Unified Applications of Generative and Transformer-Based Models Across Modalities

Huhu Team
Hoang Phuc Trong - SE183203 + Le Quang That - SE183256
AI1801, DAT301m, SU25
FPT University

Abstract

This project investigates the applicability and robustness of state-of-the-art generative and transformer-based models across various modalities including audio, video, and biological data. We evaluate three representative tasks: music source separation using Demucs, real-time video matting with MODNet, and protein structure prediction with AlphaFold2. Each model is tested on a domain-specific dataset to assess generalization, robustness, and computational trade-offs. We demonstrate how these architectur...

1 Introduction

Transformer-based and generative models have become foundational in modern AI systems due to their scalability and versatility. They achieve robust performance across diverse domains such as natural language processing, computer vision, and structural biology. However, applying these models to new data modalities remains a non-trivial challenge.

In this project, we evaluate three prominent models—Demucs (for audio), MODNet (for video), and AlphaFold2 (for biology)—on real-world tasks. Our objective is to establish a unified framework to assess the generalization and transferability of modern AI models, especially under realistic constraints like limited hardware resources in academic settings.

2 Related Work

Demucs is a U-Net-style model for music source separation that processes both waveform and spectrogram representations. It has shown strong performance in vocal and drum separation tasks while preserving audio quality.

MODNet is a real-time portrait matting model that eliminates the need for trimaps. It leverages a three-branch architecture to segment fine details such as hair, even under motion blur or cluttered backgrounds.

AlphaFold2 revolutionized protein structure prediction by integrating attention-based representations over both sequence and spatial features through its Evoformer block. It achieved state-of-the-art results in CASP14.

These models represent different extremes of modern AI—audio signal processing, real-time visual segmentation, and structural biology. Each introduces a unique architectural insight that we examine in this unified setting.

These models represent the state-of-the-art in their respective domains and offer different approaches to generalization and inference. Comparing them across modalities also allows

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

us to reflect on how deep learning abstractions adapt between signal, spatial, and structural domains.

3 Datasets and Tasks

MUSDB18 : A benchmark dataset for music source separation, containing 150 professionally produced songs with isolated stems (vocals, drums, bass, others). It covers diverse genres, enabling robust training and evaluation of music separation models.

YouTube-VOS : A video object segmentation dataset with high-quality spatio-temporal annotations. It includes varied scenarios such as object occlusions, fast motion, and background clutter—making it ideal for evaluating MODNet's real-time capabilities.

ProteinNet : Derived from PDB and CASP competitions, ProteinNet provides aligned sequences, structures, and evolutionary profiles for a wide range of proteins. It includes structured train/validation/test splits to ensure fair benchmarking.

Each dataset was preprocessed using domain-specific techniques. For instance, audio signals were normalized and downsampled; protein sequences were encoded using one-hot representation or language models; and video frames were extracted and resized to ensure consistency across GPU memory limits.

4 Methodology

We use pretrained models when available and retrain or fine-tune where feasible within computational limits. The methodology across the three tasks includes:

4.1 Music Source Separation with Demucs

We utilized a PyTorch implementation of Demucs with pretrained weights. Input waveforms were chunked and processed in batches to avoid GPU overflow. Spectral analysis was performed post-separation to calculate SDR (Signal-to-Distortion Ratio) metrics. The model outputs four tracks: vocals, drums, bass, and others. Evaluation was performed on a curated subset of 50 MUSDB18 samples.

4.2 Video Matting with MODNet

MODNet was applied frame-by-frame to 10-second clips from YouTube-VOS. Each frame was segmented into foreground and background using predicted alpha mattes. Evaluation was based on IoU (Intersection over Union) against ground-truth masks and perceptual loss to capture boundary quality. Real-time inference speed was also recorded.

4.3 Protein Folding with AlphaFold2

Using the ProteinNet dataset, we parsed input sequences and processed them via the Evoformer and structure modules of AlphaFold2. Due to hardware constraints, we only predicted a subset of protein targets. RMSD (Root Mean Square Deviation) and GDT-TS (Global Distance Test) were used to evaluate the predicted 3D structure against the ground-truth.

5 Results and Analysis

Across all models, performance aligns well with published benchmarks, validating the integrity of our pipeline. However, computation time was a limiting factor in model scalability and depth of evaluation.

Task	Model	Duration	Metric Result
Music Separation	Demucs	24h	SDR: vocals = 5.2 dB, drums = 4.7 dB
Video Matting	MODNet	12h	IoU: 86.1%, matte loss ≈ 0.031
Protein Folding	AlphaFold2	10h	RMSD: 3.1Å, GDT-TS: 0.634

Table 1: Experimental results across all three domains.

6 Discussion

6.1 Cross-Modal Insights

While the models were not designed for cross-domain applications, their behavior reveals patterns. For instance, both MODNet and Demucs rely on U-Net principles with skip connections. This shared structural pattern suggests that insights in image segmentation can inform audio separation and vice versa.

6.2 Limitations

The largest barrier was computational capacity. Training AlphaFold2 or Flamingo from scratch was not feasible. We relied on inference-only pipelines, which limited exploration of model robustness under fine-tuning.

6.3 Future Work

To overcome these challenges, we suggest exploring:

- Lightweight transformer variants (e.g., Linformer, Performer)
- Model compression techniques (e.g., quantization, pruning)
- Cross-modal transfer learning via shared encoder pretraining

7 Conclusion

We conducted a unified benchmarking of generative and transformer-based models across three highly distinct domains: audio, video, and protein folding. Each model demonstrated strong performance in its native modality and revealed opportunities for architecture reuse across tasks.

By identifying the limitations tied to computational resources and task-specific preprocessing, we pave the way for more generalized and accessible AI applications. The full codebase is available at: `https://github.com/thatlq1812/dat301m_project`.