Supplementary material
OmniVec2 - A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning

Siddharth Srivastava, Gaurav Sharma
Typeface
{siddharth.srivastava, gaurav}@typeface.ai

1. More implementation Details

Task specific tokenizers. Each tokenizer's output is directed to a shared transformer encoder. Text is processed using a BPE tokenizer [37], similar to the approach in Uniperceiver-v2, converting text into word embeddings. Image modalities like RGB, infrared, and X-Ray are tokenized using an image patch tokenizer [11]. Video processing employs the method from [6], while point clouds are handled as per [48]. For audio, the spectrogram is tokenized using the same technique as for images [11]. Time series data tokenization follows [44], and for tabular data, the approach from [23] is utilized.

Task heads. For the heads of downstream tasks, we employ ViT-Tiny, coupled with standard loss functions tailored to various tasks. More specifically, we utilize different loss functions depending on the task: (i) For classification involving images, text, and point clouds, we follow the approach outlined in [10]. For video, the methodology from [4] is applied, and for audio, we adopt the loss function from [21]. (ii) In the case of image and point cloud segmentation tasks, we utilize the loss function described in [35]. (iii) For text summarization tasks, we incorporate the strategies from [38].

Task balancing strategy. We follow [13] for task balancing, and introduce noise parameter for each task and loss. We also experimented with several other task balancing mechanisms such as Equal Weighting, Nash-MTL [32], Random Loss Weighting [28], and observed similar performances across tasks and datasets.

2. More ablations

We conduct experiments to demonstrate the impact of task balancing leading to our method enabling random selection of modalities, instead of careful selecting them for pretraining. We report the results using fine-tuning and pretraining the full multimodal multi task pretraining strategy. The results are shown in Table 1.

Method	Modality/Task Selection	Task Balancing	iN2018	K400	ESC50
Ours-1	Pairs	No	81.8	84.1	87.4
Ours-2	Random	No	78.2	74.6	82.3
Ours-3	Pairs	Yes	90.9	89.5	95.8
Ours-4	Random	Yes	**94.6**	**93.6**	**99.1**

Table 1. Ablation on modality selection and task balancing

Impact of random vs paired modality and task selection. We compare against the paired task and modality selection strategy of OmniVec, while keeping the architecture as our method (row 1) i.e. pairs of modalities and tasks are carefully chosen vs. when the modalities and tasks are chosen randomly (row 2). We observe that with random selection results in a relatively inferior performance on all the three datasets which belong to different modalities (image, video, audio).

Impact of task balancing. We can observe from that if task balancing is enabled, then the performance of random selection of modalities and task (row 4) is better than if we select the pairs carefully (row 3). This could potentially be due to careful selection introducing bias, whereas, random selection allows exploring multiple combinations of modalities and task, where task balancing enables leveraging the varying complexity of modalities and tasks.

Performance of pretrained model against similar methods. We report results using the multitask multimodal training in Table: 3. To adapt to the tasks, we follow the settings in earlier works [20, 38]. We observe that our method, with only pretraining, performs better than the competing methods having capability to process multiple modalities, on all the compared datasets.

Impact of feature fusion strategy. We evaluate the proposed method on three additional feature fusion strategies, (i) addition (ii) average pooling (iii) max pooling. The results are shown in Tab. 2. We observe that our cross attention based method significantly outperforms other feature fusion strategies on the evaluated datasets.

Dataset	Metric	CA	Add.	Avg. Pool	Max Pool
iNaturalist-2018	Top-1 Acc.	**69.9**	56.2	58.5	60.4
YouCook2	Recall@10	**94.6**	80.2	83.2	86.8

Table 2. Comparison of various feature fusion strategies, cross-attention (CA), addition (Add.), Average Pooling (Avg. Pool), Max pooling (Max Pool).

3. More experimental results

Finetuning on pretraining datasets. We report result by fine tuning the complete model on respective training sets with corresponding tasks, and report result in Table 4. We observe that the proposed method outperforms the state of the art method on these datasets.

Adaptation on unseen datasets. We report detailed results and comparison to state of the art methods on UCF-101, HMDB51, Oxford-IIIT Pets, ScanObjectNN, NYUv2, SamSum datasets in Tables 4-9. The proposed method outperforms the competing methods on all of these datasets.

Adaptation on unseen modalities. We include detailed results on Tabular data in Table 11, time-series data in Table 12, and X-Ray recognition in Table 14. We observe that we achieve state of the art results on X-Ray image recognition and time series data. On Tanular data we outperform competing methods on Adult dataset, while achieve second best performance on Bank Marketing dataset.

Adaptation on additional datasets. We also report results on ADE-20K (Table 15), and MS-COCO dataset (Table 16), using the settings explained in 'Section 4-Adaptation on unseen datasets' (main manuscript). We observe that we lag behind only by a margin of $\sim$ 10% on these, despite using only 10% of the training dataset. Further, both of these datasets, contain images which are considered difficult and contain closer to noise in the real world data, hence demonstrating that our method not only provides good generalization but also adapts relatively better than competing method with significantly lesser in-domain training data.

Cross-modal generalization on additional datasets. In Table 13, we observe that the proposed method outperforms the state of the art methods on respective datasets i.e. 1.2% on VGGSound and 0.9% on AVSBench.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 3, 4

[2] Maoet al. Multimodal variational auto-encoder based audio-visual segmentation. In *ICCV*, 2023. 4

[3] Faris Almalik, Mohammad Yaqub, and Karthik Nandakumar. Self-ensembling vision transformer (sevit) for robust medical image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 376–386. Springer, 2022. 4

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1

[5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 3

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 1

[7] Joao Carreira, Skanda Koppula, Daniel Zoran, Adria Recasens, Catalin Ionescu, Olivier Henaff, Evan Shelhamer, Relja Arandjelovic, Matt Botvinick, Oriol Vinyals, et al. Hierarchical perceiver. *arXiv preprint arXiv:2202.10890*, 2022. 3

[8] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *arXiv preprint arXiv:2305.11487*, 2023. 4

[9] Z Chen, Y Duan, W Wang, J He, T Lu, J Dai, and Y Qiao. Vision transformer adapter for dense predictions. arxiv 2022. *arXiv preprint arXiv:2205.08534*. 4

[10] 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. 1, 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[12] Chen et al. Vggsound: A large-scale audio-visual dataset. In *ICASSP*. IEEE, 2020. 4

[13] Kendall et al. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 1

[14] Zhou et al. Audio–visual segmentation. In *ECCV*, pages 386–403. Springer. 4

[15] Zhu et al. Multiscale multimodal transformer for multimodal action recognition. 2022. 4

[16] Zhang et al. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 4

[17] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 4

Method/Dataset	Supp. Modalities	AudioSet (A+V.)	AudioSet (A)	SSv2	GLUE	ImageNet1K	Sun RGBD	ModelNet40
Omni-MAE [19]	Image, Video	-	-	73.4	-	85.5	-	-
Perceiver [25]	Modality Agnostic	43.4	38.4	-	-	78.6	-	-
Heirarchical Perceiver [7]	Modality Agnostic	43.8	41.3	-	-	81.0	-	80.6
data2vec [5]	Modality Agnostic	-	34.5	-	82.9	86.6	-	-
Omnivore [20]	Image, Video, Depth map	-	-	71.4	-	84.0	65.4	-
VATT [1]	Image, Video, Audio, Text	-	39.4	-	-	-	-	-
Perceiver IO [24]	Modality Agnostic	-	-	-	-	79.0	-	77.4
OmniVec [38]	Image, Video, Audio, Text, Depth map, Point Clouds	48.6	44.7	80.1	84.3	88.6	71.4	83.6
Ours (pretrained)	Modality agnostic (w/ tokenizers)	**51.6**	**47.1**	**83.2**	**87.1**	**89.3**	**74.6**	**86.2**

Table 3. **Comparison of our framework with similar methods that work on multiple modalities**. We compare our method with masked pretraining with the best reported results from respective publications of the compared methods. Supp. Modalities indicates the modalities supported by respective methods. Our method supports any modality (modality agnostic) if a suitable tokenizer is present. It could also support universal tokenizers similar to Meta-Transformer, Autoformer, with slight drop in performance compared to base method as discussed in Sec. 4.5 (main manuscript)

Dataset	Metric	Ours	SOTA
AudioSet(A)	mAP	55.8	54.8 (OmniVec [38])
AudioSet(A+V)	mAP	56.4	55.2 (OmniVec [38])
SSv2	Top-1 Acc	86.1	85.4 (OmniVec [38])
ImageNet1K	Top-1 Acc	93.6	92.4 (OmniVec [38])
Sun RGBD	Top-1 Acc	75.9	74.6 (OmniVec [38])
ModelNet40	Overall Acc	97.1	96.6 (OmniVec [38])

Table 4. **Comparison with state of the art** after fine tuning on respective training sets of datasets used for pretraining the network.

[18] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 4

[19] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022. 3

[20] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 1, 3, 4

[21] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 1

[22] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1451–1459, 2021. 4

[23] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020. 1

[24] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 3

[25] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 3

[26] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020. 4

[27] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *NeurIPS*, 2019. 4

[28] Baijiong Lin, Feiyang Ye, Yu Zhang, and Ivor W Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *arXiv preprint arXiv:2111.10603*, 2021. 1

[29] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 4

[30] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *ICLR*, 2021. 4

[31] Yahui Liu, Bin Tian, Yisheng Lv, Lingxi Li, and Feiyue Wang. Point cloud classification using content-based transformer via clustering in feature space. *arXiv preprint arXiv:2303.04599*, 2023. 4

[32] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022. 1

[33] Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[34] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. *arXiv preprint arXiv:2302.02318*, 2023. 4

[35] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of*

Method	U101
VATT [1]	87.6
Omnivore [20]	98.2
Text4Vis [45]	98.2
SMART [22]	98.6
VideoMAE V2-g [40]	99.6
OmniVec [38]	99.6
Ours	**99.7**

Table 5. **UCF-101 Action Recognition**. Metric is 3-fold accuracy.

Method	HMDB51
VATT [1]	66.4
DEEP-HAL [39]	87.56
VideoMAE V2-g [40]	88.10
OmniVec [38]	91.6
Ours	**92.1**

Table 6. **HMDB51 Action Recognition**. Metric is 3-split accuracy.

Method	Pets (top-1)	Pets (top-5)
Omnivore [20]	95.1	99.1
IELT [46]	95.28	-
DINOv2[33]	96.70	-
EffNet-L2 [18]	97.10	-
OmniVec [38]	99.2	**99.7**
Ours	**99.3**	**99.7**

Table 7. **Fine grained image classification on Oxford-IIIT Pets dataset**. The metrics are top-1 and top-5 accuracy.

Method	SO-NN
PointConT [31]	90.3
ReCon [34]	91.3
ULIP-2 [47]	91.5
PointGPT[8]	93.4
OmniVec [38]	96.10
Ours	**96.9**

Table 8. **ScanObjectNN 3D point cloud classification**. Metric is Overall Accuracy.

Method	NYUv2
Omnivore [20]	56.8
CMN [29]	56.9
OmniVec [38]	60.8
Ours	**62.5**

Table 9. **NYU v2 semantic segmentation**. Metric is mean IoU.

Method	R-1	R-2	R-L
Pegasus [50]	54.37	29.88	45.89
MoCa [49]	55.13	30.57	50.88
OmniVec [38]	58.81	31.1	53.4
Ours	**59.3**	**32.7**	**54.8**

Table 10. **SamSum dataset meeting summarization**. Metric is ROGUE scores.

Method	Adult Accuracy	Bank Marketing Accuracy
LightGBM	87.8	-
Tabmlp	87.2	-
Tabnet	87.0	-
Tabtransformer	87.1	**93.4**
Meta-Transformer-B16F [16]	85.9	90.1
Ours	**88.1**	92.3

Table 11. **Tabular data** understanding. We report Accuracy (%).

Method	Exchange	ETTh1	Traffic	Weather
Pyraformer [30]	0.827	0.878	0.946	1.913
Informer [51]	1.040	0.764	0.634	1.550
LogTrans [27]	1.072	0.705	0.696	1.402
Meta-Transformer [52]	0.994	0.694	0.797	1.430
Reformer [26]	1.029	0.741	0.803	1.280
Ours	**0.399**	**0.601**	**0.210**	**0.330**

Table 12. **Time Series data**. We report MSE.

Dataset	Task	Metric	SoTA	Ours
VGGSound [12]	Audio-Visual Classification	Top-1 Acc.	66.2 [15]	**68.4**
AVSBench(S4) [14]	Audio Visual Segmentation	mIoU	81.74 [2]	**82.62**

Table 13. Cross-modal generalization on more modalities and tasks.

Method	Accuracy
ViT [10]	96.3
SEViT [3]	94.6
Meta-Transformer-B16F [16]	94.1
Ours	**98.1**

Table 14. **X-ray image recognition**. We conduct experiments on the ChestX-Ray dataset. We report the Accuracy (%)

Method	mIoU
BEiT-3[41]	**62.8**
EVA [17]	61.5
FD-SwinV2-G [43]	61.4
ViT-Adapter-L [9]	58.4
Ours	58.5

Table 15. **Adaptation to Semantic Segmentation**. We conduct experiments on the ADE-20K dataset. We report the mIoU. Unlike competing methods, we finetune our method using 10% of the respective training set.

Method	mIoU
Co-DETR [52]	**66.0**
InternImage-H [42]	65.4
Focal-Stable-DINO [36]	64.8
Ours	60.1

Table 16. **Adaptation to Object Detection**. We conduct experiments on the MSCOCO dataset. We report the mAP. Unlike competing methods, we finetune our method using 10% of the respective training set.

the IEEE/CVF International Conference on Computer Vision, pages 12179–12188, 2021. 1

[36] Tianhe Ren, Jianwei Yang, Shilong Liu, Ailing Zeng, Feng Li, Hao Zhang, Hongyang Li, Zhaoyang Zeng, and Lei Zhang. A strong and reproducible object detector with only public datasets. *arXiv preprint arXiv:2304.13027*, 2023. 4

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv*

preprint arXiv:1508.07909, 2015. 1

[38] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. *arXiv preprint arXiv:2311.05709*, 2023. 1, 3, 4

[39] Lei Wang and Piotr Koniusz. Self-supervising action recognition by statistical moment and subspace descriptors. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4324–4333, 2021. 4

[40] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 4

[41] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 4

[42] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 4

[43] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 4

[44] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In *NeurIPS*, 2021. 1

[45] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. *Proceedings of the AAAI, Washington, DC, USA*, pages 7–8, 2023. 4

[46] Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, 2023. 4

[47] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martin, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023. 4

[48] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022. 1

[49] Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei. Momentum calibration for text generation. *arXiv preprint arXiv:2212.04257*, 2022. 4

[50] Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022. 4

[51] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021. 4

[52] Z Zong, G Song, and Y Liu. Detrs with collaborative hybrid assignments training. arxiv 2022. *arXiv preprint arXiv:2211.12860*, 2022. 4