Synergistic Prompting for Robust Visual Recognition with Missing Modalities – Supplementary Material –

Zhihui Zhang[1] Luanyuan Dai[3] Qika Lin[4] Yunfeng Diao[5] Guangyin Jin[6]
Yufei Guo[7] Jing Zhang[8] Xiaoshuai Hao[2,*]

[1]Beijing Institute of Technology [2]Beijing Academy of Artificial Intelligence
[3]Nanjing University of Science and Technology [4]National University of Singapore
[5]Systems Laboratory of Anhui Province, Hefei University of Technology [6]National Innovative Institute of Defense Technology
[7]Intelligent Science & Technology Academy of CASIC [8]School of Computer Science, Wuhan University
3220231441@bit.edu.cn xshao@baai.ac.cn

This supplementary material provides additional details on the proposed SyP and experimental results that could not be included in the main manuscript due to page limitations. Specifically, this appendix is organized as follows:

- Sec. A presents details of the training datasets.
- Sec. B presents additional details of training strategies.
- Sec. C complements more experiment results and analysis.
- Sec. D shows more visualization results to prove the effectiveness of ***SyP***.

A. Details of Training Datasets

To evaluate the effectiveness of the proposed SyP, we conduct experiments on three widely-used multimodal datasets: **MM-IMDb**, **Hateful Memes**, and **UPMC Food-101**. These datasets represent diverse tasks, including movie genre classification, hateful meme detection, and food classification, and are designed to challenge models in handling multimodal data under varying conditions. Below, we provide an overview of each dataset, highlighting their key characteristics, modalities, and task-specific challenges. A summary is presented in Tab. 1, and visualization of the samples in the three datasets is shown in Fig. 1.

MM-IMDb is the largest publicly available dataset for movie genre classification, featuring 25,959 movies annotated with both poster images and textual metadata. It is a multi-label classification task, as movies can belong to multiple genres simultaneously. The dataset is split into 15,552 training, 2,608 validation, and 7,799 test samples.

Hateful Memes, curated by Facebook AI, is designed for hateful meme detection. It contains 10,000 multimodal examples where both text and image contribute to the overall meaning. The dataset is structured to challenge unimodal models, with 8,500 samples for training, 500 for validation, and 1,500 for testing.

UPMC Food-101 is a large-scale food classification dataset comprising 90,688 noisy image-text pairs collected from Google Image Search. It spans 101 food categories and aligns with the ETHZ Food-101 dataset. The dataset includes 67,972 training samples and 22,716 test samples, with no designated validation set. Each sample consists of an image and an accompanying textual description.

B. Training Details

All experiments are performed on four NVIDIA RTX 3090 GPUs. We employ the AdamW optimizer [5] with an initial learning rate of $1e^{-3}$ and a weight decay of $2e^{-2}$. The learning rate is warmed up for 10% of the total training steps and then decays linearly to zero. We set the prompt depth as 36. Experiments are conducted on three datasets: Hateful Memes, Food101, and MM-IMDb, using a frozen backbone strategy to efficiently fine-tune CLIP for multimodal tasks while preserving knowledge from large-scale vision-language pretraining.

- **Hateful Memes**: The dataset is trained with a batch size of 256 for 20 epochs (10,000 steps). The validation check interval is 11%, and the optimizer uses a learning rate of 1e-2. The maximum text length is 128.
- **Food101**: The dataset is trained with a batch size of 256 for 200 epochs (20,000 steps). The validation check interval is 20%, and the optimizer uses a learning rate of 1e-2. The maximum text length is 512.
- **MM-IMDb**: The dataset is trained with a batch size of 256 for 100 epochs (25,000 steps). The validation check interval is 20%, and the optimizer uses a learning rate of 1e-4 with a weight decay of 0.01. The maximum text length is 40, and a prompt-based method is used to handle missing modalities. The image encoder is ViT-B/16, with a vocabulary size of 30,522. Whole word masking is disabled,

*Corresponding author: Xiaoshuai Hao.

Dataset	Task	Modalities	Samples	Train/Val/Test Split	Key Features
MM-IMDb [1]	Movie genre classification	Image, Text	25,959 movies	15,552 / 2,608 / 7,799	Multi-label classification; each movie has a poster image and textual metadata.
Hateful Memes [3]	Hateful meme detection	Image, Text	10,000 memes	8,500 / 500 / 1,500	Challenges unimodal models; text and image must be jointly analyzed.
UPMC Food101 [8]	Food classification	Image, Text	90,688 image-text pairs	67,972 / - / 22,716	Aligns with ETHZ Food-101; noisy image-text pairs from Google Image Search.

Table 1. Overview of the multimodal datasets used in this work.

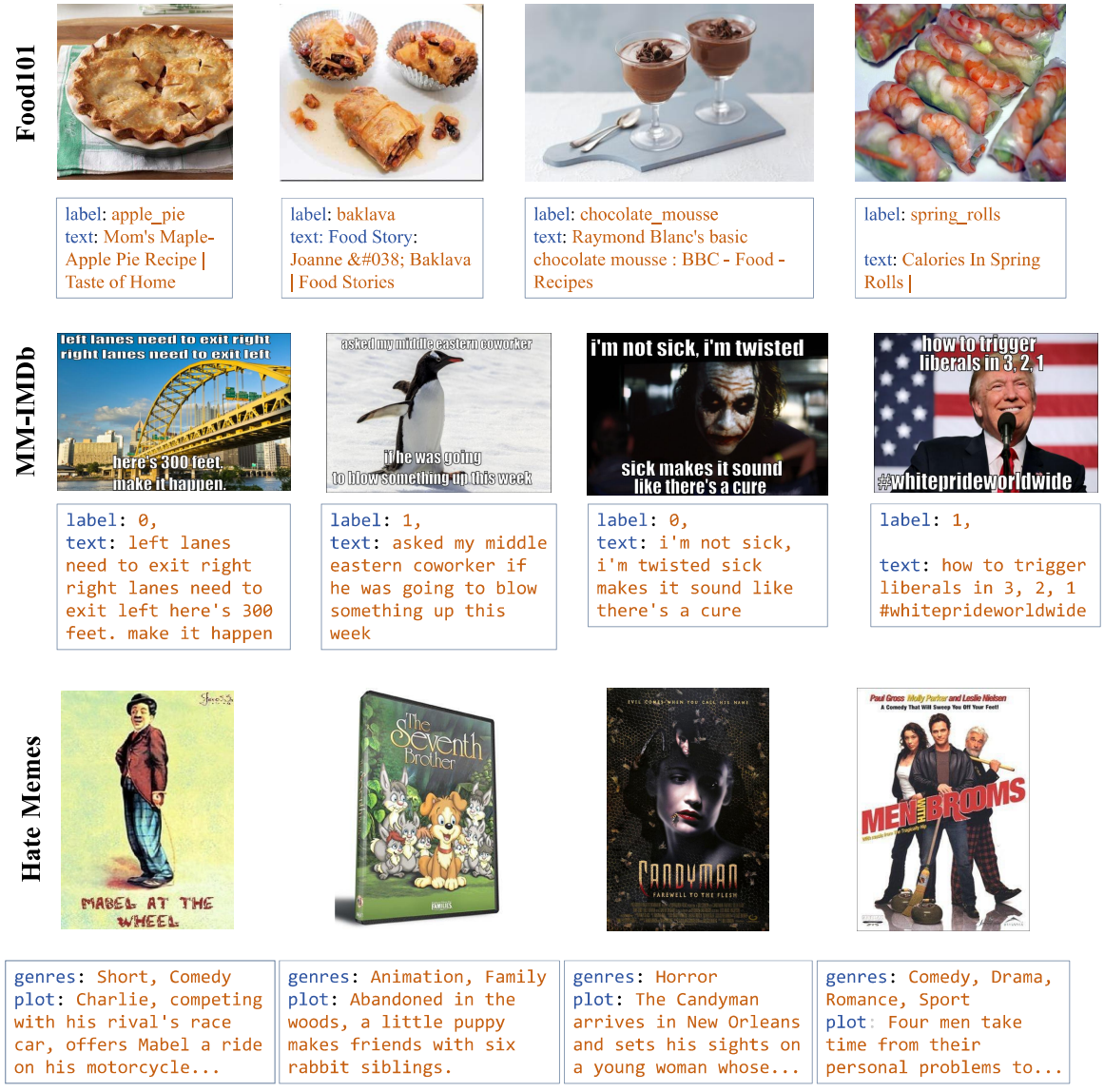


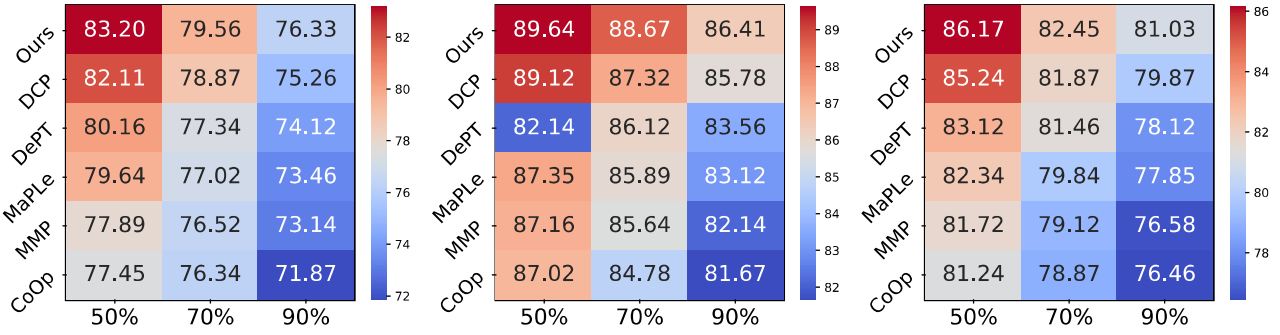
Figure 1. Visualization of the samples in three dataset.

the masked language modeling probability is set to 15%, and both training and validation image transformations use CLIP_transform.

This setup ensures efficient fine-tuning of CLIP for diverse multimodal tasks while leveraging the benefits of large-scale pretraining. The frozen backbone strategy reduces computational costs and maintains the model's generalization capabilities.

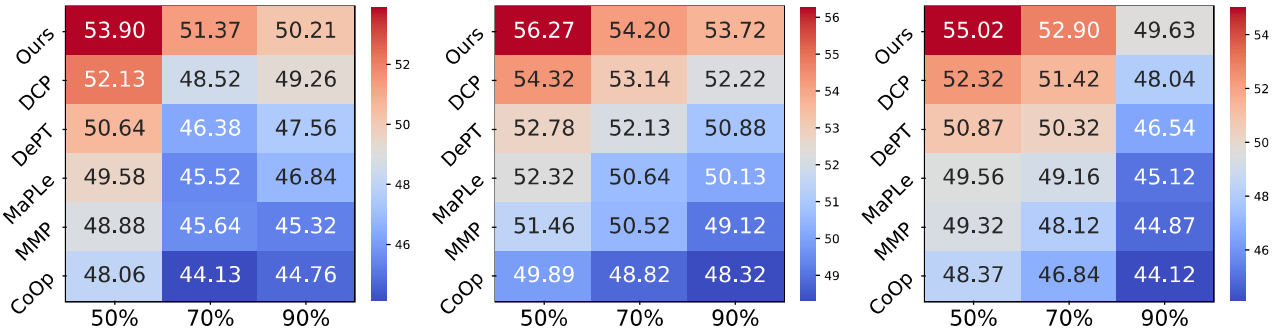
C. Additional Experiments and Analysis

Model Generalizability We evaluate the model's generalizability on Food101 and MM-IMDb datasets under varying



(a) Text Missing (b) Image Missing (c) Both Missing

Figure 2. Generalization analysis on the Food101 dataset across various missing rates in terms of Accuracy.



(a) Text Missing (b) Image Missing (c) Both Missing

Figure 3. Generalization analysis on the MM-IMDb dataset across various missing rates in terms of F1_Macro.

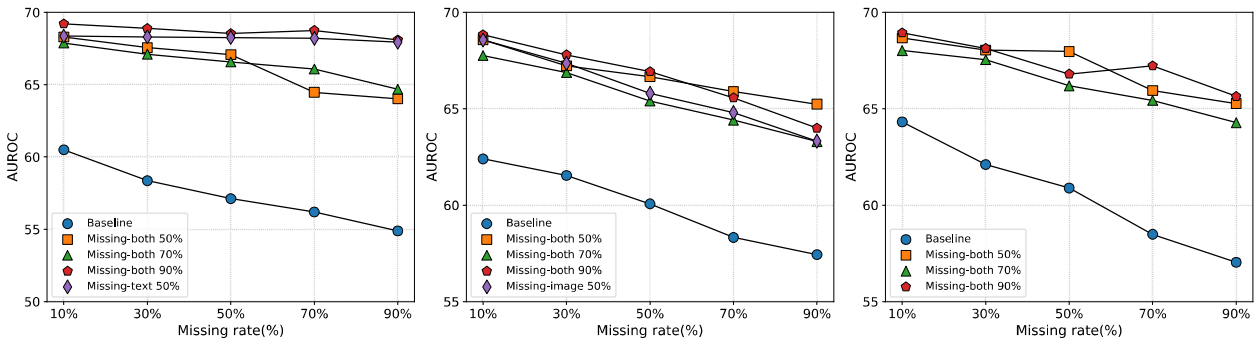
missing modality rates (50%, 70%, 90%). Results are shown in Fig. 2 (Accuracy for Food101) and Fig. 3 (F1_Macro for MM-IMDb).

For the Food101 dataset (Fig. 2), the model demonstrates strong performance across all missing conditions. In text-missing scenarios, it maintains high accuracy by effectively leveraging image features, even at a 90% missing rate. Similarly, in image-missing cases, the model relies on textual data to sustain robust performance. Even when both modalities are missing, the model adapts well, inferring missing information from available data, showcasing its dynamic adaptation capabilities. For the MM-IMDb dataset (Fig. 3), the model exhibits consistent performance despite missing data. In text-missing conditions, it effectively uses visual cues, while in image-missing scenarios, it compensates with textual features. In the most challenging case of both modalities missing, the model still performs competitively by intelligently utilizing remaining data, highlighting its resilience.

These results underscore the model's adaptability to incomplete multimodal data, making it suitable for real-world applications where data completeness is often uncertain. The consistent performance across datasets and conditions validates the effectiveness of its dynamic adapter and synergistic prompts strategy.

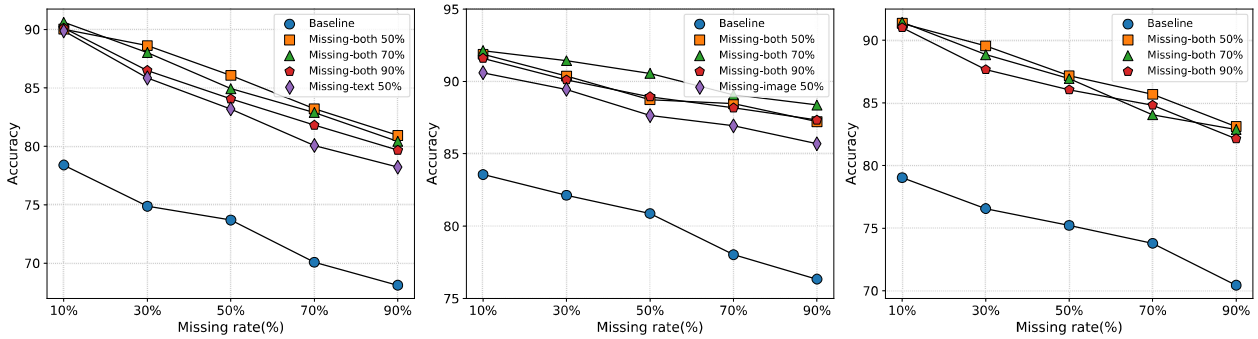
Generalizability to Different Missing Rates To validate the generalizability of the proposed SyP across different missing rates, we conduct experiments on two distinct multi-modal datasets: the *Hateful Memes* dataset and the *Food101* dataset. The experimental results, presented in Figs. 4 and 5, provide valuable insights into the performance of SyP across various missing-modality scenarios, including text missing, image missing, and both missing.

On the *Hateful Memes* dataset, as depicted in Fig. 4, SyP variants consistently outperform baseline models across all missing rates (10%–90%). This significant improvement underscores the robustness of the proposed SyP in handling incomplete data. Notably, models trained both on a single missing modality and under missing-both conditions demonstrate strong and stable performance across all scenarios, especially in the text missing scenarios. This indicates that the combination of static and dynamic prompts enables flexible adaptation, particularly when trained with higher missing rates. Similarly, on the *Food101* dataset, SyP demonstrates remarkable performance. As shown in Fig. 5, SyP achieves the highest accuracy across all missing rates and scenarios. The dynamic adapter effectively adjusts the prompt weights,



(a) Text Missing (b) Image Missing (c) Both Missing

Figure 4. **Generalizability Analysis of Our Method to Different Missing Rates on Hate Memes dataset.** (a) Models are trained on missing-both or missing-text cases, and evaluated on missing-text cases with different missing rates. (b) Models are trained on missing-both or missing-image cases, and evaluated on missing-image cases with different missing rates. (c) All models are trained on missing-both cases, and evaluated on missing-both cases with different missing rates.



(a) Text Missing (b) Image Missing (c) Both Missing

Figure 5. **Generalizability Analysis of Our Method to Different Missing Rates on Food101 dataset.** (a) Models are trained on missing-both or missing-text cases, and evaluated on missing-text cases with different missing rates. (b) Models are trained on missing-both or missing-image cases, and evaluated on missing-image cases with different missing rates. (c) All models are trained on missing-both cases, and evaluated on missing-both cases with different missing rates.

Reduction Ratio	Hateful Memes (AUROC)	Food101 (Accuracy)	MM-IMDb (F1_Macro)
$r = 5$	67.98	**86.17**	54.72
$r = 10$	67.69	85.83	**55.02**
$r = 16$	**68.16**	85.58	53.98

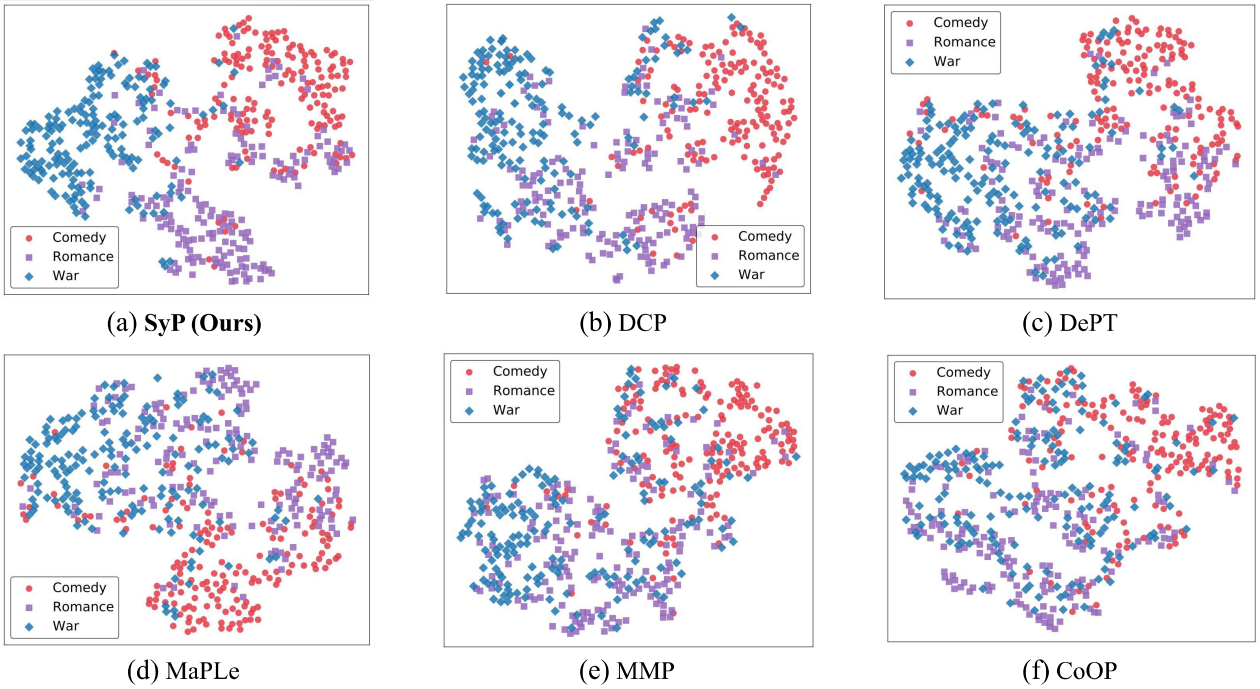
Table 2. Hyper-parameter analysis of reduction ratio r under a 50% both missing case on three datasets.

allowing the model to handle diverse missing-modality situations with high accuracy. These results highlight the adaptability and effectiveness of SyP in real-world applications where data may be noisy or incomplete.

Hyper-Parameter Analysis We analyze the impact of the reduction ratio r in the dynamic adapter under a 50% missing-modality scenario across three datasets: Hateful Memes, Food101, and MM-IMDb. As shown in Tab. 2, the reduction ratio r significantly influences model performance. For Hateful Memes, a ratio of 16 achieves the highest AUROC (68.16), indicating that larger ratios better capture subtle visual features. Food101 performs best with a ratio of 5 (86.17% accuracy), suggesting smaller ratios are more effective for fine-grained classification. For MM-IMDb, a ratio of 10 yields the highest F1-Macro (55.02), demonstrating that a balanced approach is optimal for text-heavy datasets. These results underscore the importance of tuning the reduction ratio to adapt the dynamic adapter to diverse task requirements, ensuring robust performance across varying data characteristics and modalities.

D. Visualization

Fig. 6 illustrates the t-SNE [7] visualization of the embedding distributions for three genres (Comedy, Romance, and War) in the MM-IMDb test set under a 50% both missing



(a) **SyP (Ours)** (b) DCP (c) DePT

(d) MaPLe (e) MMP (f) CoOP

Figure 6. t-SNE visualization of our model and other baselines on the MM-IMDb dataset under a 50% both missing rate.

rate. The points corresponding to ***SyP*** are more tightly clustered and clearly separated than those of DCP [6], DePT [9], MaPLe [2], MMP [4], and CoOp [10]. This indicates that the proposed SyP can effectively manage multi-modal missing problems. The model can accurately understand and reason when facing different modalities, maintaining high performance and stability. Specifically, the t-SNE visualization of ***SyP*** has several characteristics. First, the high concentration of points indicates the model's accurate recognition of similar samples. Second, the clear separation between different classes reduces misclassification. Third, the uniform distribution of points shows that the proposed SyP can balance the relationships between classes, leading to more stable model outputs. This demonstrates that the proposed SyP is better at maintaining robust genre distinctions, despite the challenge posed by missing data, showcasing its strength in dynamic multi-modal adaptation.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[2] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[3] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, pages 2611–2624, 2020.

[4] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[6] Tongkai Shi, Wei Feng, Fanhua Shang, Liang Wan, et al. Deep correlated prompting for visual recognition with missing modalities. *Advances in Neural Information Processing Systems*, pages 67446–67466, 2025.

[7] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[8] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo Workshops*, pages 1–6, 2015.

[9] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024.

[10] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 2337–2348, 2022.