Multimodal AI for Predicting the Success of Creative Content

Mansi Sunil Jadhav
240087524
Dr. Dimitrios Kollias
Dr. Odysseus Kaloidas
MSc Big Data Science
School of Electronic Engineering and
Computer Science, QMUL

***Abstract*—This dissertation investigates whether multimodal learning can improve the prediction of YouTube video success, operationalised as view count. We present an end-to-end framework that integrates visual content, language, and platform metadata via late fusion, and we evaluate the task in two complementary forms: continuous regression (forecasting expected views) and classification into discrete view-range bands. The study establishes strong unimodal baselines, then quantifies the added value of multimodal fusion and analyses modality contributions and failure modes. To ensure rigour, we adopt channel-disjoint and channel-mixed protocols. Findings highlight the dominance of platform context and motivate learnable cross-modal attention and end-to-end adaptation to unlock further gains. We conclude with practical implications for creators/platforms and a roadmap for future multimodal forecasting.**

Keywords—multimodal learning; transformers; VideoMAE; BERT; YouTube; popularity prediction; engagement forecasting.

I. Introduction

Online video platforms, such as YouTube, have transformed the way information and entertainment are produced, distributed, and consumed. Anticipating whether a video will “succeed” is both practically valuable and scientifically interesting for creators, studios and platforms. Yet, predicting success is challenging, as performance depends on a complex mix of content quality, presentation (title/thumbnail/metadata), timing, audience fit, and external dynamics. This dissertation investigates whether a multimodal learning approach—combining video, text, and metadata—can improve the prediction of YouTube success, quantified here as view count.

In this work, “success” is modelled in two complementary ways. First, as a regression problem that predicts the expected number of views. Second, as a classification problem that assigns each video to a view-range category. Tackling both allows us to (i) estimate continuous outcomes and (ii) estimate an interpretable view band used for planning. A key hypothesis is that late fusion—learning modality-specific representations and combining them at a higher decision layer—yields better generalisation than any unimodal model, because each modality captures distinct, complementary signals: visual semantics and style (video), topical and rhetorical cues (text), and platform-facing context (metadata).

Despite abundant interest in forecasting social media performance, several challenges remain. View distributions are heavy-tailed, with a few viral outliers and many low-exposure cases (Szabo, G. & Huberman, B.A., 2010); metadata can be noisy or strategically optimised; transcripts vary in quality and completeness; and visual content is high-dimensional and temporally structured. Moreover, exogenous factors (e.g., real-world events) are difficult to observe. These characteristics motivate careful target design (regression vs. class ranges), robust evaluation metrics, and architecture choices that respect modality structure.

1) Research Questions (RQs)

RQ1. To what extent can each modality—metadata, video, text—predict view counts on its own?

RQ2. Does a late-fusion multimodal model outperform the best unimodal baselines for both regression and classification?

RQ3. Which modality (or interactions) contributes most to predictive performance, and how stable are these contributions across data splits?

2) Approach Overview

We construct a dataset of 500 YouTube videos from the years 2023 and 2024 using platform APIs and automated pipelines. The metadata modality includes attributes such as video category, duration, publish time, and channel statistics like number of subscribers and number of views, likes and comments on previous videos. For text, we use automatic speech recognition (ASR) transcripts extracted using Whisper to capture topical content and narrative structure along with the title and description. For video, we extract spatiotemporal features from sampled frames/clips using the pre-trained VideoMAE model. Each modality is first modelled with a task-specific head (regression and classification). A weighted late-fusion then combines modality predictions to produce final predictions.

3) Contributions

- A practical multimodal pipeline for YouTube success prediction that integrates metadata, transcript text, and visual features with a late-fusion strategy applicable to both regression and classification.
- Systematic baselines and ablations, comparing unimodal vs. multimodal models, analysing modality importance, and quantifying gains across metrics (MSE/PC for regression; accuracy/F1/ROC-AUC for classification).
- Target design and evaluation under skew, including view-range classes and log view count, with discussion of trade-offs for interpretability and robustness.

II. Related Work

A. Background Research

Early studies on video popularity prediction often relied on metadata and early engagement metrics rather than content features. For example, factors such as uploader reputation, view counts, likes, and social network signals were used to train simple supervised models. (Szabo, G. & Huberman, B.A., 2010) While these approaches provided some predictions, they mostly reflected only initial user reactions and failed to capture the deeper contextual and content-based nuances of the videos. (Szabo, G. & Huberman, B.A., 2010) Overall, traditional models struggled with multimodal data, leaving intrinsic video qualities (visual, audio, textual content) underutilized in predicting long-term popularity.

Recent research has increasingly turned to multimodal deep learning techniques to address these shortcomings. Many works now incorporate heterogeneous features — combining visual content, textual data, audio, and other metadata — into unified models. (Chen, X. et al 2013)

B. *Literature Review*

In the context of YouTube videos, researchers have now adopted multimodal feature integration. Le *et al.* (2024) present a deep multimodal framework (dubbed *EnTube*) that integrates the video's title and textual metadata, audio track, thumbnail image, and the video content itself (frames) — to predict audience engagement on YouTube. (Le et al., 2024) Their model classifies videos into high, medium, or low engagement categories ("Engage", "Neutral", "Not Engage"), and experimental results showed that combining all these modalities yields higher accuracy than any single-modality model. (Le et al., 2024)

Another line of work has explored attention-based multimodal fusion for video popularity. Cho et al. (2024) developed an attention-enhanced BiLSTM model called AMPS (Attention-based Multi-modal Popularity prediction System) tailored for short-form videos (e.g. YouTube Shorts). (Cho M. et al, 2024) This led to a higher accuracy and F1-score over baseline classifiers, and ablation studies confirmed that each modality (visual, textual, etc.) contributed meaningfully to the prediction performance. (Cho M. et al, 2024) These findings illustrate that capturing cross-modal interactions is crucial for improving popularity forecasts.

Beyond custom-designed networks, researchers are now leveraging large pre-trained models for this task. Sun *et al.* (2025) recently investigated the use of large multimodal transformers for short video engagement prediction. (Sun et al, 2025) In their approach, a foundation model (VideoLLaMA2) was used to jointly process key video frames, textual metadata (titles/descriptions), and even background audio, enabling end-to-end learning of engagement predictors. (Sun et al, 2025) This large model outperformed prior state-of-the-art methods, especially when all three modalities were present, indicating that powerful pre-trained multimodal models can effectively capture the complex factors behind viewer engagement.

VideoMAE is one such model. In the paper by Tong et al, they show that video masked autoencoders (VideoMAE) are data-efficient learners for self-supervised video pre-training (SSVP). VideoMAE (Video Masked Autoencoder) is a transformer-based model for video understanding that extends the masked autoencoder (MAE) paradigm to video data. It can learn effectively from relatively small video datasets without requiring huge pre-training corpora. (Tong et al, 2022)

III. Methodology

A. *Data Collection*

We assembled a dataset of 500 YouTube videos using automated pipelines around the YouTube Data API (metadata/statistics) and scripted downloaders for assets. The collection window spans 2023–2024, with content drawn primarily from the US and UK. To reduce channel-specific leakage and over-representation, we limited sampling to ≤3 videos per channel across 17 categories and balanced channel sizes (small 38.8%, medium 36.6%, large 24.6%; ≈40–40–20). Videos cover a range of durations: short (<4 min) 39.2%, medium (4–20 min) 41.8%, long (>20 min) 19.0%.

Acquired modalities and fields:

- Video (mp4): Downloaded videos via yt-dlp (with fallbacks where required) for visual feature extraction.
- Text: Video title, description, and transcripts; transcripts were generated with a Whisper ASR model when creator captions were unavailable.
- Metadata: Standard identifiers, content descriptors, timing, engagement, and channel stats, including *video_id, video_url, region_code, video_category_id/title, upload_date/time, video_duration, like_count, comment_count, channel_id/title/url, subscriber_count, channel_size, video_count, previous_video_count, average number of_views on_previous 10 videos, average_likes_on previous 10 videos and average number of_comments_on previous 10 videos, view_count.*

We derived some columns from the above extracted data. We split the duration into bins to get short, medium and long videos and the number of subscribers to get small, medium and large channels.

B. *Dataset Split*

We evaluated models under two protocols with an overall 60/10/30 train/validation/test partition. In both cases, the test set was sealed first and never used for binning, preprocessing, or hyperparameter tuning. For classification, five view-range classes were defined via equal-frequency quantiles learned on the train data (see below); for regression, bins were used only for stratification—the target remained log(view_count+1). In both cases, we have tried to split in such a way that the classes are balanced across all three sets.

1) Person-Independent Protocol (PI) (Channel-Disjoint)

Goal: zero channel overlap across train/val/test.

1. Outer hold-out by channel (70/30): We applied GroupShuffleSplit with groups = channel_id to obtain a train set (70%) and a sealed test set (30%) with no channel overlap.

2. Class bins learned without test leakage: On the train set, we derived five quantile bins from view_count to define view_range. We use these same view ranges for validation and

test set too. This label encoding is denoted view_range_enc_ind.

3. Validation via stratified, group-aware split: From the train set, we used StratifiedGroupKFold with groups = channel_id and stratify = view_range. We set n_splits = 7 and reserved one fold as validation, producing ~60/10/30 overall while keeping channels disjoint.

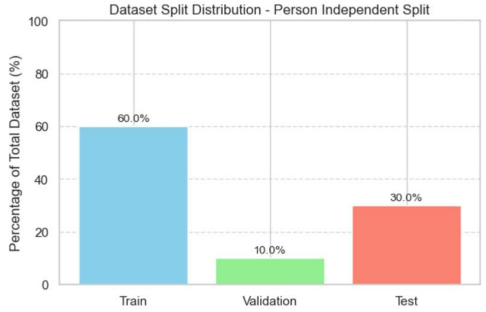

Fig. 1: Person-Independent Dataset Split Distribution

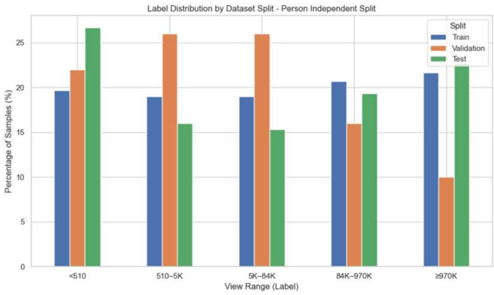

Fig. 2: Label Distribution - Person-Independent Dataset Split

2) Person-Dependent Protocol (PD) (Channel-Mixed)

Goal: Channels may appear in multiple splits; preserve class balance.

1. Outer random hold-out (70/30): We used train_test_split to create a train set (70%) and a sealed test set (30%) at the video level (no grouping).

2. Class bins from train data only: We computed five quantile bins on the train set to assign view range. These same view ranges were applied on the validation and test sets. The resulting encoding is view_range_enc_dep.

3. Validation via stratified split to hit 60/10/30: From the train set, we ran train_test_split with stratify = view_range and test_size ≈ 0.1429 (i.e., 1/7 of the dev pool), yielding an overall ~60% train / 10% val / 30% test.

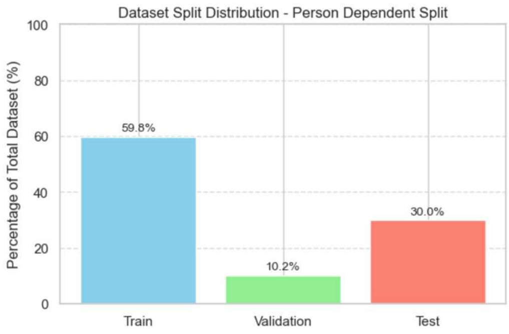

Fig. 3: Person-Dependent Dataset Split Distribution

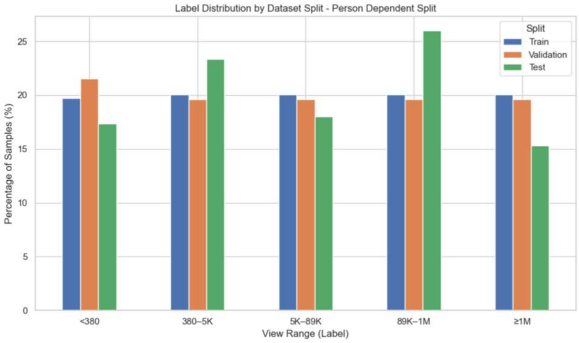

Fig. 4: Label Distribution - Person-Dependent Dataset Split

C. Data Preprocessing

1) Video Modality

a) Frame Sampling:

Each video is uniformly sub-sampled to 700 frames to cover the full timeline. If a source contains fewer than 700 frames, indices are repeated to reach a fixed length. Frames are exported as 256×256 JPEGs, yielding a fixed-length sequence while preserving global temporal structure.

b) Feature Extraction:

We use the pretrained VideoMAE-Base backbone (MCG-NJU/videomae-base) via the Hugging Face VideoMAEFeatureExtractor and VideoMAEModel. Because VideoMAE consumes 16-frame clips, frames are grouped into overlapping windows of 16 with overlap 2 (stride 14), producing ≈49 clips per video. Each clip is processed independently; we discard the [CLS] token and average patch tokens per frame to obtain one 768-D embedding per frame. Overlapping windows are merged into a single 700×768 matrix per video via weighted blending (we give lower weights at clip edges for smooth transitions). Features are cached as .npy files. Note that labels are not used during extraction.

2) Text Modality:

Inputs comprise of title, description, and transcript. When creator captions are unavailable, transcripts are produced with Whisper (small) ASR model. We form a single input string per video by concatenating the title, description, and transcript. Empty fields are skipped, so missing descriptions/transcripts simply yield shorter inputs. The text is tokenized with the BERT-base-uncased AutoTokenizer and padded/truncated to a fixed maximum length L =256. (Devlin et al, 2019) Padding tokens are masked out using the standard attention mask.

3) Metadata Modality:

We used tabular metadata comprising categorical variables—duration_type, channel_size, country, video_category_id—and numerical variables—subscriber_count, prev_video_count, avg_views_prev10, avg_likes_prev10, avg_comments_prev10, upload_hour.

Categorical features are one-hot encoded, and numerical features use log(1+x) followed by standardisation. The preprocessing transformer is fitted on the training data and applied unchanged to validation/test sets. Time-derived statistics (e.g., avg_*_prev10) are computed per channel using only prior uploads.

The correlation heatmaps in Fig. 5 and Fig. 6 show the correlation between view count and the metadata features. The strongest positive correlations are with the subscriber count and the average number of views and likes on the previous 10 videos, especially in case of the PD scenario. This pattern indicated that channel scale and short-term audience response are key predictors of subsequent video reach, showing that recent activity drives viewership.

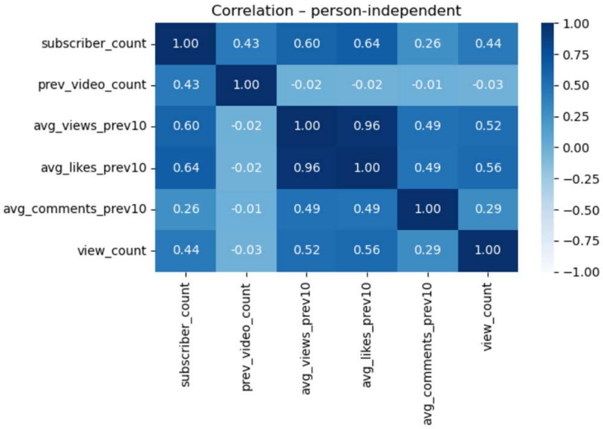

Fig. 5: Correlation Heatmap – Person-Independent

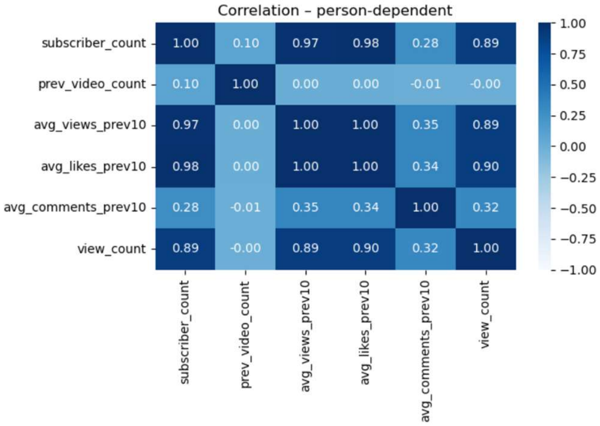

Fig. 6: Correlation Heatmap – Person-Dependent

D. *Model Architectures*

1) Baseline: CNN–RNN (ResNet-18 + GRU)

CNN-RNN pipelines are widely used as video baselines (e.g., CNN+LSTM/GRU). (Joe et al, 2015) Hence, to provide a content-only baseline, we extract per-frame features with a pretrained ResNet-18 (classifier head removed, global-average-pool features; 512-D per frame) and pass the 700-step sequence to a GRU. The final hidden state feeds a linear regressor to predict log view count. ResNet-18 is kept frozen; only the GRU and head are trained.

Shape: video → 700 frames → ResNet-18 → B,T=700,512 → GRU (hidden=256, layers=1) → Linear → B,1.

For classification, we use the same CNN–RNN pipeline: ResNet-18 frame features (512-D) → GRU (2 layers, hidden=136) → Linear(·, 5), trained with cross-entropy on logits under the same PI/PD splits.

2) Video Modality (Transformer Model)

Given per-frame features $X \in R^{T \times D}$ from VideoMAE (here T=700, D=768), we feed sequences directly to a lightweight Transformer encoder as shown in Fig. 7. Inputs shaped (B,T,D) are permuted to (T,B,D) to match PyTorch's nn.TransformerEncoder convention. A stack of L encoder blocks (parameter num_layers) performs multi-head self-attention and position-wise feed-forward transformations with residual connections and layer normalization. Global average pooling over time compresses the sequence to a single video embedding $z \in R^{D}$.

- Regression: Linear → $\hat{y} \in R$ predicting log-views.
- Classification: Linear → class logits in R^{C} with softmax.

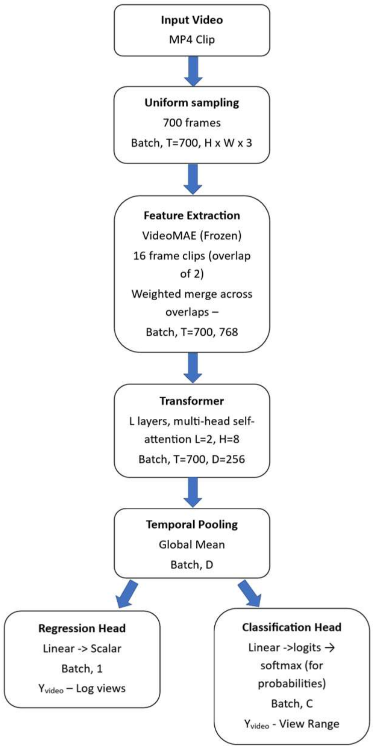

Fig. 7: Model Architecture – Video Modality

3) Text Modality (BERT)

We fine-tune BERT-base end-to-end (12 layers, hidden size H=768). From token embeddings, a small additive-attention pooler computes weights over tokens (padding masked), producing a pooled vector z. This pooling is more selective than [CLS] and more flexible than mean pooling, which helps when the useful signal is concentrated in a few sentences. It lets the model focus on the most informative parts of long transcripts while remaining simple and stable. (Zichao et al, 2016)

- Regression head: Dropout + Linear → ŷ (log-views).
- Classification head: Dropout + Linear → logits; softmax for class probabilities.

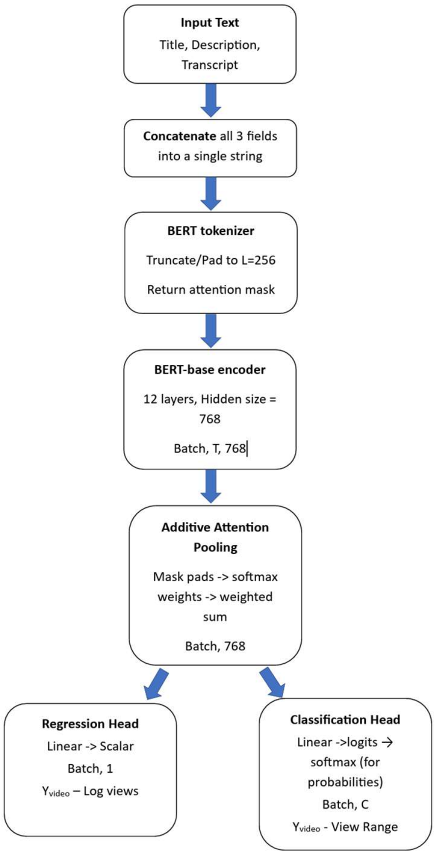

Fig. 8: Model Architecture – Text Modality

4) *Metadata (SVR/SVC)*

On pre-processed features, we train SVR (linear kernel) for regression and SVC (rbf kernel) for classification. Since we have tabular data, traditional models worked better than a deep learning training loop.

5) *Late Fusion*

We fuse per-modality predictions separately for regression and classification, and separately for the person-independent (PI) and person-dependent (PD) protocols. In both cases, weights are learned on validation only and then held fixed for test. This ensures that more accurate modalities contribute more.

- Regression (inverse-MSE weights): On validation, compute each modality's MSE on log-views; set weights w∝1/MSE. Weights are normalised if a modality is missing. The fused prediction is given as:
  $\hat{Y}_{fused} = a.\ \hat{y}_{video} + b.\ \hat{y}_{text} + c.\ \hat{y}_{metadata}$

- Classification (weighted vote / probability average): On validation, compute each modality's accuracy and set weights w∝Acc. For label decisions we use a weighted vote over hard predictions; for ROC we form probability fusion:
  $p_{fused} = w_v.p_{video}(c) + w_t.p_{text}(c) + w_m.p_{meta}(c)$

E. *Training and Optimisation*

We train two tasks (regression and classification) per modality. The shared protocol is: (i) Adam or AdamW with weight decay, (ii) early stopping on the validation metric (regression: MSE on log-views; classification: accuracy), (iii) fixed random seeds and deterministic settings where supported, and (iv) hyperparameters chosen with Optuna (TPE) on training/validation only. We use Huber loss for regression as it is more robust to outliers, and Cross Entropy Loss for classification. For text we use a linear warm-up and decay scheduler. The baseline is trained under the same PI/PD splits, losses, optimiser, early stopping, and seed controls as the proposed models (Huber loss on log-views; Adam/AdamW), ensuring a fair comparison.

In the case of Metadata, SVR/SVC hyperparameters are tuned on training/validation. For regression, we use a 'linear' kernel and 'rbf' for classification.

F. *Evaluation*

For both PI and PD protocols, we (1) train on the training split, (2) select hyperparameters and checkpoints on the validation split, (3) retrain the chosen configuration on train+validation, and (4) evaluate on the held-out test set.

Reported metrics are:

- Regression: MSE and Pearson correlation on log-views.
- Classification: accuracy, macro-F1, and ROC-AUC (micro and macro).

Fusion weights are computed on validation and reused unchanged on test (no leakage).

IV. Results

The results are reported in terms of individual performance metrics for each modality, along with the corresponding metrics for the multimodal fusion. Results are presented for both regression and classification tasks, under person-independent and person-dependent evaluation settings. A comparison with the baseline model is also presented.

A. *Regression*

As shown in Table I, metadata outperformed other single modalities in both PI/PD protocols. The MSE in PD is much lower than PI, confirming the stronger correlation between view counts and previous channel information as seen in the heatmap in section III. Text improves substantially from PI to PD indicating creator wording/recurring topics carry channel-level information. Video alone is less stable across splits and shows minimal difference in PI vs PD. In case of PD, the drop in PC for test suggests over-fitting to channel-specific patterns despite lower squared error.

Combining video and text data slightly improves the MSE over both splits. In PI split, video+metadata and text+metadata are close to the metadata MSE, however metadata is still slightly better. However, in PD, video+metadata and text+metadata perform better than metadata alone. Video+metadata gives best results among all. This shows that multimodal fusion in case of video and metadata proved to be effective.

The three-way late-fusion shows better results than video or text data alone but underperforms the best metadata-based models in both protocols. This indicates that static, validation-weighted late fusion is not sufficient to exploit the extra stream and may overfit noisier modalities.

The test MSE in terms of views is as follows:
Person-Dependent Split - 878271637213288.12
Person-Independent Split – 1176228581229917.75.

TABLE I: EXPERIMENT RESULTS: REGRESSION

Modality	Dataset Split	Val MSE – Log Views	Val PC	Test MSE	Test PC
Video	Person-Independent	13.3547	0.1905	14.8672	0.5365
Video	Person-Dependent	11.6350	0.6951	14.1122	0.3458
Text	Person-Independent	12.2181	0.3549	15.9045	0.4858
Text	Person-Dependent	10.5767	0.7003	11.9949	0.5356
Metadata	Person-Independent	5.9065	0.776	6.0593	0.844
Metadata	Person-Dependent	2.8223	0.931	5.6057	0.817
Video + Text	Person-Independent	11.8125	0.3665	13.7075	0.6138
Video + Text	Person-Dependent	7.1192	0.8411	10.8053	0.5514
Video + Metadata	Person-Independent	5.6414	0.7711	6.3186	0.8413
Video + Metadata	Person-Dependent	2.6112	0.9395	5.1965	0.8160
Text + Metadata	Person-Independent	5.4234	0.7791	6.4144	0.8393
Text + Metadata	Person-Dependent	3.1761	0.9284	5.4554	0.8094
All Modalities	Person-Independent	7.4539	0.7304	8.8002	0.8069
All Modalities	Person-Dependent	4.3591	0.9200	7.1284	0.7452

As per the scatter plots in Fig. 9 and Fig. 10, showing the true vs predicted values on the test dataset, it can be seen that across both splits, points tighten around the diagonal in the mid-range but fan out at the extremes. This shows the model under-predicts very high-view videos. The PD points are slightly tighter around the diagonal matching it's lower MSE as compared to PI.

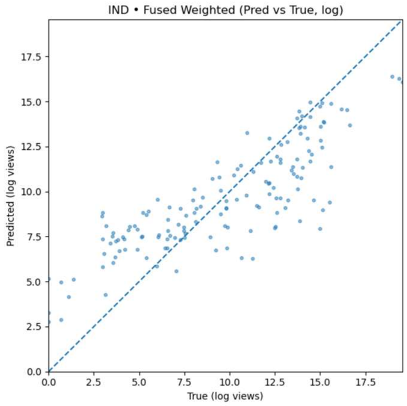

Fig. 9: True v/s Predicted Values Plot: Person-Independent Split (Test)

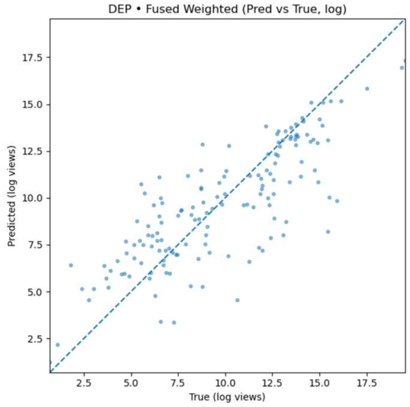

Fig. 10: True v/s Predicted Values Plot: Person-Dependent Split (Test)

B. *Classification*

Metadata is the strongest on the test set in both protocols. It can be seen in the graph in Fig. 11 and Fig. 12 that micro-ROC AUC for metadata is the highest as well. Text improves with PD suggesting recurring topics/phrasing transmit channel priors better than visuals alone. It's AUC shown in Fig. 11 and Fig. 12 follows the same trend, higher AUC for PD. Video is the weakest on test, consistent with the task being driven more by audience/context than by frame appearance.

Video+Text beats either stream alone. However, adding video or text to metadata does not improve test performance. All three modalities fused perform better than video or text alone, but not on metadata. This indicates that, although the modalities are complementary, the validation-derived weights did not surpass the already strong metadata signal.

It is worth noting that the accuracy on most modalities drops significantly from validation to test, especially in case of PD split. This shows that the models are overfitting on the channel data. This can be mitigated by training the models on a larger and more diverse dataset. Interestingly, in case of metadata, the accuracy increases on the test set. This could be because we're using one-fold validation, and the validation fold could be difficult to predict.

Additionally, in case of video and metadata, the accuracy on PI split is higher than PD split on test. This indicates the model is overfitting on the channel data. This is also confirmed by the significant drop in accuracy from validation to test (for PD). In both these cases, better regularisation could help to improve the accuracy on the test set.

TABLE II: EXPERIMENT RESULTS: CLASSIFICATION

Modality	Dataset Split	Val Acc	Val F1	Test Acc	Test F1
Video	Person-Independent	0.54	0.4381	0.3533	0.3058
Video	Person-Dependent	0.5686	0.5549	0.2933	0.2115
Text	Person-Independent	0.42	0.4212	0.3533	0.3401
Text	Person-Dependent	0.5686	0.5706	0.3800	0.3788
Metadata	Person-Independent	0.46	0.43	0.5933	0.5635
Metadata	Person-Dependent	0.76	0.76	0.5067	0.5053
Video + Text	Person-Independent	0.54	0.4381	0.32	0.317
Video + Text	Person-Dependent	0.5686	0.555	0.3867	0.3885
Video + Metadata	Person-Independent	0.54	0.4381	0.5333	0.4941
Video + Metadata	Person-Dependent	0.7647	0.7590	0.5	0.5015
Text + Metadata	Person-Independent	0.46	0.4332	0.38	0.3661
Text + Metadata	Person-Dependent	0.7647	0.7590	0.4867	0.4856
All Modalities	Person-Independent	0.48	0.3962	0.4267	0.4165
All Modalities	Person-Dependent	0.8431	0.8453	0.4800	0.4789

It can be seen in Fig 11. and Fig. 12, that the metadata (green) and fusion (red) curves rise more steeply near the origin, giving higher TPR at low FPR—useful for conservative gating (e.g., only flag likely "High/Very High" when false alarms must be low). Video/text curves climb more slowly, reflecting weaker separability. Fusion improves markedly over video/text but sits just below metadata across thresholds.

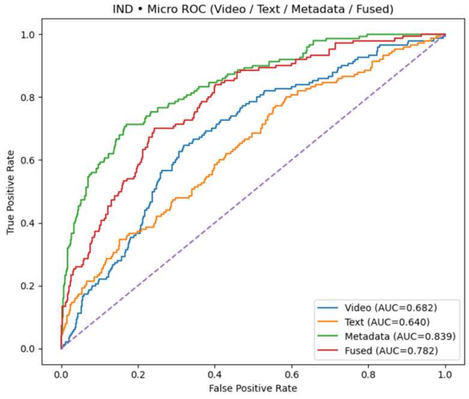

Fig. 11: Micro ROC Curve for all modalities: PI Split

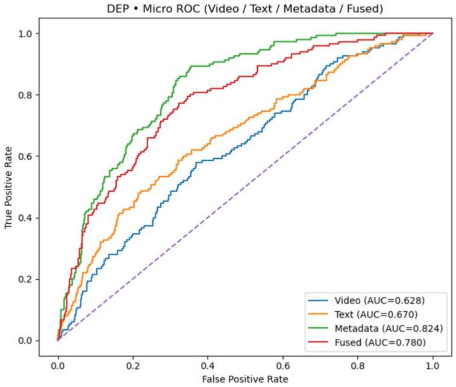

Fig. 12: Micro ROC Curve for all modalities: PD Split

It can be seen in the confusion matrix in Fig. 13 and Fig. 14, that in both PI and PD, the dominant errors are adjacent-bin confusions (e.g., Low vs Medium, High vs Very High), which is expected for an ordinal target. Very Low and Very High show the sharpest diagonals (large counts on the main diagonal), while mid-range bins diffuse into neighbours. PD shows a somewhat stronger diagonal for several bins (e.g., Low, High), but PI has crisper recognition of some extremes, mirroring the small PI>PD gap in AUC for fusion.

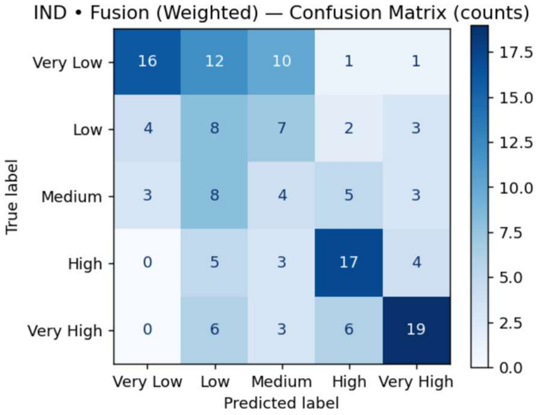

Fig. 13: Confusion Matrix – Fusion – PI Split (Test)

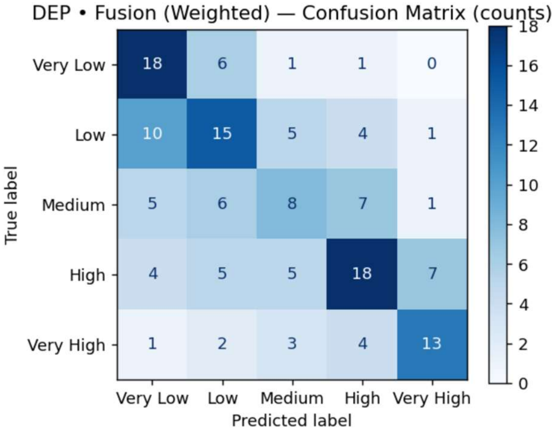

Fig. 14: Confusion Matrix – Fusion – PD Split (Test)

C. *Comparison with baseline model*

The CNN–RNN baseline consistently shows slightly lower results than our best single modality (metadata) and late fusion under both person-independent and person-dependent protocols. This gap is expected: (i) ResNet-18 provides static image features and lacks motion cues that VideoMAE captures; (ii) a GRU compresses long sequences into a single state and struggles with very long-range dependencies; and (iii) the baseline omits text and metadata, which our analysis shows carry a strong predictive signal. Consequently, while the baseline is useful for grounding results, the late-fusion model achieves better accuracy on the held-out test sets.

TABLE III: EXPERIMENT RESULTS: BASELINE MODEL

Dataset Split	Test MSE	Test Accuracy
Person-Independent Split	15.2536	0.42
Person-Dependent Split	14.9823	0.47

V. CONCLUSIONS

This work delivered a multimodal fusion pipeline for predicting YouTube video success in two complementary forms: regression on log view counts and classification into five view-range tiers. Each modality is modelled with an architecture suited to its signal—VideoMAE + Transformer for video, BERT-base + attention pooling for text, and SVR/SVC for metadata—and their outputs are combined with validation-derived weighted late fusion. Across both person-independent (PI) and person-dependent (PD) protocols, the weighted fusion consistently outperforms a CNN–RNN baseline (ResNet-18 + GRU), confirming that video-centric pretraining and multimodality add practical value over a conventional frame-CNN + RNN pipeline.

Among single modalities, metadata achieves the highest accuracy and correlation, reflecting the predictive strength of channel history and audience context. This is in line with previous research done. Prior work routinely finds that metadata/early-engagement dominates pure content models for view prediction. (Lu et al, 2025) Subscriber base and recent performance (avg_views_prev10, etc.) are highly predictive of the scale a new upload can reach. These priors generalise even to unseen channels (PI) because "larger/more active channels → more views" is a stable relationship. (Lu et al, 2025) Additionally, tabular stats are compact and well-behaved after log/standardisation. By contrast, text and video are high-dimensional, noisier (ASR errors, truncation), and their link to views is weaker and more context-dependent.

Although video and text alone did not yield optimal results, video combined with metadata outperformed all other combinations in regression. Additionally, text combined with metadata also gave close results. However, the combination of all three was not on par with even metadata alone. This shows that the fusion technique used for prediction is not as effective and a better fusion technique like using a MLP head could improve the results on all three modalities combined.

Comparing PI and PD, results show that having prior knowledge of a channel (PD) generally improves absolute error (lower MSE), while PI better reflects cross-channel generalisation and often preserves rank ordering (higher or comparable correlation/AUC). Together, the two protocols bound realistic deployment scenarios. Using both tasks—regression for fine-grained, continuous forecasting and classification for stakeholder-friendly tiering—gives a more complete picture of predictive power and decision support.

Overall, we present a robust, reproducible pipeline that can run either regression or classification with all three modalities, uses strict anti-leakage handling (bins and fusion weights learned on validation), and saves artefacts for repeatable evaluation. There is clear scope for improvement: expanding to a larger and more diverse dataset, end-to-end video fine-tuning when compute allows, calibrated or learned fusion (e.g., temperature scaling, stacking, ordinal-aware losses), stronger feature extraction and temporal augmentation, and cross-validated evaluation for tighter generalisation estimates.

VI. FUTURE WORK

- End-to-end training for video data: Instead of frozen backbones, fine-tuning a spatiotemporal encoder (e.g., VideoMAE) end-to-end may capture dataset-specific motion/style cues and give better results. However, this requires substantially more compute.
- Advanced multimodal fusion techniques: Explore early/mid fusion can prove to give comparatively better results. Instead of just weighted fusion, using a MLP head after the fusion layer can also improve results.
- Audio, Thumbnail, Title Formatting as additional modalities: Besides visual frames and text, audio can also prove to be essential in view prediction. Incorporating thumbnail images and title formatting features (length, casing, emojis, punctuation) can capture presentation effects that strongly influence engagement.
- ASR quality and text modelling: Replacing Whisper-small with stronger ASR (e.g., Whisper-medium/large-v3 or a domain-adapted wav2vec2/Conformer) to reduce WER and hallucinations could give better results on the text modality.
- Interpretability and Explainability: As models become more complex, it is crucial to provide explanations for their predictions – both for researchers and for end-users like content creators. Future work could incorporate model interpretability techniques specifically tailored to multimodal inputs. This might involve extending methods like Grad-CAM or attention visualization to highlight which parts of a video or which words in the title/description were most influential in the predicted engagement.
- Deployment in Recommendation systems: By integrating this content-based popularity predictor with a recommendation system, users can estimate

a video's likely success at upload time and use that in ranking algorithms. Moreover, using multimodal content signals in recommendations can complement user-centric signals.

- Creator-Facing Dashboards and Feedback Tools: Dashboards for creators could display predictions and insights about a video's expected performance before or shortly after publication.

References

[1] Szabo, G. & Huberman, B.A. (2010). *Predicting the popularity of online content*. Communications of the ACM, 53(8), pp. 80–88.

[2] Chen, X., Chen, J., Ma, L., Yao, J., Liu, W., Luo, J. & Zhang, T. (2018). *Fine-grained video attractiveness prediction using multimodal deep learning on a large real-world dataset*. In Proceedings of the Web Conference 2018 (WWW '18), pp. 671–678.

[3] Le, T., Nguyen-Thi, M.-V., Le, M.-T., Nguyen-Thi, H.-V., Le, T. & Nguyen, H.T. (2024). *EnTube: Exploring key video features for advancing YouTube engagement*. SSRN preprint 4868570 (June 2024).

[4] Cho, M., Jeong, D. & Park, E. (2024). *AMPS: Predicting popularity of short-form videos using multi-modal attention mechanisms in social media marketing environments*. Journal of Retailing and Consumer Services, 78, Article 103778.

[5] Sun, W., Cao, L., Cao, Y., Zhang, W., Wen, W., Zhang, K., Chen, Z., Lu, F., Min, X. & Zhai, G. (2025). *Engagement Prediction of Short Videos with Large Multimodal Models*. arXiv preprint arXiv:2508.02516.

[6] Tong, Z., Song, Y., Wang, J. & Wang, L. (2022). *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

[7] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 4171–4186.

[8] *Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici*; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4694-4702

[9] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

[10] Lu, J. *et al.* (2025) *Multi-modal and metadata capture model for micro video popularity prediction*. https://arxiv.org/abs/2502.17038.

[11] Pratik Kayal, Pascal Mettes, Nima Dehmamy, and Minsu Park. 2025. Large Language Models Are Natural Video Popularity Predictors. In Findings of the Association for Computational Linguistics: ACL 2025, pages 11432–11464, Vienna, Austria. Association for Computational Linguistics.

[12] Bielski, A. & Trzcinski, T. (2018). *Understanding multimodal popularity prediction of social media videos with self-attention*. IEEE Access, 6, pp. 74277–74287.

[13] Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Baird, A., *et al.* (2024). *The 6th Affective Behavior Analysis in-the-wild (ABAW) Competition*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2024).

[14] Kollias, D., Zafeiriou, S., Kotsia, I., Dhall, A., Ghosh, S., Shao, C. & Hu, G. (2024). 7th ABAW Competition: Multi-Task Learning and Compound Expression Recognition. *In* Computer Vision – ECCV 2024 Workshops. Springer.

[15] Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S. & Kotsia, I. (2023). ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges. *arXiv preprint* arXiv:2303.01498.

[16] Kollias, D. (2023). ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation (5th ABAW). *In* CVPR 2023 Workshops (CVPRW). Where to cite: Related Work—evidence of sustained progress in multimodal, multi-task video understanding; pointers to open benchmarks and baselines.

[17] Kollias, D. (2022). ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Multi-Task Learning Challenges (3rd ABAW). *In* CVPR 2022 Workshops (CVPRW).

[18] Kollias, D. (2022). ABAW: Learning from Synthetic Data & Multi-Task Learning Challenges (4th ABAW). *In* ECCV 2022 Workshops (LNCS). (also arXiv:2207.01138).

[19] Kollias, D., Kotsia, I., Hajiyev, E. & Zafeiriou, S. (2021). Analysing Affective Behavior in the Second ABAW2 Competition (ICCV 2021 Workshop). *In* ICCV 2021 Workshops (ABAW).

[20] Kollias, D., Schulc, A., Hajiyev, E. & Zafeiriou, S. (2020). Analysing Affective Behavior in the First ABAW 2020 Competition. *In* FG 2020 (IEEE).

[21] Kollias, D. & Zafeiriou, S. (2019). Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv preprint* arXiv:1910.04855.

[22] Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G. & Zafeiriou, S. (2019). Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *arXiv preprint* arXiv:1804.10938.

[23] Kollias, D. & Zafeiriou, S. (2020). Exploiting Multi-CNN Features in CNN-RNN based Dimensional Emotion Recognition on the OMG in-the-wild Dataset. *University of Greenwich / preprint repository*.

[24] Kollias, D. (2020). *Affect Recognition & Generation in-the-wild* (PhD thesis). BMVA Thesis Series.

[25] Amin, D.D., Moscholios, S., Papadopoulou, E., Yang, X., Kaloidas, O. & Kollias, D. (2025). Multi-Modal AI for Predicting the Success of Creative Content. ResearchGate preprint.

[26] Kollias, D. & Zafeiriou, S. (2018). *Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition*. arXiv preprint arXiv:1811.07770.

[27] Kollias, D. & Zafeiriou, S. (2018). A Multi-Task Learning & Generation Framework: Valence-Arousal, Action Units & Primary Expressions. arXiv preprint arXiv:1811.07771.

[28] Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. & Darrell, T. (2015). *Long-Term Recurrent Convolutional Networks for Visual Recognition and Description*. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2625–2634.

[29] Huber, P.J. (1964). *Robust Estimation of a Location Parameter*. Annals of Mathematical Statistics, 35(1), pp. 73–101.

[30] Wan, D., Wang, H., Stengel-Eskin, E., Cho, J. & Bansal, M. (2025). *CLAMR: Contextualized Late-Interaction for Multimodal Content Retrieval*. arXiv preprint arXiv:2506.06144.