# UnIVAL: Unified Model for Image, Video, Audio and Language Tasks

Anonymous ICCV submission

Paper ID 1

## Abstract

*This work proposes to investigate the following question: is it possible to efficiently build a unified model that can support all modalities?. To this end, we propose UnIVAL, a step further towards this ambitious goal. Without relying on fancy datasets sizes or models with billions of parameters, the $\sim$ 0.25B parameter UnIVAL model goes beyond two modalities and unifies text, images, video, and audio into a single model. With better multitask pretraining paradigm, based on task balancing and multimodal curriculum learning, UnIVAL shows competitive performance to existing state-of-the-art approaches. The representation learned from image and video-text modalities, allows the model to achieve competitive performance on audio-text tasks, despite not being pretrained on audio. Thanks to the unified model, we propose a novel study on multimodal model merging via weight interpolation, showing its benefits for out-of-distribution generalization. The model weights and the code will be open-source.*

## 1. Introduction

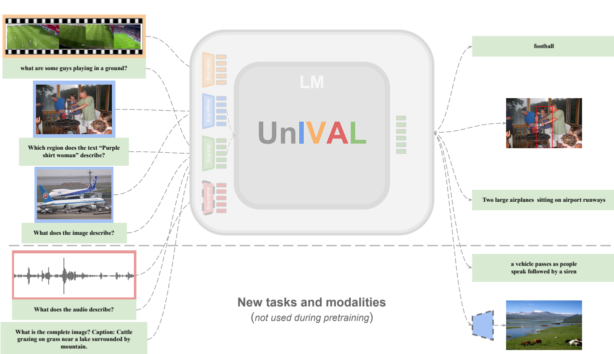


**Figure 1:** UnIVAL. A seq-to-seq unified model for multimodalities.

Large Language Models (LLMs) have made significant advances in text understanding and generation tasks, based on the Transformer architecture and a single next-token prediction objective. However, their current limitation to text modality restricts their broader understanding and interaction with the world. To address this, recent research has focused on developing multimodal models [4] that surpass task-specific approaches, particularly in image-text tasks. While some progress has been made in incorporating more than two modalities [1], such as image/video-text, the majority of current small to mid-scale vision-language models [17, 3] are still specialized per modality, rely on task-specific modules and have limited support for downstream tasks due to variations in input/output formats. Large-scale approaches, such as Flamingo [1], have helped alleviate these limitations to some extent. Promising advancements have been made with sequence-to-sequence models (OFA [20]), which support a wide range of image and image-text tasks. Similarly, LAVENDER [10] has unified pretraining tasks as Masked Language Modeling (MLM) for video-text tasks. However, these models are still confined to a maximum of two modalities, either image-text or video-text.

Unified models offer numerous advantages. (a) They harness the collaborative strengths of different pretrained tasks, facilitating knowledge transfer across various tasks and modalities. (b) They can seamlessly handle new tasks or modalities, due to the unified input/output format. (c) They benefit from a wide range of diverse data, enabling them to generalize effectively to novel tasks and modalities. Moreover, (d) these models are straightforward to scale and manage, simplify training objectives and input/output format, and involve a single model without the need for task-specific modules/heads.

Once pretraining is done, the model can be finetuned on many different datasets, producing many models with the same set of parameters, each specialized in a particular task. The shared pretraining and unified architecture of all these finetuned models pave the way to recycle, repurpose and leverage (*e.g.* by merging different models [14]) the collaboration between diverse skills across tasks and modalities, to obtain new models that are more robust and generalize better. Thus, in addition to multitask pretraining, merging different finetuned models is another way to leverage the diversity of multimodal tasks.

Here, we ask the following question.

*Is it possible to efficiently build a unified model that can support all modalities?*

A positive answer will pave the way for building generalist models that can potentially solve any task.

To answer this question, we propose UnIVAL, a step further towards generalist modality-agnostic models. UnIVAL (illustrated in Fig.1) goes beyond two modalities and unifies text, images, video, and audio into a single model. Our main contributions are two-fold: (a) To the best of our knowledge, UnIVAL is the first model, with unified architecture, vocabulary, input/output format and training objective, that is able to tackle image, video and audio language tasks, without relying on large scale training or large model size. Our model achieves competitive performance to existing modality-customized work and new SoTA on some tasks (*e.g.* +1.4/+0.98/+0.46 points accuracy on RefCOCO/RefCOCO+/RefCOCOg Visual Grounding, +3.4 CIDEr on Audiocaps), (b) We show the benefits of multimodal curriculum learning with task balancing, for efficiently training the model beyond two modalities. (c) thanks to our unified model, we propose an novel study on multimodal model merging via weight interpolation. We show that, even when the model is trained with different multimodal tasks, weight interpolation can effectively combine the skills of the different models and improve out-of-distribution generalization, without any inference overhead. (d) We show the importance of multitask pretraining, and study the synergy and knowledge transfer between pretrained tasks and modalities. In addition, we find that pretraining on more modalities makes the model generalizes better to new ones. In particular, without any audio pretraining, UnIVAL is able to attain competitive performance to SoTA when finetuned on audio-text tasks.

## 2. Pretraining of UnIVAL

Our model's core is an encoder-decoder LM [9] model designed to process abstract representations. To optimize data and compute requirements, we map different modalities, using lightweight CNN-based encoders, to a shared representation space, before feeding them into the encoder of the LM. In the following we describe how we pretrain our model.

**Unifying tasks and input/output format.** To train a single model on many tasks, a unified representation of these tasks is necessary. As our model's core is a language model, we transform all tasks into a sequence-to-sequence format, where each task is specified by a textual prompt (*e.g.*, "what does the video describe?" for video captioning). The input/output of all tasks consists of sequence of tokens, where we use a unified vocabulary that contains text, location and discrete image tokens. For pretraining tasks, we pretrain only on relatively small public datasets, such as image captioning (COCO, Visual Genome, SBU, CC3M and CC12M (only in the first stage)), VQA (VQAv2, GQA, VG), Visual Grounding and referring expression comprehension (RefCOCO, RefCOCO+, RefCOCOg), video captioning (WebVid2M) and video question answering (WebVidQA). Note that we only use the training sets of different benchmarks during pretraining.

| Model | COCO Captioning test CIDEr | VQAv2 test-std Acc. | RefCOCO testB Acc. | RefCOCO+ testB Acc. | RefCOCOg test-u Acc. |
|---|---|---|---|---|---|
| UniTAB [22] | 119.8 | 71.0 | 83.75 | 71.55 | 84.70 |
| GIT-L [19] | 138.5 | 75.5 | - | - | - |
| $OFA_{Base}$ [20] Our Ft | 138.1 | 77.1 | 83.30 | 74.29 | 82.31 |
| OmniVL [18] | 133.9 | 78.4 | - | - | - |
| UnIVAL (ours) | 137.0 | 77.1 | 85.16 | 75.27 | 85.16 |

**Table 1: Finetuning on Image-Captioning.**

**Unifying training objective.** We follow other approaches and optimize the model for conditional next token prediction.

Besides the unification of our model, in the following, we detail different techniques that lead to more efficient pretraining.

**Multimodal Curriculum Learning (MCL).** Other works train the model on all tasks and modalities simultaneously [20]. However, we have observed that models trained on more modalities tend to exhibit better generalization to new ones. To capitalize on this insight, we employ a different strategy wherein we gradually introduce additional modalities during training. This approach facilitates a smoother transition to new modalities by providing a better initialization for the newly added modality. Furthermore, this paradigm significantly reduces computational requirements compared to training on the entire dataset at once.

## 3. UnIVAL on downstream tasks

Due to space constraints, we discuss only part of the experiments.

**Image-Text tasks.** We evaluate the model for VQA, Image Captioning, and Visual Grounding (VG). For VQA and VG we cast the task as sequence generation (bbox location for VG). Table 1 shows that we achieve SoTA results on VG compared to all previous approaches. On VQA and captioning, we obtained comparable performance to other work, including the previous unified OFA model.

| Method | #PT im./vid. | MSRVTT-QA (Acc.) | MSVD-QA (Acc.) | MSR-VTT (CIDEr) |
|---|---|---|---|---|
| MERLOT [23] | -/180M | 43.1 | - | - |
| VIOLET [7] | 3.3M/182M | 43.9 | 47.9 | - |
| OmniVL [18] | 14M/2.8M | 44.1 | 51.0 | - |
| GIT-L [19] | 14M/- | 42.7 | 55.1 | 64.1 |
| MV-GPT$^T$ [16] | -/53M | - | - | 60.0 |
| LAVENDER [10] | 14M/14.4M | 45.0 | 56.6 | 60.1 |
| UnIVAL (ours) | 14M/2.5M | 43.48 | 49.55 | 60.5 |

**Table 2: Finetuning for VideoQA and Video Captioning**. The text-generation based UnIVAL model is competitive with SoTA models customized for videos or trained on significantly larger datasets.

**Video-Text tasks.** We evaluate the model for VideoQA and Video Captioning. Table 2 shows that UnIVAL attains competitive scores compared to other task/video-customized approaches, trained on larger datasets.

| Dataset | Method | $BLEU_1$ | $BLEU_2$ | METEOR | CIDEr | SPICE |
|---|---|---|---|---|---|---|
| Audiocaps | [12] | 0.647 | 0.488 | 0.222 | 0.679 | 0.160 |
| | [11] | 0.671 | 0.498 | 0.232 | 0.667 | 0.172 |
| | UnIVAL (ours) | **0.690** | **0.515** | **0.237** | **0.713** | **0.178** |
| Clotho v1 | [5] | **0.590** | 0.350 | **0.220** | 0.280 | - |
| | [21] | 0.556 | 0.363 | 0.169 | 0.377 | **0.115** |
| | [8] | 0.551 | **0.369** | 0.165 | **0.380** | 0.111 |
| | UnIVAL (ours) | 0.569 | 0.367 | 0.178 | **0.380** | 0.114 |

**Table 3: Finetuning for audio-captioning.**

**Audio-Text tasks.** Even though we do not pretrain on audio-text data, we evaluate the generalization ability of our model to the new audio modality. We use an additional audio encoder pretrained on audio classification and finetune UnIVAL directly for Audio Captioning. Table 3 shows a comparison with other approaches that takes only the audio as input. Interestingly, we significantly outperform other approaches on Audiocaps, and we are competitive with current SoTA on the small Clotho v1 dataset.

## 4. Weight interpolation of UnIVAL models

We merge models by employing weight interpolation techniques as outlined in prior literature [13]. While previous studies focused on merging models trained for a specific task (classification) within the same modality (images or text), we aim to extend weight averaging to a more challenging scenario. Our main objective is to determine whether merging models trained on highly diverse multimodal tasks can outperform individual task-specific models. Our framework is well-suited for this investigation due to its unified architecture, which supports many tasks, and shared pretraining, which promotes linear mode connectivity across weights [6]. This direction offers multiple benefits, including potential improvements in out-of-distribution (OOD) settings [15] and the ability to recycle open-source finetunings of our multimodal model [2, 14].

We explore various approaches within this framework. (1) **Weight interpolation** where we linearly interpolate between the weights of 2 models. (2) **Fusing finetuning** [2] where the average of auxiliary models' weights serve as the initialization for the last finetuning on the target task and (3) **Ratatouille finetuning** [14] where each auxiliary model is finetuned independently on the target task, and then all the finetued weights are averaged.

### 4.1. Knowledge transfer via weight interpolation of models finetuned on multimodal tasks.

Here we validate the effectiveness of WA for multimodal image-text tasks (Image Captioning, VQA, and VG). For Fusing and Ratatouille finetuning, the other tasks, besides the target one are considered auxiliary.

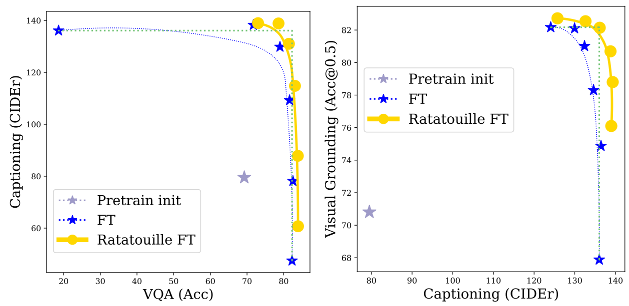


**Figure 2: Weight interpolation between models trained on different multimodal tasks.**

**Towards pareto-optimality.** The interpolation curves in Fig.2 show that we can effectively combine the skills of expert models finetuned on different tasks. While task-finetuned models perform very well on their specific target task, they suffer from severe performance degradation when evaluated on other tasks. This suggests that the different tasks are in tension. Fortunately, weight interpolation reveals convex fronts of solutions to efficiently trade-off between the different abilities. Actually, it is even possible to find an interpolating coefficient $\lambda$ so that the interpolated model outperforms the specialized one (*e.g.*, in Fig.2 the CIDEr score of the model obtained from $0.8 \times Cap + 0.2 \times VQA$ is 138.51 vs 136.52 for the Captioning model). We speculate this model benefits from the synergy between different tasks. Besides, the performances on transfer and OOD generalization are further improved in Ratatouille. Specifically, for OOD ($\lambda = 0$ or 1) Ratatouille reaches (Fig.2) 57.80/121.29 compared to 45.64/118.0 for vanilla FT on VQA to Captioning/VG to Captioning respectively.

## 5. Conclusion

In this study, we introduce UnIVAL, the first unified model capable of supporting image, video, and audio-text tasks. We achieve competitive performance while training a $\sim 0.25$B parameter model on relatively small dataset sizes. Our unified model paves the way to leverage model merging via weight interpolation, that can further exploit the diversity of these tasks. We aspire that our work will inspire the research community and accelerate the progress towards constructing modality-agnostic generalist assistant agents.

## References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. Fusing finetuned models for better pretraining. *arXiv preprint arXiv:2204.03044*, 2022.

[3] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*, 2021.

[4] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[5] Ayşegül Özkaya Eren and Mustafa Sert. Audio captioning based on combined audio and semantic embeddings. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 41–48, 2020.

[6] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *ICML*, 2020.

[7] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

[8] Andrew Koh, Xue Fuzhao, and Chng Eng Siong. Automated audio captioning using transfer learning and reconstruction latent space similarity regularization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7722–7726. IEEE, 2022.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[10] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022.

[11] Xubo Liu, Xinhao Mei, Qiushi Huang, Jianyuan Sun, Jinzheng Zhao, Haohe Liu, Mark D Plumbley, Volkan Kilic, and Wenwu Wang. Leveraging pre-trained bert for audio captioning. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1145–1149. IEEE, 2022.

[12] XINHAO MEI, XUBO LIU, QIUSHI HUANG, MARK DAVID PLUMBLEY, and WENWU WANG. Audio captioning transformer. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2021)*.

[13] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[14] Alexandre Rame, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model Ratatouille: Recycling diverse models for out-of-distribution generalization. In *ICML*, 2023.

[15] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. In *NeurIPS*, 2022.

[16] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022.

[17] Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. Efficient vision-language pretraining with visual concepts and hierarchical alignment. In *33rd British Machine Vision Conference (BMVC)*, 2022.

[18] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*, 2022.

[19] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

[20] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

[21] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Zeyu Xie, and Kai Yu. Investigating local and global information for automated audio captioning with transfer learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 905–909. IEEE, 2021.

[22] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *arXiv preprint arXiv:2111.12085*, 2021.

[23] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.