1%	74.0%	69.5%	71.5%
LAVISH [27]	Shared ❄		21.1M	249.8M	75.7%	80.4%	70.4%	74.0%
DG-SCT [8]	Non-shared ❄		110.4M	520.2M	77.4%	81.9%	70.7%	74.8%
STG-CMA (BASE) [45]	Shared ❄		26.9M	221.9M	77.1%	80.8 %	70.7%	74.5%
STG-CMA (LARGE) [45]	Shared ❄		83.0M	278.1M	**78.7%**	83.0%	72.3%	76.2%
CoPL [53]	Shared ❄		1.8M	240.2M	77.3%	77.6%	**76.3%**	76.7%
AV-STFP (ours)	Shared ❄		9.3M	312.7M	77.3%	**84.8%**	73.2%	**77.0%**

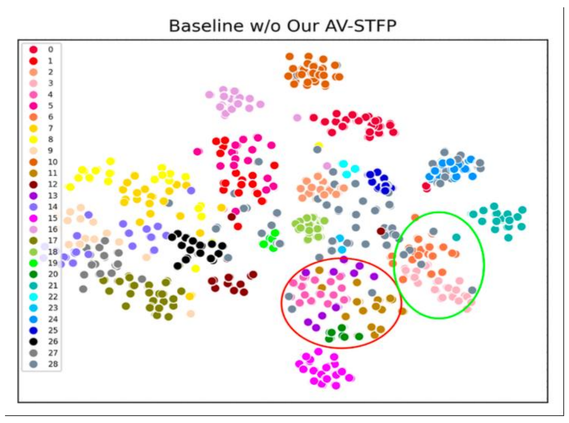
Table3: Comparison with state-of-the-art AVAR methods

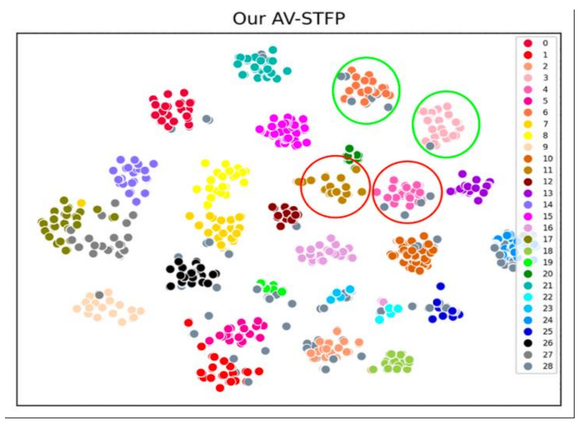
Method	Pretrain	Backbone	Param. ↓	Acc. ↑
AVSlowFast [48]	Kinetics-400 [17]	ResNet-50 🔥	38.5M	83.7%
CrissCross [36]	Kinetics-400 [17]	R(2+1)D-18 🔥	15.4M	66.7%
MoMA [43]	ImageNet [4]	ViT-B/16 ❄	16.5M	90.6%
	ImageNet [4]	ViT-L/14 ❄	57.7M	92.2%
	CLIP	ViT-B/16 ❄	16.5M	91.1%
	CLIP	ViT-L/14 ❄	57.7M	**94.2%**
AV-STFP (ours)	CLIP	ViT-B/16 ❄	**1.10M**	85.6%
	CLIP	ViT-L/14 ❄	1.82M	92.6%

Ablation Studies

Prompting			Param.	Acc.
Temporal	Spatial	Fusion		
✘	✘	✘	1.09M	78.2%
✔	✘	✘	1.28M	80.7%
✘	✔	✘	1.58M	81.4%
✘	✘	✔	1.14M	79.5%
✔	✔	✘	1.78M	82.4%
✘	✔	✔	1.63M	81.2%
✔	✘	✔	1.33M	81.5%
✔	✔	✔	1.82M	**83.0%**

Model	A.P.	V.P.	F.P.	Param.	Acc.
Audio Branch					
w/o Prompt	✘	✘	✘	0.06M	44.0%
w/ Prompt	✔	✘	✘	0.4M	65.5%
Visual Branch					
w/o Prompt	✘	✘	✘	0.06M	75.8%
w/ Prompt	✘	✔	✘	0.4M	79.4%
Audio-Visual Branch Without Fusion					
w/ Audio Prompt	✔	✘	✘	1.43M	80.8%
w/ Visual Prompt	✘	✔	✘	1.43M	79.0%
w/ AV Prompt	✔	✔	✘	1.8M	81.6%
Ours vs Finetuning					
Full Finetuning	✘	✘	✘	606.8M	57.6%
AV-STFP(ours)	✔	✔	✔	1.8M	**83.0%**





Conclusions

- We proposed an efficient and novel prompt learning method to transfer the learnt knowledge from the pre-trained image models into the audio-visual domain
- Our proposed AV-STFP achieves better performance than existing SOTA methods in various audio-visual tasks via a small set of trainable parameters.
- In the future, we will evaluate our model in more challenging audio-visual environment.