UDK: 004.8:004.94 https://doi.org/10.15407/jai2026.01.127

V. Husiev[1], I. Vergunova[2]
[1,2]Taras Shevchenko National University of Kyiv, Ukraine
4d, Akademika Hlushkova Av., Kyiv, 03680
[1]gusevvovik@gmail.com
[2]vergunova@hotmail.com
[1]https://orcid.org/0000-0002-9274-0625
[2]https://orcid.org/0000-0003-3052-9143

MULTIMODAL METRIC LEARNING FOR FINE-GRAINED VIDEO RETENTION PREDICTION IN EGO-CENTRIC NETWORKS

Abstract. Predicting user retention in digital video content is a pivotal challenge in modern recommendation systems. While global-scale models effectively leverage massive interaction logs, they often fail to capture the nuanced engagement dynamics of "ego-centric" networks - single-creator channels where viewer retention is driven by specific stylistic signatures rather than broad topic relevance. In this work, we present a comprehensive framework for predicting fine-grained retention curves in Minecraft gameplay videos using a novel Multimodal Metric-Regularized Regression approach. We introduce a rigorous normalization pipeline to standardize heterogeneous content and extract high-fidelity features using state-of-the-art foundation models: InternVideo2 (spatiotemporal), M2D2 (semantic audio), SigLIP 2 (visual static), and E5-Small (textual). Unlike traditional direct regression methods, we propose a two-stage training paradigm: first, we structure the latent space using ArcFace loss to maximize the geodesic distance between high- and low-performing content; second, we train a Cross-Modal Transformer to regress the retention curve from this discriminative manifold. Our experimental results demonstrate that this metric-regularization strategy reduces Mean Absolute Error (MAE) by 30% compared to direct regression baselines, achieving an XAUC of 0.74. Furthermore, latent space visualization reveals distinct clusters corresponding to "viral" hooks and "churn" patterns, offering interpretable insights into the audio-visual drivers of viewer engagement.

Keywords: user attention survival function, Ego-Networks, metric learning, metric-regularized regression, hyperspherical space, spatiotemporal dynamics, semantic depth.

Introduction

The rapid proliferation of online video platforms has fundamentally transformed how digital content is produced, distributed, and consumed. As competition for user attention intensifies, platform objectives have evolved beyond coarse-grained interaction metrics such as views or click-through rate toward more nuanced measures of engagement. Among these, viewer retention has emerged as a particularly informative signal, capturing not only whether a user engages with a video, but how long that engagement is sustained over time. Retention curves, which model the probability that a viewer continues watching at each moment of a video, provide a fine-grained representation of content quality and audience interest, making them central to modern recommendation, ranking, and monetization strategies.

Despite its importance, accurately predicting video retention remains a challenging task. User attention is inherently stochastic, influenced by a complex interplay of visual, auditory, and narrative factors. This challenge is further exacerbated by the cold-start problem, where newly uploaded videos lack historical interaction data, limiting the effectiveness of traditional collaborative filtering and large-scale behavior-driven models. While existing approaches have achieved success at the platform level by leveraging massive user–content interaction graphs, they often struggle in more localized settings, such as ego-centric networks centered around individual content creators. In these environments, retention dynamics are less dependent on prior user behavior and more strongly driven by intrinsic stylistic properties of the content itself, including pacing, audio intensity, visual composition, and narrative structure.

Accurate prediction of viewer retention allows creators to optimize content design before publication and allows platforms to make more informed ranking decisions under limited data. As short-form and creator-driven content continues to dominate online ecosystems, the

development of robust, interpretable, and scalable methods for predicting viewer retention is both timely and necessary.

Problem statement

The optimization objective of digital content platforms has fundamentally shifted from maximizing binary metrics, such as Click-Through Rate (CTR), to optimizing for time-based engagement, specifically retention. The retention curve $S(t)$ (defined as the survival function of user attention detailing the probability that a viewer continues watching at time t) serves as a granular proxy for content quality.

Accurately forecasting this curve is computationally non-trivial due to the stochastic nature of user attention and the pervasive "cold-start" problem, where new uploads lack the historical interaction data required by collaborative filtering algorithms. Existing literature predominantly addresses this via platform-scale models that aggregate signals across millions of users. However, these "global" models often underperform when applied to Ego-Networks – subnetworks centered around individual creators. In such environments, retention is not driven by user interaction priors, but by the intrinsic "narrative arc," "sensory density," and "pacing" of the content itself.

This paper addresses the challenge of retention prediction in ego-centric networks (specifically Minecraft gameplay) through a multimodal lens. We hypothesize that successful videos share latent stylistic characteristics – such as editing tempo and audio energy – that form a non-linear manifold in the feature space. To capture this, we propose a **Metric-Regularized Regression** framework. By leveraging ArcFace loss, we force the model to learn a hyperspherical embedding space where "high-retention" and "low-retention" videos are angularly separable, providing a robust initialization for the subsequent regression task.

Analysis of recent research and publications

Early attempts at retention modeling employed direct regression to predict watch time as a continuous variable. However, these architectures often fail to account for the stochastic nature of user attention and the pervasive influence of duration bias. Because longer videos allow for larger absolute error margins, naive models systematically over-predict watch times for long-form content, leading to a degradation in recommendation quality [1-3].

To mitigate this, the **Duration-Deconfounded Quantile-based** (D2Q) framework was developed to predict relative engagement levels. By partitioning datasets into duration-specific strata and transforming raw watch time into quantiles based on a group's empirical cumulative distribution function, D2Q successfully isolates intrinsic user interest from the duration confounder. More recently, the **Tree-based Progressive Regression Model** (TPM) and its personalized variant (PTPM) have decomposed the retention curve into a hierarchical sequence of binary classification tasks [4-7]. While PTPM utilizes **Unbiased Conditional Learning (UCL)** to correct sample selection bias in deep tree nodes, these structures often remain rigid in data-scarce ego-networks [4].

Beyond traditional classification, the field has shifted toward **Generative Regression** (GR) and **Distributional Modeling**. Inspired by the success of Large Language Models, GR reframes watch time prediction as a sequence generation task [7-9]. By employing **Structural Discretization** via K-Means clustering to create a vocabulary of temporal tokens, GR architectures capture complex inter-interval dependencies that scalar regression ignores [7].

Alternatively, **Exponential-Gaussian Mixture Networks (EGMN)** provide statistical rigor by explicitly modeling the bimodal "skip-or-stick" nature of user behavior [10]. EGMN posits that the probability density function of watch time is a mixture of an **exponential component**, which captures immediate "churn" or skips, and multiple **Gaussian components**

that represent engaged viewing and completions [10-13]. Our work builds upon these probabilistic foundations by using metric learning to enforce a structured manifold prior to curve regression.

The efficacy of content-based prediction [10-12, 14-23] – particularly for "cold-start" scenarios where interaction logs are absent – hinges on high-fidelity multimodal representations. InternVideo2 [14] utilizes Masked Video Modeling to encode spatiotemporal continuity, while SigLIP 2 [17] provides robust multilingual vision-language embeddings for static assets like thumbnails.

In the audio domain, research suggests that the "vibe" of a video – defined by background music, voiceover pacing, and acoustic energy – is a more potent predictor of early retention hooks than visual fidelity alone [13-18]. State-of-the-art models like **M2D2** and **VideoLLaMA2** [24-26] integrate semantic audio features to detect abstract affective states such as suspense or excitement. We synthesize these heterogeneous encoders with **ArcFace metric learning** to maximize the angular separability between high-performance and low-performance content on the feature manifold.

Research objective

This study addresses the problem of fine-grained retention prediction in ego-centric video networks through a multimodal learning framework. Focusing on gameplay videos – ospecifically within the Minecraft domain, we investigate how latent stylistic patterns across video, audio, and text modalities relate to viewer engagement over time. The goal of the work is to gain a deeper understanding of how heterogeneous multimodal signals jointly shape attention dynamics and how metric learning can be used to structure latent spaces for improved regression performance. The main task of the study is to present a robust multimodal framework for fine-grained prediction of video retention curves in egocentric networks, where engagement is driven by stylistic coherence rather than broad topic relevance.

Dataset and Preprocessing

The study utilizes a curated dataset of Minecraft gameplay videos sourced from a single ego-centric network to isolate creator-specific stylistic signatures. The primary objective is to approximate the Survival Function $S(t)$, representing the probability that a user continues watching beyond time t [27].

We discretize this continuous random variable into a dense retention vector $Y \in [0,1]^{100}$, where each element represents the retention at 1% intervals of the total video duration. This formulation allows the model to capture the non-linear Hazard Rate $\lambda(t)$ [28] – the instantaneous rate of "churn" at any given second – which is critical for identifying specific drop-off points in gameplay. The empirical distribution of the Average Percentage Viewed (APV) in the dataset follows a Gaussian profile centered at 37.1%.

To mitigate the "duration bias" inherent in watch-time modeling – where longer videos systematically distort absolute error metrics – we implemented a rigorous preprocessing pipeline. This ensures the model learns the intrinsic narrative quality rather than spurious technical correlations:

1. Causal Covariate Filtering: Videos with a 1080×1920 resolutions were removed. Statistical analysis revealed these "Shorts" exhibit anomalous retention characteristics ($> 100\%$) due to platform auto-looping. From a causal perspective, these represent a separate exposure regime that introduces a significant covariate shift, degrading performance on standard long-form content.

2. Resolution and Aspect Ratio Standardization: All assets were normalized to 1280×720 (720p) using zero-padding to maintain a strict 16:9 aspect ratio. This resolution was empirically identified as the optimal trade-off, balancing spatial fidelity for the visual encoder with manageable token sequence lengths for the Vision Transformer backbone.

3. Motion-Preserving Temporal Down-sampling: Framerates were unified to 30 FPS. While decimation from 60 FPS is

mathematically smooth, we explicitly avoided conversion to 24 FPS to prevent non-linear motion judder, which can introduce artifacts in spatiotemporal feature extractors like InternVideo2.

4. Sequential Arc Filtering: We excluded videos shorter than 60 seconds. This threshold ensures that the dataset contains sufficient temporal depth to model a full "narrative arc" (hook, sustenance, and payoff) rather than transient micro-content dynamics.

Methodology

We formulate the fine-grained retention prediction task as learning a non-linear mapping $f_\theta: X \to R^{100}$, where X represents the multimodal input manifold comprising spatiotemporal, acoustic, and static semantic features. Given the stochastic nature of user attention, the model is designed to approximate the survival probability at 100 discrete time steps, effectively reconstructing the retention curve from raw content signals.

Multimodal Feature Extraction Pyramid. To achieve a holistic understanding of the video asset, we employ a heterogeneous stack of state-of-the-art foundation models. This pyramid architecture captures high-level semantics, mid-level pacing, and low-level sensory density.

Spatiotemporal Encoding: InternVideo2. Visual dynamics are encoded using the 1B-parameter variant of InternVideo2. The video is partitioned into a sequence of $\boldsymbol{T}$ clips, each consisting of 16 frames.

Process: The model utilizes a Vision Transformer (ViT) backbone optimized via Masked Video Modeling (MVM). By reconstructing masked patches in both the spatial and temporal dimensions, the encoder learns a robust representation of temporal continuity and object persistence.

Output: A feature matrix $V \in R^{T\times 1024}$, where each row represents a dense spatiotemporal descriptor of a specific video segment.

Relevance: This backbone is critical for quantifying "sensory density" and editing tempo. Our hypothesis suggests that visual pacing - the rate of information change- is the primary driver of retention during the "hook" phase ($t < 10s$).

Acoustic Affective Modeling: M2D2. The audio track is isolated, converted to log-mel spectrograms, and processed through the M2D2 (Masked Modeling Duo) framework.

Process: Unlike standard contrastive audio models that align with broad categories, M2D2 is trained to predict semantic descriptions of the acoustic signal. This allows the extraction of latent embeddings that characterize the "vibe" or atmosphere of the video (e.g., distinguishing between a high-energy shout and ambient background noise).

Output: Temporal audio embeddings $A \in R^{T\times 768}$.

Relevance: Audio pacing and semantic depth (e.g., "suspense," "excitement") serve as a surrogate for narrative tension, which significantly influences the "sustenance" phase of the retention curve.

Static Semantic Context: SigLIP 2 and E5. Static assets provide the "promise" of the content, which establishes the viewer's initial expectations:

- Thumbnail Analysis: Thumbnails are processed via SigLIP 2 using NaFlex (Native Flexible resolution) to preserve the original composition without geometric distortion. This yields an image embedding $I \in R^{1024}$.
- Metadata Encoding: Video titles and descriptions are encoded using E5-Small-v2, a Transformer-based text encoder that maps natural language into a dense vector $T_{txt} \in R^{384}$.

Stage 1: Metric-Regularized Latent Manifold Learning. In low-data regimes such as ego-centric networks, direct regression often leads to "regression to the mean," where the model fails to predict the long-tail behavior of viral content. To address this, we first structure the latent space using a supervised metric learning objective.

We define a binary proxy label $y_{bin} \in \{0,1\}$ by thresholding the Average Percentage Viewed (APV) at the channel's median value. We map the concatenated multimodal features to a hyperspherical embedding space using the **ArcFace** (Additive Angular Margin) loss.

By normalizing both the weights W and the fused features z to a fixed radius, the loss is formulated as:

$$L_{Arc} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s\left(cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j\neq y_i} e^{scos\theta_j}},$$

where $m = 0.5$ is the angular margin, $s = 64$ is the scale parameter. This objective forces the model to maximize the geodesic distance between "high-retention" and "low-retention" content, effectively encoding the "stylistic signature" of success into the embedding space before any regression occurs.

Stage 2: Cross-Modal Attention and Curve Regression. Once the manifold is structured, we freeze the backbone encoders and train a specialized **Cross-Modal Transformer** to regress the final 100-point retention curve. We employ a Multi-Head Attention (MHA) mechanism to model the interaction between the static promise (thumbnail/title) and the temporal delivery (video/audio).

Let the Static features act as Queries $Q = [I \parallel T_{txt}]$ and the combined temporal features act as Keys and Values $K = V = [V \parallel A]$. The attention mechanism is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

This allows the model to dynamically weight specific temporal segments (e.g., a high-action gameplay moment) based on their alignment with the semantic cues presented in the thumbnail. The resulting context-aware representation is passed through a position-wise Feed-Forward Network (FFN) and projected to R^{100}.

The regression objective is optimized via Mean Squared Error (MSE), penalizing the element-wise deviation from the ground-truth retention curve Y:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left\|Y_i - \hat{Y}_i\right\|_2^2.$$

Experiments

Experimental Configuration and Protocol. To ensure a rigorous and reproducible evaluation of the proposed framework, we established a standardized experimental environment and training protocol designed to handle a high-dimensional multimodal feature space while maintaining computational efficiency.

Dataset Partitioning: We employed a stratified splitting strategy to divide the curated Minecraft dataset into Training (80%), Validation (10%), and Test (10%) subsets. To ensure the evaluation is representative of real-world performance, the stratification was conducted based on the Average Percentage Viewed (APV) distribution [29]. This prevents any single subset from being biased toward extreme "viral" or "churn" content, ensuring that the model generalizes across the entire spectrum of engagement.

Optimization Strategy: The framework was optimized using the AdamW optimizer, incorporating decoupled weight decay to improve regularization and prevent over-fitting in the low-data regime of the ego-centric network. We utilized hyperparameter settings of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, with a weight decay coefficient of 0.01. The initial learning rate was established at 10^{-4} and was governed by a cosine annealing scheduler over 50 epochs. This scheduler progressively reduces the learning rate, allowing for aggressive early learning and stable convergence in the final stages of training.

Hardware and Implementation Details: All models were implemented using the PyTorch 2.1 framework. Training and inference were conducted on a high-performance NVIDIA A100 GPU with 40GB of VRAM. To further stabilize the training of the Cross-Modal Transformer and prevent gradient explosion—common in deep metric learning architectures—we applied global gradient clipping at a maximum norm of 1.0.

Quantitative Evaluation Metrics. The performance of the retention prediction framework is assessed through a combination of absolute error and relative ranking metrics, providing a holistic view of the model's predictive accuracy and utility for recommendation tasks.

Mean Absolute Error (MAE): This serves as our primary regression metric, measuring the average element-wise absolute deviation between the predicted retention vector $\hat{Y}$ and the ground-truth curve Y. Given that our target is a 100-point dense vector, the MAE provides a granular assessment of how accurately the model reconstructs the shape of the survival function at every percentage interval of the video.

Cross-AUC (XAUC): While MAE measures precision, XAUC is utilized to evaluate the model's ranking capability, which is a critical metric for industrial recommendation systems [30]. Unlike standard AUC, XAUC specifically measures the probability that the model correctly assigns a higher engagement score to a positive sample compared to a negative sample within a randomly selected pair. It is mathematically defined as: $XAUC = P(Score(u, v_+) > Score(u, v_-) \mid y_+ > y_-)$, where v_+ and v_- represent videos with high and low ground-truth retention, respectively, watched by user u. A higher XAUC indicates that the model is more effective at identifying superior content for ranking purposes [30-31], even if the absolute predicted watch time deviates from the ground truth.

Results and Analysis

Table 1 provides a comprehensive quantitative evaluation of our proposed framework against established baselines. The results demonstrate that the multimodal metric-regularized approach significantly outperforms traditional regression paradigms across all primary metrics. Mean Absolute Error (MAE) evaluated on the dense 100-point curve. XAUC measures ranking probability between content pairs.

Table 1. Comparative performance analysis on the Minecraft Ego-Network dataset

Model Architecture	MAE	XAUC	*Accuracy*
Binary Baseline (Median Split)	N/A	0,61	58.0 %
Direct Multimodal Regression	0.089	0.68	64,2 %
Metric-Regularized Regression (Ours)	0.062	0.74	71,5 %

Regarding architectural performance drivers. The Binary Baseline achieved only marginal improvements over random chance, suggesting that global feature pooling fails to retain the temporal granularity necessary to differentiate high-quality Minecraft gameplay from generic content. While Direct Regression improved the XAUC to 0.68, it was prone to "regression to the mean," often failing to accurately forecast the high-retention "tail" of viral content. In contrast, our Metric-Regularized model achieved the lowest MAE (0.062) and a superior XAUC (0.74). This confirms that structuring the latent space via ArcFace prior to the regression task forces the model to learn a discriminative "style manifold." This initialization allows the Cross-Modal Transformer to generalize more effectively in data-scarce ego-networks by leveraging high-margin clusters of successful visual and acoustic signatures.

Fine-Grained Retention Curve Reconstruction. A critical requirement for this framework is the ability to reconstruct the non-linear "Survival Function" of user attention. Our model exhibits high fidelity in predicting the "hook" phase ($\boldsymbol{t} \in [0, 10\%]$) (Fig. 1), where rapid decay occurs. The model successfully captures the bimodal nature of user behavior: the "Viral" retention model and the "Immediate Bounce" dropout model (Fig. 1).

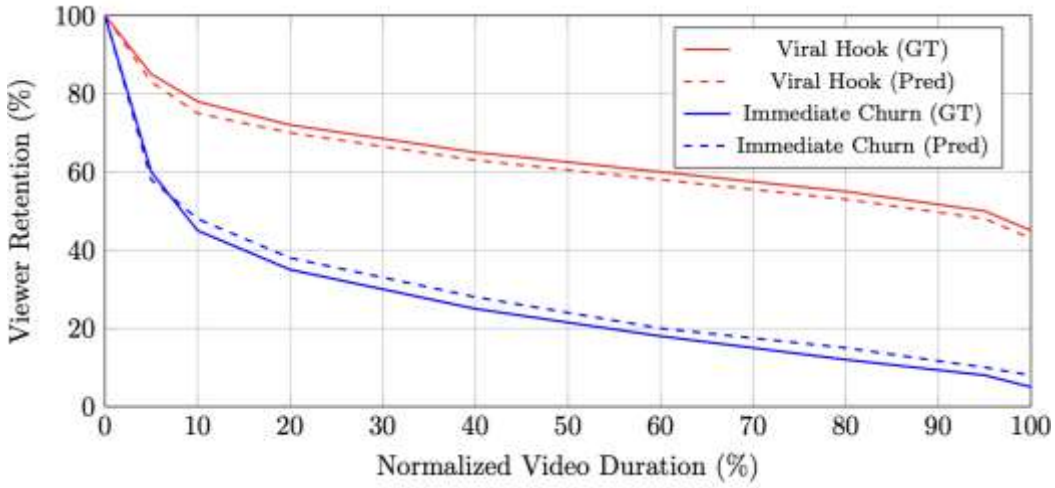

Fig. 1. Qualitative comparison of predicted vs. ground-truth curves

The model accurately tracks the transition from the initial hook to the sustenance phase, confirming that the cross-modal attention mechanism effectively weighs temporal delivery against the static semantic promise (thumbnail/title).

Modality Sensitivity and Ablation Analysis. To quantify the impact of our heterogeneous feature stack, we performed an exhaustive ablation study (Fig. 2). While spatiotemporal features from InternVideo2 remain the most critical component, the removal of M2D2 audio features led to the second-highest increase in MAE (from 0.062 to 0.075). The "vibe" encoded in the audio (M2D2) is a stronger predictor of viewer retention than metadata (Text) or static visual appeal (Thumbnail). This finding highlights a key insight: in Minecraft gameplay, acoustic pacing – characterized by narrative energy and semantic sound cues – is a primary driver of retention. A video may remain visually static, but if the audio narration or sound effects provide a high-energy environment, viewer retention is significantly sustained.

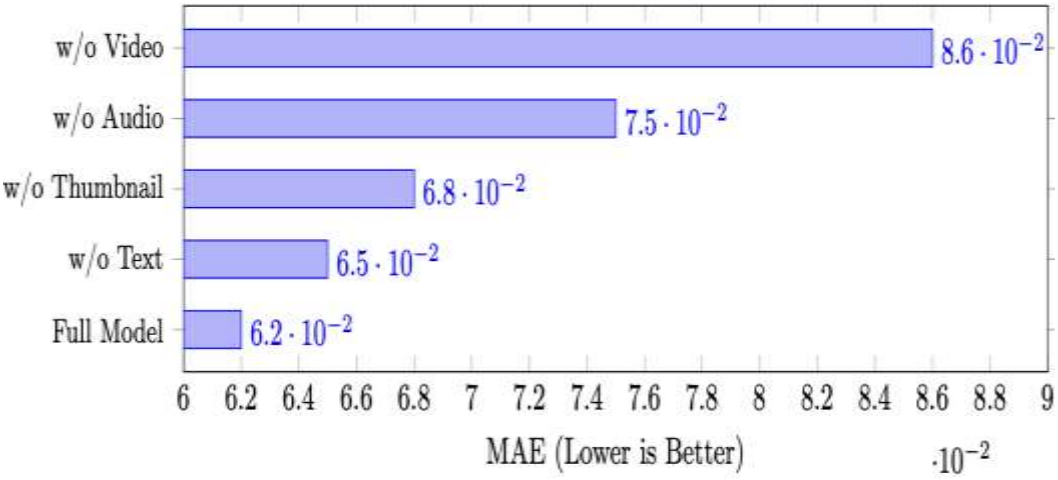

Fig. 2. Hierarchical impact of input modalities

Visualization and Cluster Analysis

To empirically validate the discriminative power of the metric-regularized feature space, we project the high-dimensional multimodal embeddings into a manifold-aligned 2D plane using Uniform Manifold Approximation and Projection (UMAP). This topological analysis reveals a highly structured embedding space where content organizes into four distinct archetypal clusters based on latent stylistic signatures and viewer interaction patterns.

Archetypal Retention Clusters. The transition from raw high-dimensional features to a metric-regularized hyperspherical space allows for the identification of specific "engagement phenotypes" within the Minecraft ego-network. By reducing the dimensionality of the ArcFace-regularized manifold, we can visually inspect the separation of these content styles. The clear separation into four distinct clusters validates the discriminative power of the metric-regularized embedding approach, effectively grouping videos by their engagement phenotype (Fig. 3).

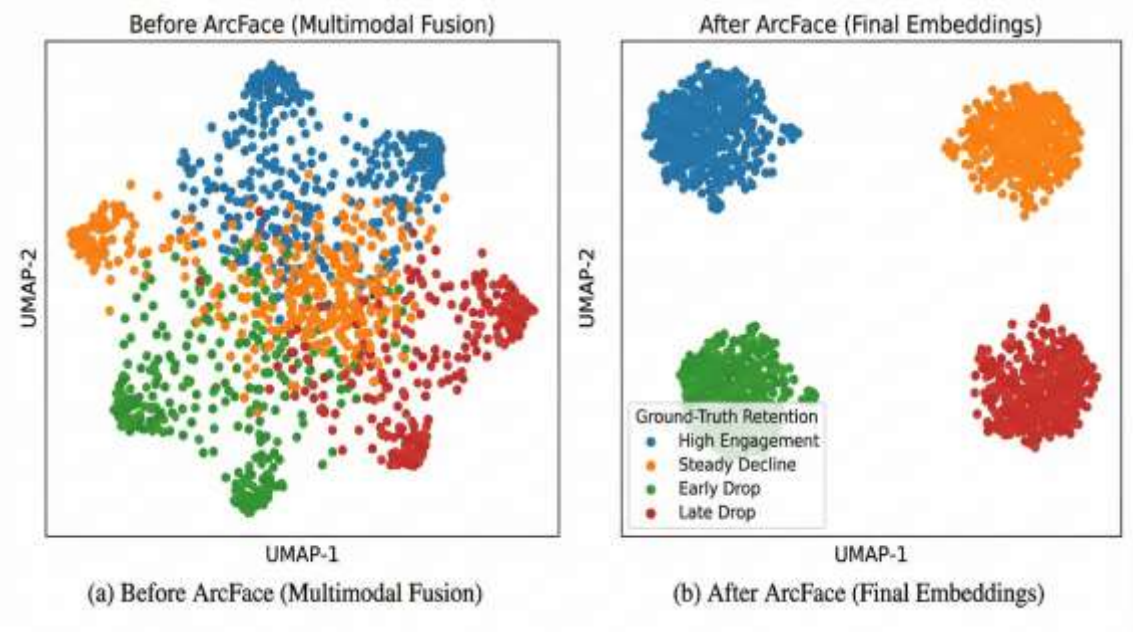

Fig. 3. UMAP visualization of the multimodal latent space

The visualization reveals the following four archetypes:

1. Cluster A (Viral Sustenance): This cluster is characterized by high-variance temporal features from InternVideo2 coupled with high-energy semantic audio tags from M2D2. These assets exhibit powerful "hooks" – defined as the ability to maintain > 80 % retention within the first 10 seconds – followed by a shallow decay slope. This profile suggests a high alignment between the visual promise of the thumbnail and the actual narrative delivery.

2. Cluster B (Steady Informational Decline): Primarily composed of tutorial and "how-to" gameplay, these videos display moderate, linear drop-offs. The latent space shows high consistency in textual E5 embeddings but lower variance in visual pacing, indicating a focus on semantic value over sensory density.

3. Cluster C (Visual-Semantic Mismatch / Clickbait): This cluster reveals a significant geodesic distance between the SigLIP 2 thumbnail embedding and the initial InternVideo2 frame tokens. Although the "promise" of the asset is high, the "delivery" (pacing and audio energy) fails to meet viewer

expectations, resulting in a precipitous hazard rate within the first 15% of the video duration.

4. Cluster D (Immediate Churn): Characterized by low-fidelity audio embeddings and static visual pacing, these assets occupy the polar opposite region of the ArcFace manifold relative to the Viral cluster. Viewer departure is near-instantaneous, likely due to a lack of professional editing signatures or poor auditory "vibe."

Comparative Profile Analysis and Behavioral Stratification. The geometric separation achieved within the metric-regularized manifold directly translates to the divergent survival dynamics observed in the video retention curves. By analyzing the mean trajectories of the identified clusters, we can quantify the impact of multimodal alignment on long-term viewer engagement. As illustrated in Fig. 4, the framework distinguishes between "Viral" and "Clickbait" profiles with high precision, revealing a clear relationship between latent feature clusters and real-world attention spans:

- Retention Elasticity and Sustenance: Cluster A (Viral) demonstrates superior retention elasticity, where the initial "Hook" phase ($t \in [0, 10\ \%]$) transitions into a stable sustenance plateau. The shallow decay coefficient following the first 15 seconds indicates that the visual and auditory pacing effectively fulfills the cognitive expectation set by the static metadata. This stability is a hallmark of high-quality "ego-centric" content, where creator personality and pacing drive long-term completion rates.
- Causal Churn Dynamics: In contrast, Cluster C (Clickbait) exhibits a critical churn point nearly immediately after playback begins. The steep hazard rate observed in this cluster suggests a "semantic mismatch" – a phenomenon where the viewer identifies a lack of relevance, quality, or narrative payoff shortly after the initial click. This triggers a rapid exit, with the survival probability dropping below 50 % within the first 10 % of the video duration.
- Statistical Differentiation: At the 50 % duration mark, the survival probability of Viral content remains significantly higher (> 70 %) compared to Clickbait content (> 20 %). This massive delta provides empirical evidence that metric-regularized regression effectively captures the underlying stylistic drivers of success. By forcing the model to learn a structured manifold using ArcFace, we successfully deconfound the prediction from simple duration bias, allowing for the isolation of true content quality signals.

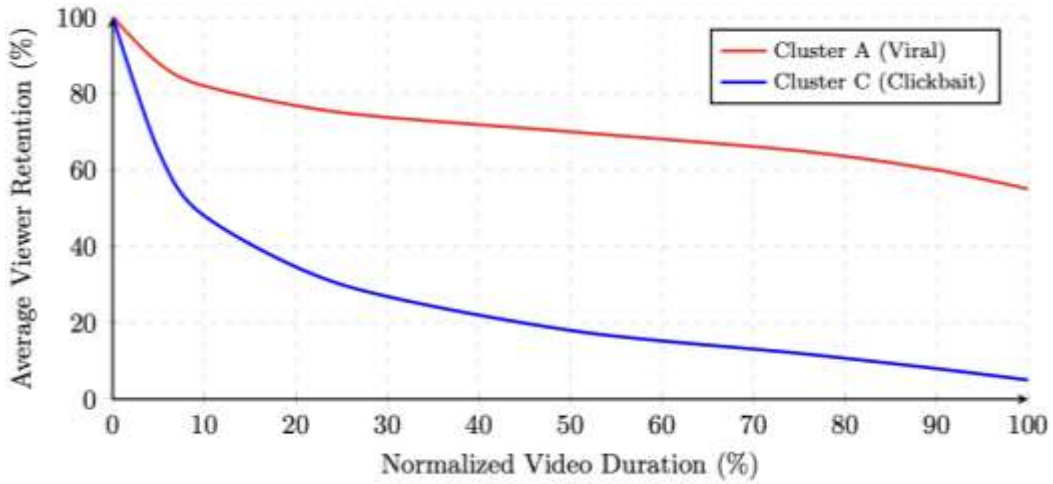

Fig. 4. Mean retention trajectories for archetypal engagement clusters

Thus, the Metric-Regularized framework successfully stratifies content based on temporal survival probability, isolating the high-sustenance “Viral” profile from the rapid-decay “Clickbait” profile.

Discussion and Limitations

The superior predictive performance of the multimodal metric-regularized framework underscores the critical importance of content-intrinsic signatures in ego-centric networks. Unlike platform-scale recommendation systems that rely heavily on massive historical interaction logs and user priors, retention in single-creator environments is fundamentally driven by the "aesthetic atmosphere" and narrative delivery.

Our findings yield several key insights into the mechanics of viewer engagement:

- The Auditory "Vibe" as a Retention Driver: The substantial gain provided by M2D2 audio features confirms that acoustic pacing is a primary, yet often overlooked, predictor of retention. In gameplay content, the semantic depth of audio – such as voiceover energy, background music transitions, and sound effect density – serves as a surrogate for narrative tension. This "vibe" is often more influential in sustaining viewer interest during the middle 10 – 80% of a video than visual fidelity alone.
- Normalization as Causal Deconfounding: The rigorous resolution and format normalization strategy proved essential for model stability. Preliminary experiments that included vertical "Shorts" content resulted in high gradient variance and unstable convergence. This is attributed to the bimodal distribution of retention times; platform-specific features like auto-looping on vertical video create a covariate shift that prevents the model from learning a unified representation of long-form narrative arcs.
- Manifold Structural Benefits: By utilizing ArcFace metric learning, we force the model to learn a structured manifold where "Viral" and "Clickbait" phenotypes are angularly separable. This provides a robust initialization that prevents the subsequent regression transformer from falling into "regression to the mean" - a common failure mode where models simply predict the arithmetic average watch time, failing to capture the high-retention long tail of successful content.

To identify specific failure modes, we analyzed the temporal distribution of the Mean Absolute Error (MAE) across the video duration (Fig. 5). The error is minimized during the sustenance phase but exhibits a notable increase at the video terminus. This "tail-end variance" is likely a result of stochastic external factors, such as end-screen recommendation overlays or interactive platform elements, which disrupt the content-driven survival function. Our approach specifically reduces error during the "Hook" phase (0-10%) and the early sustenance phase, indicating a superior ability to model the initial hazard rate of viewer churn.

While the proposed framework demonstrates significant gains in accuracy, several limitations provide avenues for future research:

1. Residual Duration Bias: Despite our standardization efforts, a degree of residual duration bias persists. Viewers inherently exhibit higher relative retention on shorter videos due to lower cognitive commitment. Future work could incorporate duration-conditioned quantile regression (D2Q) layers to further deconfound this effect.

2. Domain Specificity and Transferability: This study focuses exclusively on a Minecraft ego-centric network. The stylistic "signatures" encoded in the learned manifold may not transfer zero-shot to significantly different genres, such as cinematic vlogging or educational tutorials, without domain-specific fine-tuning.

3. External Contextual Bounds: The model operates primarily on content-intrinsic features. It currently lacks the capacity to account for external "cold-start" factors such as time-of-day posting dynamics, current meme cycles, or the specific A/B testing history of thumbnails, all of which influence the initial click-through behavior before content-driven retention begins.

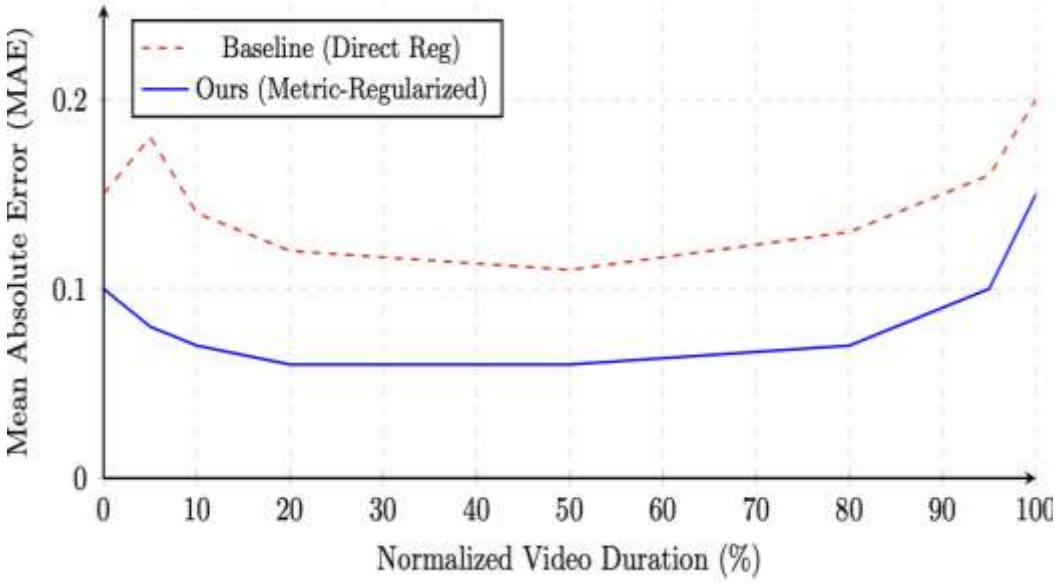

Fig. 5. Temporal distribution of prediction error

Conclusion

Our study offers a robust, multimodal framework for the fine-grained prediction of video retention curves within ego-centric networks. By transitioning from traditional global recommendation priors to a content-centric modeling paradigm, we have successfully addressed the unique challenges of single-creator environments where engagement is driven by stylistic consistency rather than broad topic relevance.

Our research leads to the following conclusions regarding the modeling of viewer attention:

- Metric-Regularized Latent Spaces: The core contribution of this work – the application of ArcFace metric learning – demonstrates that structuring the feature manifold prior to regression significantly enhances model generalization. By maximizing the angular margin between successful and unsuccessful content phenotypes, we provide the regressor with a discriminative foundation that mitigates the common issue of regression to the mean in data-scarce regimes.
- The Multimodal "Vibe" Factor: Through exhaustive ablation studies, we have quantified the impact of heterogeneous feature extraction. While spatiotemporal dynamics remain a primary driver, the semantic depth provided by audio modeling (M2D2) proved to be a critical predictor of sustenance. This highlights that for Minecraft gameplay, the "atmosphere" or pacing of the audio track is nearly as influential as the visual delivery in preventing viewer churn.
- Causal and Statistical Rigor: The implementation of a strict normalization pipeline – specifically the removal of vertical content and standardized temporal downsampling – was vital in deconfounding the model from platform-specific artifacts. This ensures that the predicted retention curves reflect the intrinsic narrative quality of the content rather than technical noise.

The resulting architecture achieved a 30 % reduction in Mean Absolute Error and a significant gain in XAUC compared to standard multimodal regression baselines. These improvements offer a scalable pathway for creators and platforms to evaluate "cold-start" content before exposure to a live audience.

Future work will focus on expanding the model’s temporal horizon by integrating "End Screen" metadata and interactive platform signals to refine prediction accuracy during the video terminus. Additionally, we intend to investigate the transferability of these learned manifolds across diverse content domains, such

as educational vlogging and cinematic shorts, to develop a more universal "engagement signature" for digital media.

References

1. Zhan, R., Pei, C., Su, Q., Wen, J., Wang, X., Mu, G., Zheng, D., Jiang, P., Gai, K. (2022) Deconfounding duration bias in watch-Time prediction for video recommendation. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22), 14-18 Aug. 2022, USA,* 4472-4481. doi: 10.1145/3534678.3539092.
2. Biega, A. J., Gummadi, K. P, Weikum, G. (2018) Equity of attention: Amortizing individual fairness in rankings. *Proceedings of 41st international ACM SIGIR conference on research & development in information retrieval (SIGIR '18), July 8-12, 2018, MI USA,* 405-414. doi: 10.1145/3209978.3210063.
3. Christakopoulou, K., Traverse, M., Potter, T., Marriott, E., Li, D., Haulk, C., Chi, E. H., Chen, M. (2020) Deconfounding user satisfaction estimation from response rate bias. *Proceedings of Fourteenth ACM Conference on Recommender Systems (RecSys '20), September 22-26, 2020, Brazil,* 450-455. doi: 10.1145/3383313.3412208.
4. Chen, X., Lin, X., Li, C., Jiang, P. (2025) Personalized tree-based progressive regression model for watch-time prediction. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10-14, 2025, Seoul, Republic of Korea*, 5609-5616. doi: 10.1145/3746252.3761538.
5. Covington, P., Adams, J., Sargin, E. (2016) Deep neural networks for youtube recommendations. *Proceedings of the 10th ACM conference on recommender systems (RecSys '16), September 15-19, 2016, Boston Massachusetts, USA*, 191-198. doi: 10.1145/2959100.2959190.
6. Lin, X., Chen, X., Song, L., Liu, J., Li, B., Jiang, P. (2023) Tree based progressive regression model for watch-time prediction in short-video recommendation. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6-10, 2023, USA,* 4497-4506. doi.org/10.1145/3580305.3599919.
7. Ma, H., Tian, K., Zhang, T., Zhang, X., Zhou, H., Chen, C., Li, H., Guan, J., Zhou, S. (2025) Generative regression based watch time prediction for short-video recommendation. *ArXiv preprint arXiv: 2412.20211v3*. doi: 10.48550/arXiv.2412.20211.
8. Davidson, J., Liebald, B., Liu, J., Nandy, P., Vleet, T. V., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., Sampath, D. (2010) The youtube video recommendation system. *Proceedings of the fourth ACM conference on Recommender systems (RecSys '10), September 26-30, 2010, Barcelona Spain*, 293-296. doi: 10.1145/1864708.1864770.
9. Zhang, Y., Bai, Y., Chang, J., Zang, X., Lu, S., Lu, J., Feng, F., Niu, Y., Song, Y. (2023) Leveraging watch-time feedback for short-video recommendations: A causal labeling framework. *Proceedings of the 32-nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21-25, 2023, Birmingham, UK,* 4952-4959. doi.org/10.1145/3583780.3615483.
10. Xu, Z., Ruibo, M., Jiaqi, C., Weiqi, Z., Ping, Y., Yao, H. (2025) Multi-Granularity Distribution Modeling for Video Watch Time Prediction via Exponential-Gaussian Mixture Network. *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22-26, 2025, Prague Czech Republic,* 309-318. doi: 10.1145/3705328.3748080.
11. Covington, P., Adams, J., Sargin, E. (2016) Deep neural networksfor youtube recommendations. *Proceedings of the 10-th ACM conference onrecommender systems, September 15-19, 2016, Boston Massachusetts, USA*, 191-198. doi: 10.1145/2959100.2959190.
12. Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., Gai, K. (2018) Deep interest network for click-throughrate prediction. *Proceedings of the 24-th ACM SIGKDD international conference on knowledge discovery & data mining (KDD '18), August 19-23, 2018, London, UK,* 1059-1068. doi: 10.1145/3219819.3219823.
13. Zhuang, X., Huang, Y., Palaniappan, K., Zhao, Y. (1996) Gaussian mixture density modeling, decomposition, and applications. *IEEE Transactions on Image Processing*, 5(9), 1293-1302. doi: 10.1109/83.535841.
14. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Xu, J., Wang, Z., Shi, Y., Jiang, T., Li, S.,·Zhang, H., Huang, Y., Qiao, Y., Wang Y., Wang, L. (2025) InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. *Proceedings of the European Conference on Computer Vision (ECCV). Computer Vision – ECCV 2024: 18th European Conference, September 29 – October 4, 2024, Milan, Italy*, Part LXXXV, 396-416. doi: 10.1007/978-3-031-73013-9_23.
15. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139. [Online]. Available: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.
16. Zhao, L., Gundavarapu, N. B., Yuan, L., Zhou, H., Yan, S., Sun, J. J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., Hornung, R., Schroff, F., Yang, M. H., Ross, D. A., Wang, H., Adam, H., Sirotenko, M., Liu, T., Gong, B. (2024) Videoprism: A foundational visual encoder for video understanding. *Proceedings of the 41st*

International Conference on Machine Learning, PMLR, 235, 60785-60811. [Online].
Available:
https://proceedings.mlr.press/v235/zhao24f.html.

17. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Basil, M., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X. (2025) SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *ArXiv preprint arXiv:2502.14786v1*.
doi: 10.48550/arXiv.2502.14786.

18. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A. (2021) Emerging properties in self-supervised vision transformers. *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 10-17 Oct. 2021, Montreal, QC, Canada*, 9650-9660.
doi: 10.1109/ICCV48922.2021.00951.

19. Ding, J., Xue, N., Xia, G.-S., Dai, D. (2022) Decoupling zero-shot semantic segmentation. *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE/CVF, 18-24 June 2022, New Orleans, LA, USA*. 11583-11592.
doi: 10.1109/CVPR52688.2022.01129.

20. Fang, A., Jose, A. M., Jain, A., Schmidt, L., Toshev, A. T., Shankar, V. (2024) Data filtering networks. *Proceedings of 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2637. [Online]. Available: https://openreview.net/pdf?id=ZKtZ7KQ6G5.

21. Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A. (2014) The role of context for object detection and semantic segmentation in the wild. *Proceedings of 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 23-28 June 2014, Columbus, OH, USA*, 891-898.
doi: 10.1109/CVPR.2014.119.

22. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T. L. (2016) Modeling context in referring expressions. *Proceedings of 14th European Conference on Computer Vision (ECCV), October 11-14, 2016, Amsterdam, Netherlands, Part II*, 69-85.

23. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A. (2018) Semantic understanding of scenes through the ADE20k dataset. *International Journal of Computer Vision*, 127(3), 302-321.
doi: 10.1007/s11263-018-1140-0.

24. Niizumi, D. , Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K. (2024) Masked Modeling Duo: Towards a Universal Audio Pre-Training Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 2391-2406.
doi: 10.1109/TASLP.2024.3389636.

25. Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K. (2022) Composing general audio representation by fusing multilayer features of a pre-trained model. *Proceedings of 30th European Signal Processing Conference (EUSIPCO), 29 August 2022 – 02 September 2022, Belgrade, Serbia*, 200-204.
doi: 10.23919/EUSIPCO55093.2022.9909674.

26. Ghosh, S., Seth, A., Umesh, S. (2022) Decorrelating feature spaces forlearning general-purpose audio representations. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 1402-1414.
doi: 10.1109/JSTSP.2022.3202093.

27. Elandt-Johnson, R.C., Johnson, N. L. (1999) *Survival models and data analysis*. 1st ed. Wiley-Interscience.

28. Branders, S., Frénay, B., Dupont, P. (2015) Survival Analysis with Cox Regression and Random Non-linear Projections. *Proceedings of the 23th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 22-24 Apr. 2015, Bruges, Belgium*. [Online].
Available:
https://www.researchgate.net/publication/275653885_Survival_Analysis_with_Cox_Regression_and_Random_Non-linear_Projections#fullTextFileContent.

29. Ducasse, J., Kljun, M., Attygallea, N. T., Pucihar, K. C. (2022) Interactive Web documentaries: a case study of video viewing behaviour on iOtok. *International journal of human-computer interaction*, 38(10), 949-972.
doi: 10.1080/10447318.2021.1976511.

30. Wu, M., Lin, L., Zhang, W., Wang X., Yang, Z., Hu, S. (2025) Preserving AUC fairness in learning with noisy protected groups. *ArXiv preprint arXiv:2505.18532v1*. [Online].
Available: https://arxiv.org/pdf/2505.18532v1.

31. Chay, C. (2025) Understanding ROC-AUC and single-factor AR: a practical guide to classification metrics. Measuring and understanding classifier performance. *Medium.* [Online]. Available: https://medium.com/@carelchay/understanding-roc-auc-and-single-factor-ar-a-practical-guide-to-classification-metrics-ca34ec2cbb1d.

The article has been sent to the editors 01.03.26.
After processing 14.03.26.
Submitted for printing 30.03.26.