Unified Applications of Generative and Transformer-Based Models Across Modalities
(Group Haha)

Hoang Phuc Trong
FPT University
tronghpse183203@fpt.edu.vn

Le Quang That
FPT University
thatlq183256@fpt.edu.vn

Abstract

This project investigates the potential of generative and transformer-based models in solving complex problems across diverse modalities, including audio, video, biology, and multimodal vision-language tasks. By leveraging advanced architectures such as Demucs, MODNet, AlphaFold2, CLIP, and Flamingo on benchmark datasets (e.g., MUSDB18, YouTube-VOS, ProteinNet, and LAION-400M), we aim to evaluate and analyze their transferability, generalization capabilities, and limitations in real-world scenarios.

1 Introduction

Transformers and generative models have redefined deep learning by enabling robust performance in low-resource and cross-domain tasks. With self-attention mechanisms and large-scale pretraining, these models excel in applications ranging from natural language processing to computer vision and biological structure prediction. Our goal is to unify these strengths and explore their real-world impact across four challenging domains: music source separation, video matting, protein folding, and multimodal few-shot classification.

This project aims to bridge the gap between theoretical capabilities and practical deployment of these models. By investigating use cases from different fields—such as audio signal separation, object-level video segmentation, molecular biology, and multimodal few-shot classification—we aim to evaluate how well these models generalize, transfer across domains, and handle noisy or limited data scenarios.

2 Datasets and Tasks

We evaluate pre-trained or fine-tuned models using the following datasets:

- **MUSDB18** (1) for music separation
- **YouTube-VOS** (2) for video matting
- **ProteinNet** (3) for structure prediction
- **LAION-400M** (4) for vision-language matching

Each dataset presents unique challenges, requiring the modeling of audio waveforms, video segmentation masks, biological sequences, and multimodal semantic relationships.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

3 Related Work

Demucs (5) combines spectrogram and waveform-level separation. MODNet (6) delivers real-time matting performance. AlphaFold2 (7) utilizes Evoformer for structural biology. In the multimodal domain, CLIP (8) and Flamingo (9) demonstrate impressive zero-shot capabilities.

4 Proposed Methodology

Our experimental framework includes four major domains, each associated with a representative task, model, and dataset:

- Preprocessing for each modality (text, audio, video, and sequence)
- Evaluation using PyTorch, HuggingFace, and BioPython
- Visualization of attention maps and analysis of learned embeddings

5 Evaluation Strategy

Metrics used per domain:

- **SDR** for music separation
- **IoU** and alpha matte loss for video matting
- **GDTM** and RMSD for protein structure
- **Accuracy** and F1 for CLIP/Flamingo classification

We will also conduct cross-validation, reproducibility checks, and training curve analysis.

6 Expected Outcomes

We aim to evaluate how generative and transformer-based models perform across diverse modalities. The project will highlight each model's strengths and limitations, and identify which architectures generalize well in low-data or cross-domain settings. For example, Flamingo will be tested on few-shot image-text matching tasks from the LAION-400M dataset to assess its zero-shot reasoning ability. Overall, our findings will serve as a practical reference for selecting and adapting models in multi-domain applications.

7 Applications

Domain	Practical Use Case	Industry/Field
Music Separation	Karaoke apps, remixing tools	Entertainment, AudioTech
Video Matting	Background replacement, AR in video calls	Film, XR, Communication
Protein Structure	Drug discovery, mutation impact prediction	Biomedical, Pharma
Multimodal Learning	Visual search, content filtering, recommendation	Web, AI search engines

8 Conclusion

By integrating state-of-the-art generative and transformer-based models into a unified benchmarking framework, this project aspires to provide a comprehensive snapshot of how these models perform across radically different data types and tasks. The results will inform researchers, practitioners, and developers about model robustness, domain adaptability, and architectural trade-offs.

This work stands as a prototype for afuture AI systems that are not just narrow in scope but capable of general intelligence across modalities—a step closer to building truly versatile, multimodal machine learning systems.

References

[1] SigSep. *MUSDB18 - A Corpus for Music Separation*. 2017. https://sigsep.github.io/datasets/musdb.html

[2] Xu, N. et al. *YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark*. ECCV 2018. https://youtube-vos.org/

[3] AlQuraishi, M. *ProteinNet: A Standardized Data Set for Machine Learning of Protein Structure*. BMC Bioinformatics, 2019. https://www.biorxiv.org/content/10.1101/625467v4

[4] Schuhmann, C. et al. *LAION-400M: Open Dataset of CLIP-filtered 400 Million Image-Text Pairs*. 2021. https://laion.ai/blog/laion-400-open-dataset/

[5] Défossez, A. et al. *Hybrid Spectrogram and Waveform Source Separation*. arXiv preprint arXiv:2106.09685. https://arxiv.org/abs/2106.09685

[6] Ke, L. et al. *Green Screen Matting with Deep Learning*. arXiv preprint arXiv:2004.08372. https://arxiv.org/abs/2004.08372

[7] Jumper, J. et al. *Highly Accurate Protein Structure Prediction with AlphaFold*. Nature, 2021. https://www.nature.com/articles/s41586-021-03819-2

[8] Radford, A. et al. *Learning Transferable Visual Models From Natural Language Supervision*. ICML, 2021. https://arxiv.org/abs/2103.00020

[9] Alayrac, J.-B. et al. *Flamingo: A Visual Language Model for Few-Shot Learning*. arXiv:2204.14198. https://arxiv.org/abs/2204.14198