
Appendix for Text as Any-Modality for Zero-Shot Classification by Consistent Prompt Tuning

A DETAILS OF PRE-TRAINED MULTIMODAL MODELS.

Our TaAM-CPT is built upon multimodal pre-trained models, including video-language model, image-language model, and audio-language model, and uses frozen text encoders for prompt tuning, as well as frozen modality encoders for object recognition predicting. In our work, we choose the pretrained multimodal models, open-sourced by the LAION (Schuhmann et al., 2022) organization, as the modality-aligned text and modality encoders. For a total of 300k text sentences on a single Tesla V100 for the Kinetic-400, MSCOCO, and ESC50 datasets, each epoch takes 12 minutes and the total training cost for 10 epochs is about 2 hours.

ViCLIP. ViCLIP is a video-language pretraining model, building upon the open-source CLIP of OpenAI. The model consists of a video encoder and corresponding text encoder, which is pretrained on the InternVid dataset containing 7 million videos, each with detailed text descriptions. We use the BASE architecture as our baseline model with 12 attention layers and 512 encoding dimensions.

CLIP. We select the open-source image-language pretraining model released by the LAION organization as our baseline model. The model comprises an image encoder and corresponding transformer-based text encoder, each with 12 attention layers and an encoding dimension of 512. The size of the input image is 224×224 , with the patch size being 32. For image modality, CLIP-ViT-B-32 (Cherti et al., 2023) is selected as the image encoder and image-text encoder.

CLAP. For the audio-language pretraining model, likewise, we select CLAP released by the LAION organization as our baseline model. The audio encoder is a transformer-based model with 4 groups of swin-transformer blocks, while the text encoder is RoBERTa. Two-layer MLPs with ReLU activation are applied to mAP both audio and text outputs into 512 dimensions. For audio modality, we select CLAP (Wu et al., 2023) from LAION (Schuhmann et al., 2022) as the audio encoder and the built-in Robert as the audio-text encoder.

B DETAILS OF DATASETS

B.1 VIDEO DATASETS

UCF101. UCF101 (Soomro et al., 2012) is a commonly used video classification dataset that contains 101 different action classes, each class contains approximately 100~300 video clips, and a total of 13,320 video clips. These video clips are collected from real data on YouTube, ranging in length from 10~30 seconds. We use all of the video data to evaluate our methods.

Kinetic-400. Kinetic-400 (Carreira & Zisserman, 2017) is a large-scale, high-quality video dataset collected from YouTube, including 400 human action classes. Each action class contains 450~1150 video clips, covering a wide range of classes, e.g., playing instruments, interactions between humans and objects, and handshakes. Each action has 250~1000 video clips for the training set, 50 video clips for the validation set, and 100 video clips for the test set. The validation set is used to evaluate our methods.

Kinetic-600. Kinetic-600 (Carreira et al., 2018) is an extension of the Kinetic-400 dataset, comprising approximately 480K video clips from 600 action classes. Each action class has at least 700 video clips. The dataset consists of 450~1000 video clips for training, 50 for validation, and 100 for testing per action class. The validation set is used to evaluate our methods.

Kinetic-700. Kinetic-700 (Carreira et al., 2019) is an extension of the Kinetic-600 dataset, covering 700 human action classes. Each action class has at least 700 video clips. Each video is a 10-second action clip extracted from original YouTube videos and labeled accordingly. There are a total of 650,000 video clips, with each action class comprising 450,100 video clips for training, 5,000 video clips for validation, and 1,000 video clips for testing. We use the validation set to evaluate our methods.

B.2 IMAGE DATASETS

MSCOCO. MSCOCO (Lin et al., 2014) is a large-scale computer vision dataset used for tasks such as object recognition, object detection, and image segmentation. It includes 80 image classes, 328,000 images, and 2,500,000 instances. It comprises 82,783 training images, 40,504 validation images, and 40,775 test images. We use the validation set to evaluate our methods.

VOC2007. VOC2007 (Everingham et al., 2010) is an image dataset containing 20 image classes that can be used to evaluate image classification, object detection, and image segmentation tasks. It consists of 9,963 images in total, with 5,011 images in the training set and 4,952 images in the test set. The test set is used to evaluate our methods.

VOC2012. VOC2012 (Everingham et al., 2010) dataset contains 20 classes, including people, animals, vehicles, indoor objects, and a background category, making a total of 20 classes. It can be used for evaluating image classification, object detection, and image segmentation tasks. It comprises 11,540 images, with 5,717 images in the training set and 5,823 images in the test set. The test set is used to evaluate our methods.

NUSWIDE. NUSWIDE (Chua et al., 2009) is an image dataset that contains 269,648 images collected from Flickr, with a total of 81 manually annotated concepts, including objects and scenes. It includes 161,789 images for the training set and 107,859 images for the validation set. We use the validation set to evaluate our methods.

ImageNet-mini. ImageNet-mini (Russakovsky et al., 2015) is derived from the ImageNet dataset and contains 100 classes with a total of 60,000 images, with 600 samples per class. The training and validation sets are typically divided into an 8:2 ratio by class. (For small sample classification, 64 classes are used for training, 16 for validation, and 20 for testing.) We use the test set to evaluate our methods.

Objects365. Objects365 (Shao et al., 2019) is a large object detection dataset that contains 638k images, 365 image classes, and 10,101k bounding boxes, far surpassing datasets like COCO. According to the paper’s annotation process, a total of 740k images were annotated, with 600k used for training, 38k for validation, and 100k for testing. We use the test set to evaluate our methods.

B.3 AUDIO DATASETS

ESC50. ESC50 (Piczak, 2015) is a standard dataset for environmental sound classification that contains 50 different environmental categories, each with 40 samples of up to 5 seconds in duration, totaling 2,000 samples. These samples cover a wide range of environments, such as animal sounds, traffic noise, indoor activities, etc. All samples are carefully balanced to ensure uniformity when training models. We use the validation set to evaluate our methods.

US8K. UrbanSound8k (Salamon et al., 2014) is a widely used open data set for automatic urban environment sound classification, which includes ten categories such as air conditioning sound and car horn sound. There are 8732 audio clips in the dataset with a length of about 4 seconds. The data set is divided into training and testing sets. We use the test set to evaluate our methods.

C TRAINING TEXT DATA CONSTRUCTION.

Here, we discuss the text training data construction for different modalities. We construct the following prompt template to input into LLaMA-2-7B for generating text description data.

TEMPLATE: Make several English sentences to describe a { **Modality** }. *Requirements: Generate 5 English sentences! Each sentence should be less than 25 words and includes: { **Labels** }.*

where { **Modality** } is replaced with video, audio, and image, { **Labels** } denotes the sampled classes. For video and audio datasets, which typically involve single classification tasks, we set the number of sampled categories to 2 to prevent too many categories from appearing in one sentence, which could interfere with the model’s learning of specific representations for each category. For image classification datasets, where multiple categories can appear on a single image, the number of sampled categories is set to 1, 2, 3, or 4 to ensure that the model not only learns the dependencies between image categories but also acquires independent representations for each category. As shown in Figure

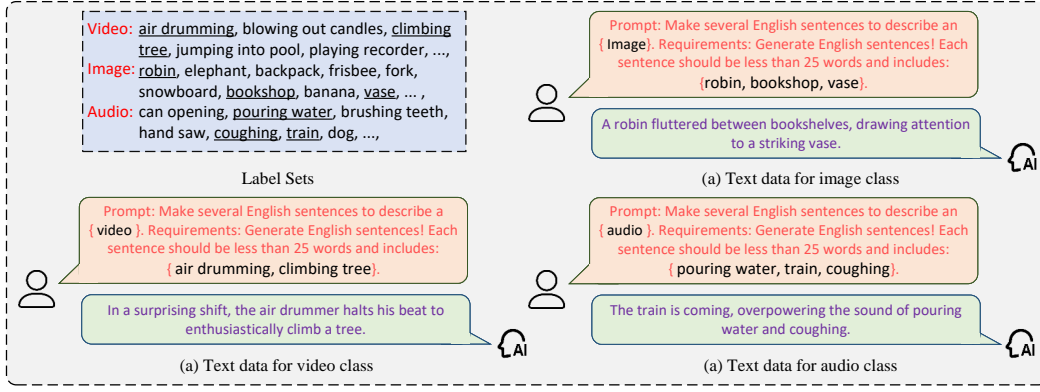


Figure 1: The candidate label set and text data generated by LLMs.

1, we randomly select several classes from the label set and construct a prompt template to query the LLMs to generate text data containing the semantic information of these classes.

D ABLATION STUDY

Prompt Design. Here, we mainly discuss the variants of consistent prompt tuning (CPT) in Table 1: a) Shared-Intra (1024), where the prompt is initialized as 1024-d vector and mapped to 512-d through a FC; b) Shared-Intra (512) represents initialization as a 512-d vector and then mapped to 512-d; c) Shared-Inter (512), where all prompts across all modalities share a FC and are mapped to 512-d. On Kinetic-400, we note a pronounced degradation of these variants. We believe the decline is mainly attributable to the numerous categories that are semantically proximate (e.g., *making pizza* and *making sandwich*). These phenomena are also observed in the MSCOCO and ESC50 datasets.

Unified Architecture. Our TaAM-CPT is designed as a general model toward unlimited modalities, exhibiting more robust object recognition capabilities than single modality-specific models. Table 7 presents the results of training each modality independently by intra-modal learning (e.g. $VP \checkmark$ with $\mathcal{L}_{Ia} \checkmark$), as well as the impact of applying the uni-directional contrastive learning (\mathcal{L}_{Ie}) across modalities. We can see that training single modality prompt by intra-modal learning has already yielded better results than the pre-trained models, and when all modalities are trained together, the performance of each modality can be further improved. In addition, applying uni-directional contrastive learning to guide the learning of video modality, not only improves the performance of the video modality but also enhances the object classification capabilities of the image and audio modalities.

Loss Weight. In this study, we design Ranking loss and uni-directional contrastive loss to perform intra-modal learning and inter-modal learning. The Ranking loss aims to learn class-specific prompt for each modality, while the contrastive loss is applied to guide the learning of weaker modalities (video) through those stronger ones (image and audio). Here, we explore the impact of setting different loss weights for these two loss functions. As shown in Figure 3, \mathcal{L}_{Ia} represents the Ranking loss for intra-modal learning, and \mathcal{L}_{Ie}

Table 1: Results of different prompt designs.

Prompt	K400	MSCOCO	ESC50
Shared-Intra (1024)	(43.1, 74.2)	55.4	90.6
Shared-Intra (512)	(47.5, 75.3)	58.7	91.9
Shared-Inter (512)	(50.1, 79.3)	62.2	92.1
TaAM-CPT(Ours)	(55.2, 80.4)	68.1	94.2

Table 2: Results of evaluating the unified architecture.

VP	IP	AP	\mathcal{L}_{Ia}	\mathcal{L}_{Ie}	K400	MSCOCO	ESC50
ZS-ViCLIP, CLIP, CLAP					(53.8, 78.7)	55.6	90.5
\checkmark	\times	\times	\checkmark	\times	(53.8, 78.9)	—	—
\times	\checkmark	\times	\checkmark	\times	—	65.8	—
\times	\times	\checkmark	\checkmark	\times	—	—	92.5
\checkmark	\checkmark	\checkmark	\checkmark	\times	(53.7, 79.1)	65.2	92.7
\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	(55.2, 80.4)	68.1	94.2

Table 3: Results of different loss weight between intra-modal learning and inter-modal learning.

\mathcal{L}_{Ia}	\mathcal{L}_{Ie}	K400	MSCOCO	ESC50
0.4	1.6	(54.9, 80.0)	67.9	94.0
0.8	1.2	(55.1, 80.2)	68.1	94.1
1.0	1.0	(55.2, 80.4)	68.1	94.2
1.2	0.8	(55.0, 80.2)	68.0	94.0
1.6	0.4	(54.5, 79.6)	68.0	93.9

represents the uni-directional contrastive loss for inter-modal learning. Our method achieves the best results when the weights of \mathcal{L}_{Ia} and \mathcal{L}_{Ie} are identical. Additionally, we notice that when the weight of \mathcal{L}_{Ie} is set to 1.0, 0.8 and 0.4, there is a significant decrease in top-1 and top-5 accuracy on the Kinetic-400 dataset, while the performance on MSCOCO and ESC50 datasets only suffer minor damage. This indicates that inter-modal learning greatly affects the learning of weaker modality, which is the video modality in this case.

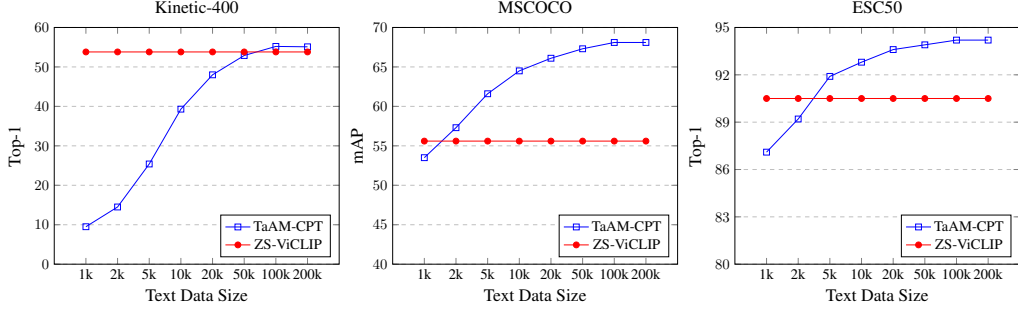


Figure 2: Results of different size of text training data on Kinetic-400, MSCOCO and ESC50 datasets.

Text Training Data Size. Our TaAM-CPT is trained with text data generated by LLMs instead of modality-specific labeled data. Therefore, we conduct various experiments with different sizes of text training data on the Kinetic-400, MSCOCO, and ESC50 datasets. As shown in Figure 2, on the Kinetic-400 dataset with text data size being 1k, the top-1 accuracy is only 9.8% due to the insufficient number of text data for each class, which hinders the learning of robust class-specific representations. However, as continuing to expand the scale of text training data, the corresponding text data for each class also increases gradually. When the text data reaches 100K, our TaAM-CPT outperforms ZS-ViCLIP. On the MSCOCO and ESC50 datasets, which contain 80 and 50 class labels, respectively, when the amount of text data is 5K, our method has already significantly surpassed ZS-CLIP and ZS-CLAP by 7% mAP and 2% top-1 accuracy. The performance on these two datasets begins to stabilize when the amount of text data is increased to 50K, indicating that datasets with more classes require a larger scale of text training data.

E VISUALIZATION OF INTRA-MODAL LEARNING.

Here, as shown in Figure 3, 4, 5, 6, 7, we present the more visualization results of the distribution of class-specific prompt learned by intra-modal learning on Kinetic-600/700, MSCOCO, ImageNet-mini, and ESC50 datasets.

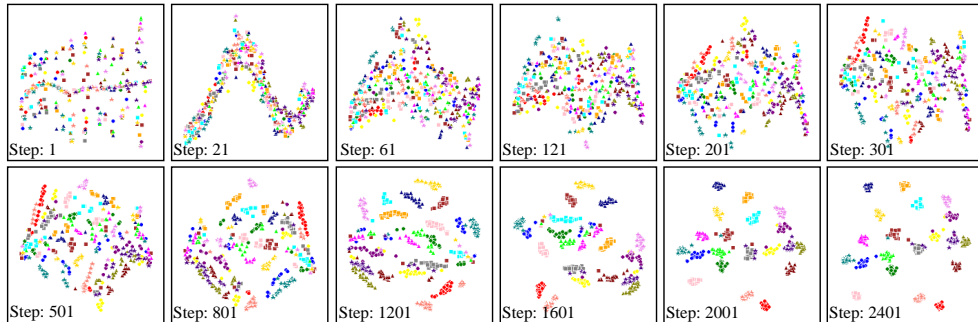


Figure 3: Visualization of the distribution of video prompt and video feature using t-SNE (van der Maaten & Hinton, 2008) for dimensionality reduction. We randomly select 20 video classes from the Kinetic-600 dataset.

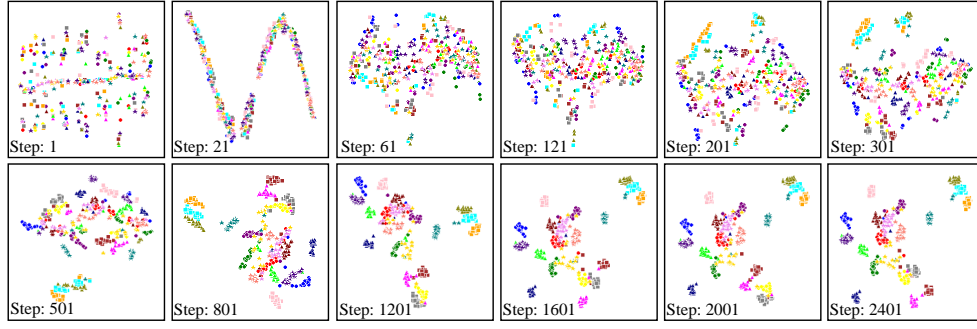


Figure 4: Visualization of the distribution of video prompt and video feature using t-SNE (van der Maaten & Hinton, 2008) for dimensionality reduction. We randomly select 20 video classes from the Kinetic-700 dataset.

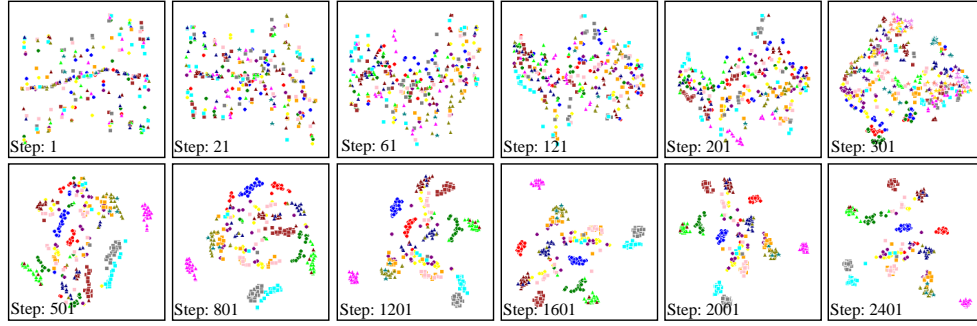


Figure 5: Visualization of the distribution of image prompt and image feature using t-SNE (van der Maaten & Hinton, 2008) for dimensionality reduction. We randomly select 20 image classes from the MSCOCO dataset.

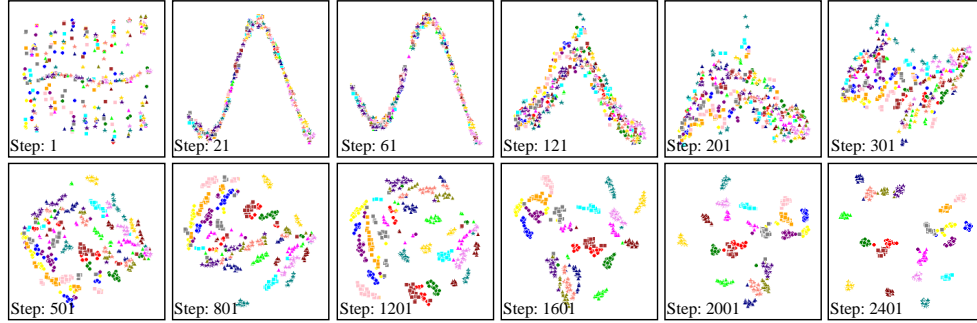


Figure 6: Visualization of the distribution of image prompt and image feature using t-SNE (van der Maaten & Hinton, 2008) for dimensionality reduction. We randomly select 20 image classes from the ImageNet-mini dataset.

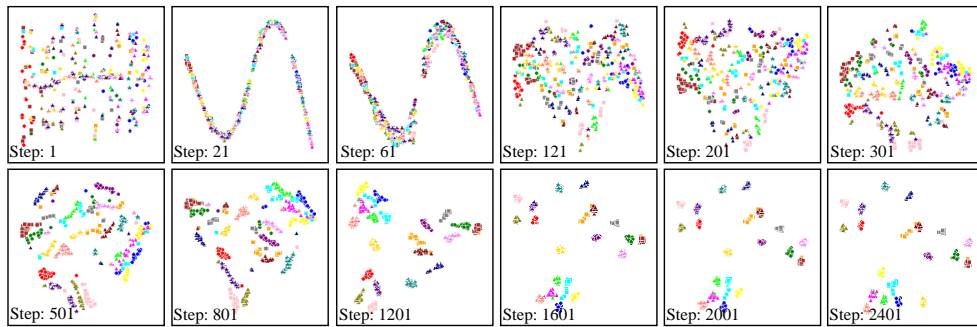


Figure 7: Visualization of the distribution of audio prompt and audio feature using t-SNE (van der Maaten & Hinton, 2008) for dimensionality reduction. We randomly select 20 audio classes from the ESC50 dataset.

REFERENCES

- Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 6299–6308, 2017.
- Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *CoRR*, 2018.
- Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *CoRR*, 2019.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pp. 2818–2829, 2023.
- Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009.
- Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes voc challenge. *IJCV*, 88(2):303–338, 2010.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, volume 8693, pp. 740–755, 2014.
- Karol J Piczak. Esc: Dataset for environmental sound classification. In *ACM MM*, pp. 1015–1018, 2015.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.
- Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *ACM MM*, pp. 1041–1044, 2014.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022.
- Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pp. 8430–8439, 2019.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(86):2579–2605, 2008.
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pp. 1–5, 2023.