Appendix

A SUMMARY

This appendix describes more details of the ICLR 2024 submission, titled *Meta-Transformer: A Unified Framework for Multimodal Learning*. The appendix is organized as follows:

- We detail utilizing Meta-Transformer on more modalities. § B.
- Then we further demonstrate the performance and merits of Meta-Transformer in dealing with multi-modal tasks (involving inputs from more than one modality to perform predictions) in § C.
- In addition, we conduct an ablation study and introduce more details of experiments on text, image, point cloud, audio, and other 8 modalities in § D.
- Beside these details, we also discuss the limitations of Meta-Transformer in § E.
- Last but not least, we discuss the impact of Meta-Transformer on the machine learning and computer vision community in § F.

B EXTENSIBILITY ON SINGLE-MODALITY PERCEPTION

In the main body of this paper, we illustrate that Meta-Transformer can simultaneously uncover the underlying patterns of natural language, 2D images, 3D point clouds, and audio spectrograms with the same network architecture and network parameters. Furthermore, we explore its ability in perceiving other modalities, like video recognition, infrared, X-Ray, and hyperspectral image recognition. In specific, we conduct experiments on UCF101 (Soomro et al., 2012) (**video**), RegDB (Nguyen et al., 2017) (**infrared** images), Chest **X-Ray** (Rahman et al., 2020), and Indian Pine (**hyperspectral** images) datasets.

B.1 VIDEO RECOGNITION

For video recognition, we follow VideoMAE (Tong et al., 2022) to modify the tokenizer by replacing the 2D embedding layer with a 3D embedding layer to simultaneously encode the spatial-temporal information from input frames. After tokenization, by leveraging the modality-shared encoder and task-specific heads, Meta-Transformer is able to extract high-level semantic features from videos and achieve favorable performance in the action recognition task of the UCF101 dataset.

Dataset. The UCF101 (Soomro et al., 2012) dataset is a common-used benchmark dataset for action recognition tasks. It is an extended version of UCF50 and contains 13,320 video clips of 101 categories. These 101 categories can be divided into 5 groups: Body motion, Human-human interactions, Human-object interactions, Playing musical instruments and Sports. All the input frames are with a resolution of 320×240 and a fixed frame rate of 25 FPS, collected from YouTube.

B.2 INFRARED IMAGE RECOGNITION

Infrared and hyperspectral image recognition poses unique challenges due to their specific characteristics. For infrared images, the Meta-Transformer framework could be adapted to capture thermal information by encoding temperature values alongside visual features, where the tokenizer for infrared images is the same as common RGB images.

Dataset. The RegDB (Nguyen et al., 2017) dataset focuses on evaluating the performance of infrared recognition algorithms in unconstrained and realistic scenarios. It includes variations in pose, expression, illumination, and occlusion. We conduct experiments on the RegDB dataset to evaluate the performance of Meta-Transformer on infrared recognition.

B.3 HYPERSPECTRAL IMAGE RECOGNITION

Similarly, for hyperspectral images, we expect that Meta-Transformer can also handle the high-dimensional spectral information by representing each spectral band in token embeddings. Compared with dealing with RGB images, the only modification is that we employ the new linear projection layer to replace the existing 2D convolution layer.

Dataset. The Indian Pine dataset is widely used in remote sensing and hyperspectral image analysis. It consists of 145×145 pixels with 145 spectral bands, which are captured in Indiana.

B.4 X-RAY IMAGE RECOGNITION

In addition, we explore the potential of the Meta-Transformer in medical image analysis. We leverage the tokenizer for RGB images here to encode raw medical images. Specifically, we conduct experiments regarding X-ray image analysis on the Chest X-Ray (Rahman et al., 2020) dataset. It is a collection of medical images commonly used for the analysis and diagnosis of various thoracic conditions. It comprises 7,000 X-ray images of the chest. The dataset is annotated with labels indicating the presence or absence of abnormalities such as lung diseases, fractures, and heart conditions.

C EXTENSIBILITY ON MULTI-MODALITY PERCEPTION

Since the modalities of text, image, point cloud, and audio are all involved in this paper, we did not conduct comprehensive multi-modal experiments as common practice such as Flamingo (Alayrac et al., 2022), OFA (Wang et al., 2022a), or BEiT-3 (Wang et al., 2022c). Instead, we conduct multi-modal experiments on a new and challenging task of Audio-Visual Segmentation (Zhou et al., 2022a), which is mainly focused on building an intelligent listener to align with fundamental visual tasks.

C.1 AUDIO-VISUAL SEGMENTATION

Audio-visual segmentation (Zhou et al., 2022a) refers to the task of segmenting objects from different audio sources within a referring image. It aims to develop algorithms that analyze both audio and visual signals simultaneously to identify and delineate distinct sources or events. It finds applications in fields like video conferencing, surveillance, multimedia analysis, and augmented reality.

We conduct experiments on the AVSS (Zhou et al., 2022a) dataset, which is recently released in the field of audio-visual research. It provides a comprehensive collection of audio and visual data captured in real-world scenarios. The dataset includes synchronized audio and visual recordings, featuring various events of human actions and natural sounds. In contrast to introducing multi-modal fusion modules as existing methods, Meta-Transformer directly concatenates visual and audio embeddings after Data-to-Sequence tokenization. After extracting representation, we employ a simple global average pooling layer to obtain the final representations of two modalities. Table 13 illustrates

Table 13: **Audio-Visual Segmentation with Meta-Transformer**. We conduct experiments on the AVSS (Zhou et al., 2022a) dataset, we report mIou (%) and F-score.

Method	**mIou (%)**	**F-score**	**Params**
AVSS (Zhou et al., 2022a) (ResNet-50)	20.18	0.252	$\tilde{8}$0M
AVSS (Zhou et al., 2022a) (ASPP)	28.94	-	$\tilde{1}$80M
AVSS (Zhou et al., 2022a) (PVT-v2)	29.77	0.352	$\tilde{1}$80M
Meta-Transformer	**31.33**	**0.387**	**86.5M**

the performance of Meta-Transformer and existing methods on the AVSS dataset for audio-visual segmentation. The evaluation metrics reported in this task are mIou and F-score. In comparison, Meta-Transformer outperforms all other methods with the highest mIou of 31.33% and the highest

F-score of 0.387. It also stands out for its significantly lower parameter count, with only 86.5 million parameters compared to the approximate 80M to 180M parameters of other methods.

Meta-Transformer offers several advantages over other methods in the field.

- **Unified architecture**. It relieves modality-specific encoders and reduces computation by leveraging a unified encode to process both audio and images, resulting in a more efficient and streamlined process.
- **Faster convergence**. Thanks to the unified architecture for processing both audio and images, the encoder can deeply align the two modalities instead of only at the output end, which leads to faster convergence. Meta-Transformer only needs 4 training epochs to reach 31.33% of mIou.
- **Superior performance**. Meta-Transformer achieves a significant improvement of 10% compared to other methods of a similar parameter scale.
- **Efficiency**. Despite its enhanced performance, Meta-Transformer achieves this with much fewer parameters, requiring only $1/3$ of the parameter amount, which makes forward and backward progress ease.

In summary, the benefits of employing the Meta-Transformer to deal with multi-modal tasks are appealing due to computational efficiency, rapid convergence, improved performance, and parameter efficiency. It reveals the significantly promising direction to apply Meta-Transformer to more multi-modal tasks.

D Experimental Details

Text understanding. For text understanding evaluation, we employ the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) which incorporates several different datasets, covering a wide range of natural language understanding tasks.

The comparison centers on paraphrasing, sentiment, duplication, inference, and answering tasks. When using frozen parameters pretrained on images, Meta-Transformer-B16$_{\text{F}}$ achieves scores of 54.6% in sentiment (SST-2), 81.1% in paraphrase (MRPC), 66.0% in duplication (QQP), 63.4% in inference (MNLI), and 56.3% in answering (QNLI) tasks.

Image understanding. 1) Classification: we conduct experiments on ImageNet-1K (Deng et al., 2009) which contains approximately 1.3 million images with 1000 categories. Following common practices (Wang et al., 2021b; Liu et al., 2021b; 2022c), base-scale models are trained for 300 epochs, while large models are pre-trained on ImageNet-22K (14.2 million images) for 90 epochs and fine-tuned on ImageNet-1K for another 20 epochs. 2) Object Detection: we conduct experiments on the MS COCO dataset (Lin et al., 2014) using Mask R-CNN (He et al., 2017) as the detector and training each model for 12 epochs. 3) Semantic Segmentation: we train the segmentation head UperNet (Xiao et al., 2018) on ADE20K (Zhou et al., 2017) for 160k iterations, providing a fair comparison with previous CNN-based and transformer-based backbones.

With the Meta-Transformer-B16$_{\text{F}}$ and Meta-Transformer-L14$_{\text{F}}$, achieving 69.3% and 75.3%, respectively. At the same time, when the pretrained parameters are further tuned, Meta-Transformer can outperform existing advanced methods.On object detection, Meta-Transformer-B16$_{\text{F}}$ and Meta-Transformer-L14$_{\text{F}}$ achieve APs of 31.7% and 43.5%, while Meta-Transformer-B16$_{\text{T}}$ and Meta-Transformer-L14$_{\text{T}}$ reach 46.4% and 56.3% AP, respectively. In semantic segmentation, the mIoUs for Meta-Transformer-B16$_{\text{F}}$ and Meta-Transformer-L14$_{\text{F}}$ are 33.4% and 41.2%, while Meta-Transformer-B16$_{\text{T}}$ and Meta-Transformer-L14$_{\text{T}}$ achieve 51.0% and 55.0%, respectively. In comparison, SwinV2-L/24‡ outperforms the Meta-Transformer in both object detection (58.8% AP) and semantic segmentation (55.9% mIoU). These results highlight that Meta-Transformer demonstrates a competitive performance in various image understanding tasks even compared to Swin Transformer (Liu et al., 2021b) and InternImage.

Infrared, X-Ray, and Hyperspectral data understanding. We conduct experiments on infrared image, X-Ray scan, and hyperspectral data recognition with RegDB (Nguyen et al., 2017), Chest X-Ray (Rahman et al., 2020), and Indian Pine [1] datasets, respectively.

Point cloud understanding. 1) Classification: to assess the performance of Meta-Transformer in 3D object classification, we use the ModelNet-40 (Wu et al., 2015) benchmark, consisting of CAD models across 40 classes, with 9,843 training samples and 2,468 validation samples. 2) Semantic segmentation: to evaluate performance in 3D point cloud segmentation, we assess the model on both S3DIS (Armeni et al., 2016) and ShapeNetPart (Yi et al., 2016) datasets. The S3DIS dataset encompasses 6 large indoor areas and 13 semantic classes, comprising 271 rooms. The ShapeNetPart dataset includes 16,880 object models across 16 shape categories.

When pretrained on 2D data, Meta-Transformer-B16$_{\text{F}}$ demonstrates competitive performance, achieving an overall accuracy (OA) of 93.6% on ModelNet-40 with only 0.6M trainable parameters, which is comparable to the best-performing models. On the S3DIS Area-5 dataset, Meta-Transformer outperforms other methods with a mean IoU (mIoU) of 72.3% and a mean accuracy (mAcc) of 83.5%, using 2.3M parameters. Moreover, Meta-Transformer excels in the ShapeNetPart dataset, achieving the highest scores on both instances mIoU (mIoU$_I$) and category mIoU (mIoU$_C$) with 87.0% and 85.2%, respectively, using 2.3M parameters.

Audio recognition. For audio recognition, we utilize the Speech Commands V2 (Warden, 2018) dataset, which consists of 105,829 one-second recordings of 35 common speech commands. Meta-Transformer-B16T model exhibits a significantly higher accuracy of 97.0% when tuning the parameters, whereas the AST model only reaches an accuracy of 92.6%. When AST is pre-trained on ImageNet and supplemented with additional Knowledge Distillation (KD), it achieves an improved performance of 98.1%, but with a higher number of trainable parameters of 86.9M. SSAST models display accuracy scores ranging from 97.8% to 98.0% while requiring 89.3M parameters. These results highlight that the Meta-Transformer performs competitively in the audio domain, demonstrating its versatility and effectiveness across different fields.

Video recognition. For video understanding, we conduct experiments on the UCF101 (Soomro et al., 2012) dataset for action recognition, with more details presented in § B.1.

Time-series forecasting. For time-series forecasting, we conduct experiments on ETTh1 (Zhou et al., 2021), Traffic[2], Weather[3], and Exchange (Lai et al., 2018) datasets. We use the tokenizer of Autoformer (Wu et al., 2021).

Graph understanding. We conduct experiments on the PCQM4M-LSC dataset (Hu et al., 2021), which is a large-scale dataset consisting of 4.4 million organic molecules with up to 23 heavy atoms with their corresponding quantum-mechanical properties. With the target of predicting molecular properties using machine learning, it has plenty of applications in drug discovery, and material science.

Tabular analysis. We conduct experiments on adult and bank marketing from UCI repository [4]. We use the tokenizer of TabTransformer (Huang et al., 2020) to encode raw tabular data.

IMU recognition. To evaluate the ability of Meta-Transformer to understand the inertial motion systems, we conduct experiments of IMU sensor classification on the Ego4D (Grauman et al., 2022) dataset.

D.1 Ablation study

we mainly conduct the ablation experiments, which are relevant to the depth of tuning transformer blocks, and pretraining on tokenizers as shown in Table 14 and Table 15.

[1] https://github.com/danfenghong/IEEE_TGRS_SpectralFormer/blob/main/data/IndianPine.mat

[2] https://pems.dot.ca.gov/

[3] https://www.bgc-jena.mpg.de/wetter/

[4] http://archive.ics.uci.edu/ml/

Models	Pretrained Tokenizer	Modality	Performance (%)
Meta-Transformer-B16	From Scratch	Video	54.22
Meta-Transformer-B16	VideoMAE	Video	57.11
Meta-Transformer-B16	From Scratch	Image	85.42
Meta-Transformer-B16	MAE	Image	85.93

Table 14: Ablation study on tokenizer components.

Models	Transformer Depth	ImageNet-1K (%)
Meta-Transformer-B16	1	42.74
Meta-Transformer-B16	2	58.91
Meta-Transformer-B16	4	75.63
Meta-Transformer-B16	8	83.98
Meta-Transformer-B16	12	85.42

Table 15: Ablation study on fine-tuning transformer blocks.

Our code is built on open-source projects including MMClassification[5], MMDetection[6], MMsegmentation[7], OpenPoints[8], Time-Series-Library[9], Graphomer [10].

We sincerely thank their great contributions. More implementation details can be found in our source code.

E LIMITATION

From the perspectives of complexity, methodology, and further application, the limitations of the Meta-Transformer are summarized as follows:

Complexity: Meta-Transformer requires $\mathcal{O}(n^2 \times D)$ computation dealing with token embeddings $[\boldsymbol{E}_1, \cdots, \boldsymbol{E}_n]$. High memory cost and heavy computation burden make it difficult to scale up.

Methodology: Compared with Axial Attention mechanism in TimeSformer (Bertasius et al., 2021) and Graphormer (Ying et al., 2021), Meta-Transformer lacks temporal and structural awareness. This limitation may affect the overall performance of Meta-Transformer in tasks where temporal and structural modeling plays a critical role, such as video understanding, visual tracking, or social network prediction.

Application: Meta-Transformer primarily delivers its advantages in multimodal perception. It's still unknown about its ability for cross-modal generation. We will work on this in the future.

F FURTHER IMPACT DISCUSSION

F.1 MODALITY-FREE PERCEPTION

We hope that Meta-Transformer can introduce new insight into both multi-modal learning and multi-modal generation fields. Meta-Transformer enables the usage of a shared encoder to encode diverse modalities, e.g. natural language, 2D images, 3D point clouds, as well as audio spectrograms., and project them into a shared representation space. This naturally reduces the modality gap across

[5]https://github.com/open-mmlab/mmpretrain/tree/mmcls-1.x

[6]https://github.com/open-mmlab/mmdetection

[7]https://github.com/open-mmlab/mmsegmentation

[8]https://github.com/guochengqian/openpoints

[9]https://github.com/thuml/Time-Series-Library

[10]https://github.com/microsoft/Graphormer

modalities and mitigates the burden of cross-modal alignment. In addition, Meta-Transformer removes the need for paired training data (such as image-text pairs), thus endowing multi-modal learning with more training flexibility.

F.2 Application Prospects

We investigate the application of Meta-Transformer on a wide range of modalities including RGB images, text, point clouds, video understanding, remote sensing (hyper-spectral images), nighttime surveillance (infrared images), and medical analysis (X-Ray images).

In video understanding, Meta-Transformer reveals the potential of enhancing the analysis and interpretation of videos by integrating information from text, audio, and image with the shared encoder. This benefits tasks such as action recognition, event detection, and video summarization. Meta-Transformer's capability to handle video-related modalities paves the way for improved video understanding applications in areas like video surveillance, video indexing, and content-based video retrieval.

In hyperspectral imaging for remote sensing, Meta-Transformer enables the analysis and understanding of hyperspectral data by extracting high-level semantic features. It enhances tasks such as classification, target detection, and land cover mapping, improving the accuracy and efficiency of remote sensing applications. The ability to process hyperspectral images using Meta-Transformer opens doors for advancements in environmental monitoring, agriculture, urban planning, and disaster management.

In medical applications, particularly X-ray image analysis, Meta-Transformer offers a promising approach to improving diagnostic accuracy and efficiency with multi-modal information. It can effectively capture and fuse information from X-ray images, clinical data, and other modalities to aid in disease detection, anomaly identification, and treatment planning by leveraging its unified learning framework. Meta-Transformer's capability to handle multi-modal data enhances the potential for more accurate and comprehensive medical imaging analysis, leading to better patient care and outcomes.

For infrared images used in nighttime recognition and surveillance, Meta-Transformer's ability to process infrared data helps extract crucial information for object detection, tracking, and recognition in low-light conditions, which opens an avenue for advancements in nighttime surveillance, security systems, and autonomous navigation in challenging environments with the cooperation between infrared cameras with RGB cameras.

F.3 Conclusion

In summary, we think that the ability of Meta-Transformer to unify multi-modal learning comes from that *neural network architectures can learn modality-invariant patterns*. The architecture of Meta-Transformer illustrates the advantages of length-variable token embeddings in multi-modal learning, which provides flexible but unified forms of multi-modal semantics. Then it's time to think about designing algorithms to train networks that generalize on *unseen* modalities. Meanwhile, it's also intriguing to design the architecture of a unified multi-modal decoder, which can decode representations into any form of a specific modality.

Although Meta-Transformer presents a surprising performance and shows a new promising direction in multi-modal perception, we are not sure whether the proposed architectures are also effective in generative tasks. And it remains mysterious how to develop modality-invariant generative models. We hope that this can inspire future research.