Prompting Segmentation with Sound is Generalizable Audio-Visual Source Localizer

Yaoting Wang[1†],Weisong Liu[2†], Guangyao Li[1], Jian Ding[3], Di Hu[1*], Xi Li[4]
[1] Renmin University of China, Beijing, China [2] Northwest Polytechnical University, Xi'an, China
[3] Wuhan University, Wuhan, China [4] Zhejiang University, Hangzhou, China
yaoting.wang@outlook.com, liuweisong@mail.nwpu.edu.cn, guangyaoli@ruc.edu.cn
jian.ding@whu.edu.cn, dihu@ruc.edu.cn, xilizju@zju.edu.cn

Abstract

Never having seen an object and heard its sound simultaneously, can the model still accurately localize its visual position from the input audio? In this work, we concentrate on the Audio-Visual Localization and Segmentation tasks but under the demanding zero-shot and few-shot scenarios. Different from existing methods that mostly employ the encoder-fusion-decoder paradigm to localize from the fused audio-visual feature, we introduce the encoder-prompt-decoder paradigm to better fit the data scarcity and varying data distribution dilemmas with the help of abundant knowledge from pre-trained models. Specifically, we construct Semantic-aware Audio Prompt (SAP) to help the visual foundation model focus on sounding objects, meanwhile, the semantic gap between the visual and audio modalities is encouraged to shrink. Then, we develop a Correlation Adapter (ColA) to keep minimal training efforts as well as maintain adequate knowledge of the visual foundation model. Extensive experiments demonstrate that this new paradigm outperforms other fusion-based methods in both the unseen class and cross-dataset settings.

1. Introduction

Audio-Visual Localization (AVL) utilizes audio input to locate sounding objects within a visual scene [1, 3, 9, 21]. However, to meet the demand for greater precision in real-world scenarios, AVL has shifted from localizing with bounding boxes or coarse heatmaps to pixel-level segmentation masks, known as Audio-Visual Segmentation (AVS) [26]. Currently, as illustrated in the upper-center of Figure 1, many methods commonly implement AVS based on the fused audio-visual representation, and we name such methods as the encoder-fusion-decoder paradigm. However, in real-world applications, the limited training data and varying data distribution hinder the segmentation performance of models when faced with unseen classes and different datasets. Hence in this study, we aim to enhance the current research on AVS, and enable effective localization on both the unseen classes and cross-datasets settings.

To examine the generalization capability of the encoder-fusion-decoder paradigm, we set cross-dataset tests on the VGG-SS dataset [3] but trained on AVS-Benchmarks [26]. The right side of Figure 1 shows that under the zero-shot setting, current methods fail to outperform the classic AVL models trained on the VGG-Sound dataset [4], which shares the same data distribution as VGG-SS. We attribute this result to explore the audio-visual correlation on specific datasets using the encoder-fusion-decoder paradigm, leading to the restricted generalization ability due to the lack of utilize prior knowledge of pre-trained models. [17] proved that simply using the prior knowledge in the pre-trained visual model can improve the generalization ability.

We argue that one of the ways to enhance generalization capability is to more effectively leverage the prior knowledge encoded in large-scale pre-trained models [23]. Many models in natural language processing (NLP) and computer vision (CV) exhibit remarkable generalization abilities [2, 7]. Some researchers [12, 24, 25] consider prompt learning to be capable of enhancing the model generalization ability. The key benefit lies in its ability to align the data distribution of downstream tasks with the prior knowledge embedded in the foundation model, as the task formats and the output space have reached a consensus [10, 22], thus enhancing the model generalization capability across various downstream tasks. Inspired by prompt learning in NLP and multimodal research, we consider that a visual foundation model with audio prompts can be a promising way to achieve generalizable AVL and AVS.

Therefore, we introduce an encoder-prompt-decoder paradigm that prompts the visual foundation model to per-

[*] Corresponding author.
[†] These authors contributed equally.

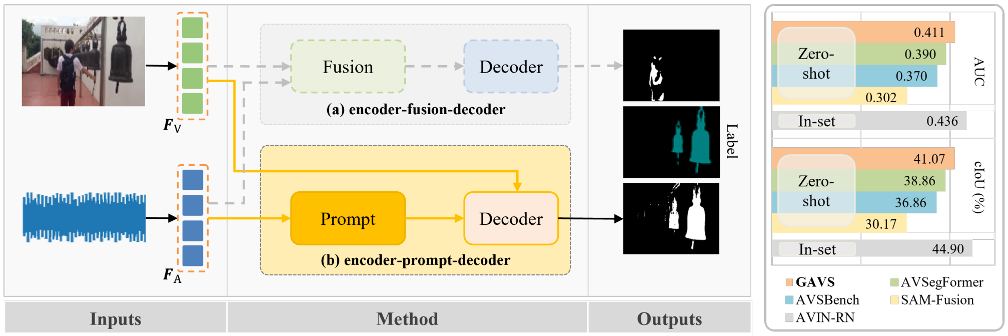

Figure 1. The AVS pipeline of encoder-fusion-decoder (the upper-center) and our proposed encoder-prompt-decoder (the lower-center) paradigms. Classical encoder-fusion-decoder methods decode mask from the fused modality while we prompting visual input with audio to adapt AVL and AVS tasks to the visual foundational model. The results on the VGG-SS dataset highlight the challenge of generalizing across different datasets. However, our approach breaks through the 40% cIoU barrier, getting the performance closer to the best method trained on in-set (VGG-Sound).

form segmentation using audio cues; rather than solely decoding from the fused modality. This paradigm enables the seamless integration of the AVS task within the underlying visual foundation model, consequently enhancing the generalization capability in AVS by effectively leveraging the prior knowledge of the pre-trained model. Firstly, we construct a Semantic-aware Audio Prompt (SAP) to bridge the semantic gap between the visual and audio modalities, aligning the semantics of the given image and audio through contrastive learning. SAP assists the visual foundation model in localization based on the provided audio cues with the same cross-modal semantics. Subsequently, we use a Correlation Adapter (ColA) to construct the audio-visual correlation to retain as much prior knowledge as possible from the visual foundation model. We use the Segment Anything Model (SAM) [11] as our visual foundation model for its generalizable segmentation capabilities.

2. Generalizable Audio-Visual Segmentation

2.1. Data Preprocessing

We split the video clip into images at one-second intervals and feed them into ViT [5] to get visual features $F_V \in \mathbb{R}^{d_V \times H \times W}$. Additionally, we process the audio using VGGish [8] to get the audio features $F_A \in \mathbb{R}^{d_m}$, then we acquire the corresponding audio-visual pairs.

2.2. Semantic-aware Audio Prompting

As shown in the left part of Figure 2, SAP serves to prompt the visual foundation model to retrieve sounding objects from the visual space with the prior knowledge.

We first obtain the comprehensive visual feature $F_{VG} \in \mathbb{R}^{d_V}$ by performing global average pooling on the visual feature F_V, then we feed F_{VG} into an MLP to achieve consistent dimension with the audio feature F_A, resulting in the visual cues $F_C \in \mathbb{R}^{d_m}$ to align the semantics with audio by contrastive learning. We also introduce a learnable adaptive noise $F_N \in \mathbb{R}^{d_N}$ as part of the audio prompt to implicitly aligns current modality features with the data distribution of the visual foundation model during the tuning process for specific downstream tasks, and enhancing the model's generalization and noise tolerance during inference.

Through the aforementioned operations, we simply concatenate the prompt components and audio input to obtain the final audio prompt $F_{A'} \in \mathbb{R}^{2d_m+d_N}$, which we also refer to as SAP. Finally, we feed the visual input and projected prompt $F_P \in \mathbb{R}^{6\times d_V}$ [1] into the Audio Source Decoder for sounding object segmentation.

2.3. Audio Source Decoder

We further construct the audio-visual correlation using the visual foundation model with the help of its prior-knowledge. However, to minimise the harm to the foundation model, instead of tuning the whole decoder or modifying the cross-modal attention modules in the middle of Figure 2 that already contain prior interactive knowledge, we propose ColA to efficiently construct the audio-visual correlation by tuning the core context engaging in different cross-modal attention modules.

Specifically, ColA is a bottleneck adapter that takes the context between different cross-modal attention modules as input and performs lightweight updates on it. In this case, the context refers to the audio prompt that has been updated through the first cross-modal (audio-visual) module, which then serves as the key and value for the next cross-modal (visual-audio) module.

After traversing through all transformer layers, the final visual output is used as the mask embedding $F_M \in$

[1] SAM provides 6 token slots including 1 IoU token, 4 query tokens and 1 prompt token, we use the first query token for mask generating.

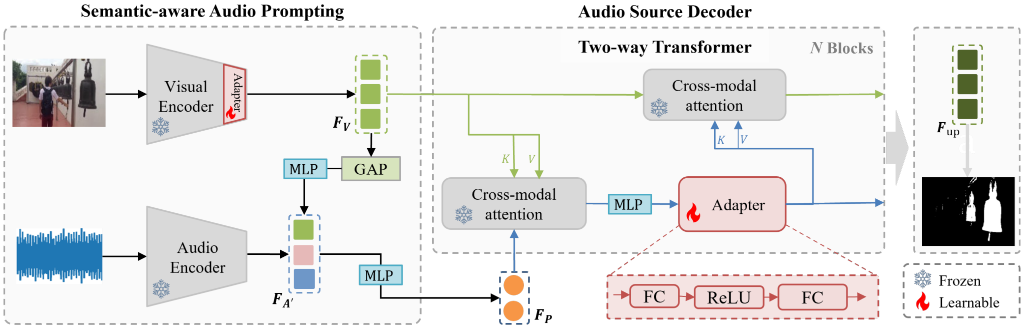

Figure 2. The overview of GAVS. (1) We firstly align the audio and visual semantics for SAP, and introduce visual features as cues (the green one in $F_{A'}$) for audio input (the blue one in $F_{A'}$). Then we further combine audio input with learnable adaptive noise (the pink one in $F_{A'}$) to construct the final SAP $F_{A'}$, and get the projected prompt F_P. (2) Next, we utilize cross-modal attention to learn the correlation between audio and visual in the Audio Source Decoder, projecting audio into the visual space.

			V1S		V1M		V2	
Method	Audio-backbone	Visual-backbone	mIoU(%)	F-score	mIoU(%)	F-score	mIoU(%)	F-score
AVSBench [26]	VGGish	PVT-v2	78.70	0.879	54.00	0.645	62.45	0.756
AVSegFormer [6]	VGGish	PVT-v2	82.06	0.899	58.36	0.693	64.34	0.759
AUSS [13]	VGGish	PVT-v2	**89.40**	**0.942**	63.50	0.752	-	-
AVSC [14]	VGGish	PVT-v2	81.29	0.886	59.50	0.657	-	-
AuTR [15]	VGGish	Swin-base	80.40	0.891	56.20	0.672	-	-
AV-SAM [19]	ResNet18	ViT-Base	40.47	0.566	-	-	-	-
Audio-SAM †	VGGish	ViT-Base	56.33	0.727	33.68	0.459	57.41	0.684
SAM-Fusion‡	VGGish	ViT-Base	71.92	0.775	50.61	0.637	60.19	0.724
GAVS (ours)	VGGish	ViT-Base	80.06	0.902	**63.70**	**0.774**	**67.70**	**0.788**

Table 1. Performance on AVS-Benchmarks. †: We only replace the sparse prompt in SAM with audio inputs, to conduct a comparative experiment with AV-SAM. ‡: Set up similar to GAVS, but fuse audio and visual without prompt before the Audio Source Decoder.

$\mathbb{R}^{d_V \times H \times W}$. We appropriately upscale F_M and then multiply it with the audio query $F_P[1]$ to generate the final mask $M_{pred} \in \mathbb{R}^{4H \times 4W}$.

3. Experiments

We evaluate the grounding segmentation performance and generalization ability on AVS-Benchmarks and VGG-SS datasets, and their subsets.

3.1. Grounding Segmentation on AVS-Benchmarks

AVS-Benchmarks [26] is a dataset specifically designed for AVS task. Refer to Table 1, our model achieves the best performance in multi-source setting (V1M and V2) and gets comparable performance in single-source setting (V1S). In the comparison with AV-SAM, where both models utilize prompts, our implemented straightforward Audio-SAM freezes all parameters except for the audio input, which is passed through an additional MLP for updating. This resulted in a performance improvement of 15% compared to AV-SAM, demonstrating the effectiveness of the encoder-prompt-decoder paradigm. Besides, we further compare the performance of various open-source models at different data volumes to demonstrate our superiority in data utilization. As shown in Figure 3, with only 50% of the data, we can achieve the best performance equivalent to using 100% of the data by other models.

3.2. Unseen Classes on AVS-V3

We design AVS-V3 to test the unseen classes generalization ability of models for the AVS task. It is set up with four settings, namely 0-shot, 1-shot, 3-shot, and 5-shot. As shown in Table 2, our model achieves the highest 0-shot performance, exhibiting superior generalization when encountering unseen objects. Meanwhile, we can observe that after 3-shot learning, our model surpasses other models' performance trained with 5-shot, indicating that our model possesses better few-shot learning ability.

	0-shot		1-shot		3-shot		5-shot	
Method	mIoU(%)	F-score	mIoU(%)	F-score	mIoU(%)	F-score	mIoU(%)	F-score
AVSBench [26]	53.00	0.707	56.11	0.754	63.22	0.767	63.87	0.783
AVSegFormer [6]	54.26	0.715	58.30	0.764	64.19	0.774	65.17	0.785
SAM-Fusion	46.25	0.630	50.39	0.671	57.05	0.719	60.82	0.741
GAVS (ours)	**54.71**	**0.722**	**62.89**	**0.768**	**66.28**	**0.774**	**67.75**	**0.795**

Table 2. Performance on AVS-V3 for testing the generalization ability on unseen objects. Our model GAVS, which is trained with encoder-prompt-decoder paradigm achieves a significant performance improvement compared to other encoder-fusion-decoder models.

3.3. Cross-datasets on VGG-SS

VGG-SS. VGG-SS [3] is a dataset designed for the AVL task performance test. In this experiment, we test models'

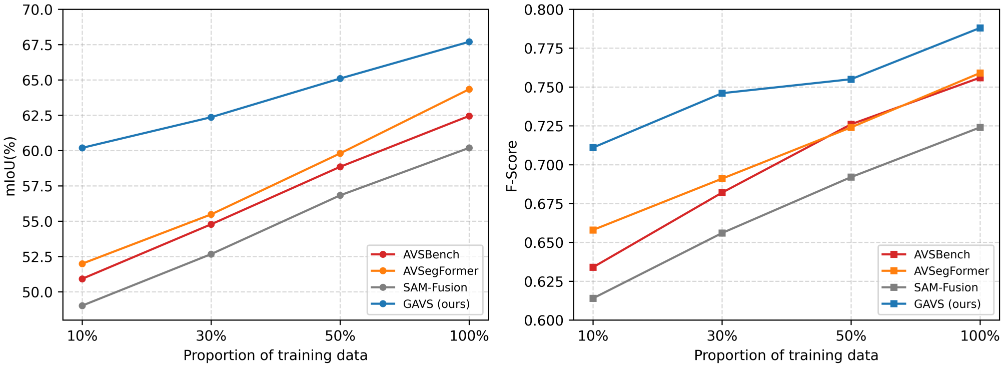

Figure 3. Visualization of performance improvements of AVS models on the AVS-V2 dataset in relation to the amount of data used for training. We compare models with subsets consisting of 10%, 30%, and 50% of the full dataset. Our results show that our method achieves better performance with only 10% of the training data compared to other models trained with 30%. Moreover, our model outperforms other models trained on the full dataset when trained with only half of the data.

Method	Train	cIoU(%)	AUC
HardWay [3]	in-set	34.4	0.382
EZ-VSL [18]	in-set	38.85	0.395
SLAVC [17]	in-set	39.80	-
MarginNCE [20]	in-set	39.78	0.400
AVIN-RN [16]	in-set	**44.90**	**0.436**
AVSBench [26]	zero-shot	36.86	0.370
AVSegFormer [6]	zero-shot	38.86	0.390
SAM-Fusion	zero-shot	30.17	0.302
GAVS (ours)	zero-shot	**41.07**	**0.411**

Table 3. The results of VGG-SS for comparing the performance of zero-shot AVS models with traditional self-supervised in-set AVL models. Our model outperforms other AVS models in cross-dataset setting.

cross-dataset generalization on the VGG-SS test set. Previous works [3, 16, 17, 18, 20] trained models on VGG-Sound 144k, we label them as “trained on in-set” because VGG-SS is extracted from VGG-Sound. In contrast, we train typical AVS models on AVS-V2 and can be labelled as “trained with zero-shot”. As shown in Table 3, models such as AVSBench and AVSegFormer perform well on AVS-Benchmarks but fail to perform as well in cross-dataset settings. Our model has better cross-dataset generalization ability and surpasses other zero-shot models, although there is still some gap compared to the best in-set model.

VGG-SS-Sub. We split VGG-SS-Sub as a subset of VGG-SS to test the cross-dataset generalization ability of fusion-based and prompt-based AVS models transfer from AVS to AVL task. Same with the AVS-V3, it is set up with zero-shot and few-shot (1, 3, 5) settings. Note that the zero-shot performance of this subset cannot be compared with the VGG-SS full set as the test set is different. From Table 4, we can observe that our model achieves better zero-shot and few-shot performance, suggesting that with SAP and ColA, our model can better fit the data distribution across different datasets.

	0-shot		1-shot		3-shot		5-shot	
Method	cIoU(%)	AUC	cIoU(%)	AUC	cIoU(%)	AUC	cIoU(%)	AUC
AVSBench [26]	37.28	0.374	53.33	0.534	56.78	0.569	57.38	0.574
AVSegFormer [6]	37.99	0.380	53.41	0.534	56.84	0.569	57.65	0.577
SAM-Fusion	31.22	0.315	40.39	0.407	45.25	0.453	48.67	0.487
GAVS (ours)	**38.62**	**0.387**	**53.70**	**0.537**	**57.41**	**0.574**	**60.14**	**0.602**

Table 4. Performance on VGG-SS-Sub for testing the generalization ability across different datasets. Our model is trained following the encoder-prompt-decoder paradigm and achieves the best zero-shot and few-shot performance.

4. Conclusion

The development of large-scale pre-trained models has greatly enhanced the generalization performance of traditional CV tasks, but little attention gives to the generalization of cross-modal AVS in zero-shot and few-shot scenarios. In this work, we introduce GAVS, the model following the encoder-prompt-decoder paradigm to address the increasing demand for precise localization with limited annotated data and varying data distribution. Compared with other models following the encoder-fusion-decoder paradigm, our proposed method achieves generalizable cross-modal segmentation, benefiting from using SAP to help the visual foundation model focus on the sounding objects and using ColA for efficient audio-visual correlation construction. Our method is only one solution and provides a reference for exploring generalizable AVS, future work can investigate more flexible methods for generalizable cross-modal audio-visual correlation learning based on large-scale pre-trained models.

Acknowledgments This research was supported by National Natural Science Foundation of China (NO.62106272), the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), in part by the Research Funds of Renmin University of China (NO. 21XNLG17) and Public Computing Cloud, Renmin University of China.

References

[1] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset, 2020.

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

[6] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *arXiv preprint arXiv:2307.01146*, 2023.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] 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.

[9] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020.

[10] Chen Jia and Yue Zhang. Prompt-based distribution alignment for domain generalization in text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10147–10157, 2022.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[12] Aodi Li, Liansheng Zhuang, Shuo Fan, and Shafei Wang. Learning common and specific visual prompts for domain generalization. In *Proceedings of the Asian Conference on Computer Vision*, pages 4260–4275, 2022.

[13] Yuhang Ling, Yuxi Li, Zhenye Gan, Jiangning Zhang, Mingmin Chi, and Yabiao Wang. Hear to segment: Unmixing the audio to guide the semantic segmentation. *arXiv preprint arXiv:2305.07223*, 2023.

[14] Chen Liu, Peike Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics, 2023.

[15] Jinxiang Liu, Chen Ju, Chaofan Ma, Yanfeng Wang, Yu Wang, and Ya Zhang. Audio-aware query-enhanced transformer for audio-visual segmentation. 2023.

[16] Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang. Induction network: Audio-visual modality gap-bridging for self-supervised sound source localization. *arXiv preprint arXiv:2308.04767*, 2023.

[17] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022.

[18] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022.

[19] Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentationsegmentation. *arXiv preprint arXiv:2305.01836*, 2023.

[20] Sooyoung Park, Arda Senocak, and Joon Son Chung. Marginnce: Robust sound localization with a negative margin. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[21] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 292–308. Springer, 2020.

[22] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[23] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*, 2023.

[24] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.

[25] Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt vision transformer for domain generalization. *arXiv preprint arXiv:2208.08914*, 2022.

[26] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022.