Generative Models and Transformer-Based Approaches on Multiple Modalities
(Group Haha)

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

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

Abstract

Recent advances in deep learning, particularly in generative modeling and transformer-based architectures, have led to breakthroughs across multiple data modalities. This project explores the application of generative and transformer-based models to a range of challenging tasks including music source separation, video object matting, protein structure prediction, and multimodal few-shot classification. We aim to understand the generalization capabilities of these models when adapted to different types of data—text, images, audio, and biological sequences—using publicly available datasets and pretrained models.

1 Introduction

Transformer-based models and generative architectures have reshaped AI research. Originally developed for NLP, these models now excel in vision, audio, biology, and multimodal tasks. Generative models like GANs and diffusion models achieve impressive synthesis results, while transformers offer generalization across domains. This project investigates their use in tasks such as music separation, video matting, protein folding, and multimodal classification.

2 Datasets and Tasks

We work on four tasks, each with dedicated datasets:

- Music Separation: MUSDB18 ?
- Video Matting: YouTube-VOS ?
- Protein Structure Generation: ProteinNet ?
- Multimodal Few-Shot Classification: LAION-400M ?

These tasks span multiple modalities with varied challenges such as noisy data, long sequences, and weak supervision.

3 Related Work

Recent models demonstrate domain-specific and cross-domain success:

- Demucs ?: audio source separation with convolutional transformers.

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

- MODNet ?: video matting using semantic priors and attention.
- AlphaFold2 ?: highly accurate protein structure prediction using transformers.
- CLIP and Flamingo ??: few-shot visual-language understanding.

These works motivate our exploration of model transferability across modalities.

4 Proposed Methodology

For each task, we will:

- Run or fine-tune existing pretrained models (e.g., Demucs, MODNet, OpenFold, CLIP).
- Compare training results with inference-only pipelines.
- Analyze latent representations and attention maps.

We use PyTorch, HuggingFace, and compatible data loaders. Evaluation is both quantitative and qualitative.

5 Evaluation Strategy

Evaluation metrics per task:

- Music: Signal-to-Distortion Ratio (SDR)
- Video: F-measure, Mean Absolute Difference
- Protein: RMSD, GDT
- Multimodal: Top-1 Accuracy, Zero-shot Precision

Learning curves and error visualization will help assess overfitting or transfer quality.

6 Expected Outcomes

We aim to demonstrate that generative and transformer-based models generalize across domains with minimal modification. We expect insights into architectural limits and the effect of pretraining, with practical takeaways for future multimodal applications.