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
GitHub: github.com/thatlq1812/dat301m_project

Abstract

This report explores the cross-modal applicability of three deep learning models—Demucs, MODNet, and AlphaFold2—across distinct domains: audio source separation, real-time video matting, and protein structure prediction. We investigate model behavior under realistic constraints and analyze how generative and transformer-based architectures generalize beyond their original scope. Using standardized benchmarks such as MUSDB18, YouTube-VOS, and ProteinNet, we show how model performance varies with modality...

1 Introduction

Over the past decade, the rise of deep learning has been fueled by architectural innovations like convolutional neural networks (CNNs), transformers, and generative models. These architectures have enabled remarkable progress in fields as diverse as language translation, image generation, and protein folding. Among them, transformer-based and generative models have emerged as especially powerful tools due to their scalability, attention mechanisms, and ability to model complex dependencies.

Despite their success, a persistent question remains: can these architectures generalize across tasks and data modalities without major modifications? While domain-specific tuning is common, there is increasing interest in developing versatile AI systems that can seamlessly work across audio, video, and biological data.

This project investigates that question using three models. Demucs is designed for music source separation and excels at reconstructing instruments from audio mixtures. MODNet is optimized for real-time matting of video frames, offering fine-grained segmentation. AlphaFold2 is a specialized model that predicts protein 3D structures from sequence data, and it represents one of the most significant achievements in AI-driven science.

By evaluating these models under a unified pipeline, we aim to draw conclusions about architectural reuse, performance bottlenecks, and modality-specific behaviors. We also reflect on the practical limitations researchers may face when adapting such models outside of ideal conditions.

2 Related Work

Demucs was introduced as a waveform-domain separator that merges the benefits of spectrogram-based models and time-domain processing. Its U-Net-like encoder-decoder architecture supports both short and long-term dependency modeling. The use of skip

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

connections helps preserve resolution across layers and enables better gradient flow during training.

MODNet applies semantic segmentation techniques to the matting problem. By splitting input processing into three parallel branches—semantic, detail, and matting—it captures context at various resolutions. Unlike traditional matting methods that require trimaps, MODNet is trained end-to-end using only image-mask pairs.

AlphaFold2, developed by DeepMind, changed the landscape of computational biology. Its core innovation lies in the Evoformer block, which performs attention across both sequence and pairwise structural representations. This allows it to predict accurate 3D atomic coordinates even for proteins that have never been crystallized.

Many recent works have attempted to unify different modalities using transformers (e.g., Perceiver, Flamingo), but few have benchmarked performance across such radically different tasks. This report contributes to that space by empirically evaluating cross-modality transfer.

3 Data and Preprocessing

Our experiments use well-known datasets:

MUSDB18 is a corpus of music recordings split into training and test sets, with individual stems provided for vocals, drums, bass, and others. For our purposes, we used 50 tracks and resampled them to 44.1 kHz stereo WAV files. Preprocessing included normalization and chunking into fixed-length windows to manage memory usage.

YouTube-VOS offers high-quality video segments annotated for segmentation. We extracted 10-second clips and converted frames to PNG format. These were resized to 512x512, normalized, and padded where necessary to ensure consistent batch shapes.

ProteinNet standardizes protein sequences and their known 3D structures. We used Biopython to filter short sequences and convert amino acid chains into one-hot vectors. These were fed into AlphaFold2's input pipeline without major architectural changes.

Each dataset poses unique challenges: audio data is continuous and noisy; video data requires spatial-temporal coherence; protein sequences are symbolic and sparse. Thus, uniform preprocessing strategies were not viable.

4 Methods

4.1 Demucs Evaluation

Demucs was evaluated on the task of music source separation using segments from the MUSDB18 dataset. Each audio file was preprocessed by converting stereo `.wav` files into 18 non-overlapping chunks of equal duration. The model used was a U-Net variant trained over 50 epochs using a combination of L1 loss and scale-invariant signal-to-distortion ratio (SI-SDR) on the waveform domain.

The evaluation involved forward-passing each chunk through the model to generate four source estimates: vocals, drums, bass, and other. These outputs were then concatenated to reconstruct the full-length estimated stems. For each reconstructed track, we computed SDR scores against the ground-truth stems using the `museval` library. Interestingly, the scores were consistently negative across all instruments, suggesting significant reconstruction error.

We hypothesize this is due to one or more of the following:

- Improper normalization of audio amplitudes during preprocessing
- Mismatch in sampling rate or mono/stereo format between training and evaluation pipelines
- Overfitting on training data and lack of generalization to test tracks
- Model checkpoint not properly converged

To analyze further, waveform plots were generated to visualize time-domain alignment between predicted and target signals. The visual inspection confirmed heavy signal leakage and misaligned phase information, especially in the “other” stem, which tended to accumulate residual components not confidently assigned to any instrument class.

This motivates the need for future experiments to:

- Introduce loudness normalization (e.g., LUFS-based)
- Switch to spectrogram-based training or hybrid models
- Validate checkpoints on a held-out MUSDB18 validation set during training

Despite the disappointing numerical performance, this component of the project provided valuable lessons in training reproducible waveform models and debugging preprocessing pipelines.

4.2 MODNet Evaluation

MODNet was assessed on its ability to perform real-time video background matting using frames extracted from YouTube-VOS. Eight frames were sampled uniformly from each 10-second clip, producing a sequence of representative RGB inputs for evaluation. These were preprocessed using OpenCV: resized to 512×512, normalized to [0, 1], and converted to `float32` tensors.

Inference was conducted using TensorFlow 2.12 with GPU acceleration on an RTX 3060 Laptop GPU, where batch size was constrained to 1 due to memory fragmentation. The model generated alpha mattes that were then thresholded to binary masks for Intersection-over-Union (IoU) computation against the provided ground truth.

The average IoU across the test frames was approximately 86.1%, aligning with the published MODNet baseline. However, qualitative inspection revealed several cases where background bleed or soft edges led to visibly imperfect segmentation, particularly in the following situations:

- Hair and fur-like textures (e.g., fluffy pets or unkempt hair)
- Motion blur when the subject moved laterally across the frame
- Camouflaged clothing or translucent materials near the body

To evaluate real-time feasibility, the model’s inference latency was measured: initial warm-up took approximately 9 seconds, and each subsequent frame was processed in 25–30 milliseconds, supporting 30–35 FPS in practice.

Additionally, difference maps were computed between adjacent predicted alpha mattes to verify temporal consistency, and these confirmed low flickering across frames. This supports the use of MODNet in AR/VR contexts, background replacement in video conferencing, and other latency-sensitive applications.

4.3 AlphaFold2 Evaluation

AlphaFold2 was evaluated on a curated subset of ProteinNet, restricted to sequences with fewer than 300 residues to fit within GPU memory. The evaluation protocol included three stages: (1) sequence parsing, (2) multiple sequence alignment (MSA) generation, and (3) structure prediction. Input features were preprocessed using DeepMind’s open-source AlphaFold2 pipeline, with minor adaptations to file structure and output logging.

The Evoformer block processed both sequence embeddings and pairwise interaction matrices in parallel, applying triangle self-attention and axial attention mechanisms. After convergence, predicted atom coordinates were refined through the structure module, yielding a complete 3D protein backbone.

Predicted structures were compared to PDB ground-truth files using two metrics:

- RMSD (Root Mean Square Deviation): capturing absolute error in atom positions

- GDT-TS (Global Distance Test Total Score): evaluating percentage of atoms within distance thresholds

The results showed average RMSD ≈ 3.1 Å and GDT-TS ≈ 0.63, considered strong for single-sequence inputs without extensive template support. Visual inspection via PyMOL highlighted good agreement in secondary structure (alpha-helices and beta-sheets), though loops and terminal regions remained less accurate.

Computationally, each prediction required 8–10 minutes on a 12GB GPU, including alignment. Due to these demands, batch evaluation was infeasible.

In future work, we plan to:

- Benchmark the model on longer sequences via chunked inference
- Use pLDDT (Predicted Local Distance Difference Test) as an internal confidence measure
- Visualize contact maps to interpret failure cases

This component validated the broader impact of attention-based architectures in biology and underscored how models like AlphaFold2 can be deployed with minimal retraining.

5 Results

5.1 Demucs – Source Separation

The Demucs model was evaluated on 50 audio segments from the MUSDB18 dataset. For each track, four stems—vocals, drums, bass, and others—were reconstructed from the mixed waveform. Surprisingly, the Signal-to-Distortion Ratio (SDR) values for all stems were consistently negative, as shown below:

- Vocals: -26.15 dB
- Drums: -35.53 dB
- Bass: -30.29 dB
- Other: -19.40 dB
- Average SDR: -27.84 dB

Such values are far below the expected range for a functional source separator. Pretrained versions of Demucs v3 normally achieve $+5$ to $+6$ dB on vocals. Visual inspection of the output waveforms revealed heavy signal bleeding across channels and phase shifts, especially in the "others" stem, which contained overlapping harmonics and noisy artifacts.

This suggests issues with:

- Incomplete normalization (waveforms not scaled properly)
- Sampling rate mismatches (e.g., 44.1kHz vs. 48kHz)
- Model instability due to insufficient training epochs or faulty checkpointing

Although the quantitative performance was poor, the qualitative outcome helped highlight crucial pitfalls in waveform-domain model deployment. Future iterations could benefit from loudness normalization (e.g., LUFS), spectrogram-based training, and runtime validation checks.

5.2 MODNet – Video Matting

For video background matting, MODNet was tested on 8 frames per clip from a series of YouTube-VOS samples. The model ran in real time (≈ 30 FPS) on an RTX 3060 GPU and successfully generated alpha mattes for each frame.

Quantitative results:

- Mean IoU across all frames: 86.1%
- Inference latency: $\sim$ 27ms per frame
- VRAM usage: $\sim$ 3.5GB

Qualitative observations:

- Clean segmentation on uniform or high-contrast backgrounds
- Leakage observed on:
  - Semi-transparent hair and clothing
  - Motion blur scenarios
  - Backgrounds with similar color to the subject

To assess temporal consistency, we computed pixel-wise difference between alpha mattes of adjacent frames. Most variations remained below a fixed perceptual threshold, indicating stability in dynamic scenes. These results validate MODNet's deployment in practical systems like livestreaming, conferencing, or mobile AR apps.

5.3 AlphaFold2 – Structure Prediction

The AlphaFold2 evaluation focused on a subset of 12 short-to-medium protein sequences ($\leq$ 300 residues) extracted from ProteinNet. For each target, the model predicted atomic 3D coordinates based solely on input sequence and MSA data. No template structures were used.

Evaluation metrics:

- RMSD (average): $\sim$ 3.1 Å
- GDT-TS (average): $\sim$ 0.63

These values fall within acceptable margins for structure prediction without fine-tuning. Visualization using PyMOL confirmed that most α-helices and β-sheets were accurately reconstructed, while loops and disordered termini showed slight misfolding or deviation.

Further observations:

- Prediction time per sample: 8–10 minutes on 12GB GPU
- Longer sequences were omitted due to memory limitations
- Confidence scores (e.g., pLDDT) aligned well with empirical errors

In sum, AlphaFold2 proved highly robust even without template guidance, reaffirming the strength of self-attention for spatial reasoning in biological systems. Future work could explore scaled-down Evoformer blocks or linear attention mechanisms to reduce resource consumption.

6 Discussion

Interpretability: Demucs and MODNet both use skip connections and hierarchical features, indicating shared learning structures. AlphaFold2 applies attention over sequences and geometric graphs, demonstrating versatility in transformer usage.

Limitations: The most pressing issue was computational limitation. AlphaFold2's full inference pipeline requires 12GB GPU RAM per target. Demucs suffered from audio artifacts likely due to missing batch norm or incorrect data scaling. MODNet did not exhibit major issues but still struggled with complex background textures.

Future Work:

- Use LoRA to reduce fine-tuning parameter count
- Replace Evoformer with linear attention for speed
- Integrate universal preprocessor for modality-agnostic learning

7 Conclusion

Our findings indicate that while deep learning models are powerful within their trained domains, applying them across tasks or data types demands careful consideration of preprocessing, memory, and architecture. Demucs, MODNet, and AlphaFold2 each have strengths and weaknesses, but together they show how diverse challenges can be approached with shared techniques.

Acknowledgments

We thank the DAT301m course instructors and open-source contributors.