POPULARITY PREDICTABILITY DEPENDS ON INFORMATION SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Micro-video popularity prediction draws on information ranging from broad context to item content and historical references. However, existing methods largely ignore that information at different levels contributes differently to predictability, and instead collapse them into a single prediction. We propose **Scale-Structured Predictability Modeling (SSPM)**, which treats popularity predictability as a function of information scale. SSPM organizes increasingly informative evidence into a nested prediction process, where each level targets the variation that becomes predictable beyond the preceding one. (1) **Context Prior Modeling** establishes the predictable component supported by population or temporal context. (2) **Multimodal Evidence Refinement** identifies the additional variation explained by item-level multimodal evidence. (3) **Historical Analogy Refinement** further captures recurrent local variation shared with similar historical instances, using out-of-fold residuals to separate genuinely new information from what earlier levels have already explained. Together, these components form a progressive decomposition of predictability rather than three independent predictors. Experiments on three micro-video popularity benchmarks show that SSPM improves prediction performance by up to 19.77% in nMSE while revealing how predictive gains emerge across information scales. These results support our main finding: popularity predictability depends on information scale, and explicitly modeling this structure leads to both stronger prediction and better understanding of where predictive gains come from.

1 INTRODUCTION

Human perception rarely resolves uncertainty from a single observation: contextual knowledge provides coarse expectations, current sensory evidence refines them, and accumulated experience resolves remaining ambiguity. Popularity prediction follows a similar principle. Micro-video popularity depends on platform or temporal context, current-item content, author and metadata cues, and related historical instances. Prior work has improved prediction through multimodal encoding and retrieval augmentation (Zhang et al., 2023; Xie et al., 2021; Zhong et al., 2024; Cheng et al., 2025b; 2024a; Wang et al., 2026), typically by mapping the available evidence to popularity through a concatenated or jointly fused predictor.

The resulting evidence does not reveal predictability all at once: predictability itself has a hierarchy. Population or temporal context supports broad regularities shared across environments; current-item multimodal evidence explains item-specific deviations; and historical analogies expose recurrent patterns within local neighborhoods. We call the generalization scope of such a regularity its *information scale*. Information scale is therefore not merely a grouping of features, but a hierarchy describing where different portions of target variation become predictable.

One-shot fusion can improve endpoint accuracy, but it entangles regularities with different generalization scopes. Consequently, feature accumulation cannot identify where predictability originates or at which scale additional target variation becomes predictable; full-target fitting at multiple stages may relearn structure captured at coarser scales; and endpoint metrics cannot distinguish gains from context, current-item evidence, or local history. Figure 1 illustrates why a scale-structured view is needed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

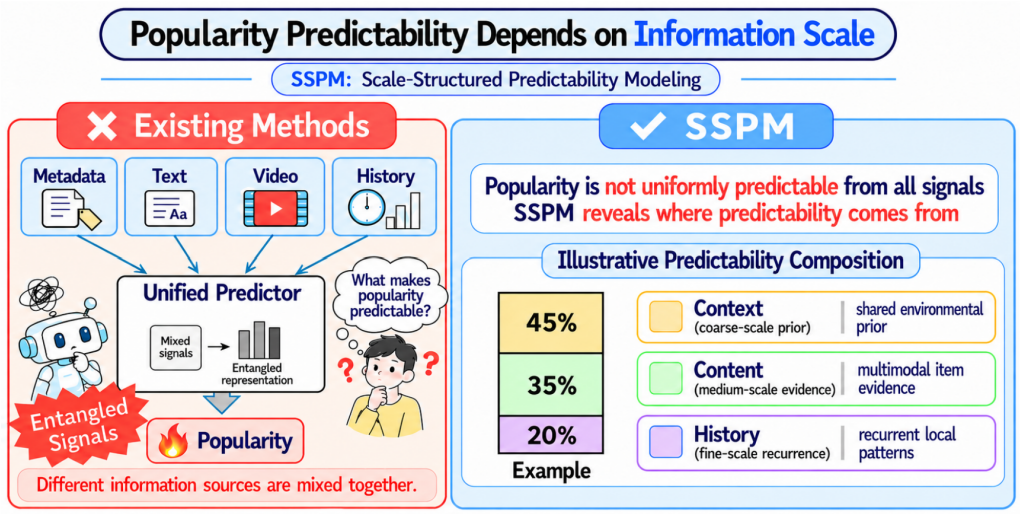

Figure 1: Illustration of the key insight behind SSPM. Unlike existing approaches that entangle heterogeneous signals into a single predictor, SSPM reveals that popularity predictability emerges progressively across information scales. It organizes prediction as nested refinements over coarse-scale context priors, medium-scale multimodal evidence, and fine-scale historical recurrence, enabling the model to reveal at which information scale predictive gains emerge.

We instead view popularity prediction as a *progressive revelation* process. Coarse context first reveals an initial predictable component and leaves an unresolved residual; current-item evidence then explains part of that residual and leaves finer uncertainty; historical analogy finally captures locally recurrent residual structure. Each stage should therefore target the variation newly resolved at its information scale.

We instantiate this view through SSPM (*Scale-Structured Predictability Modeling*), a framework for progressive decomposition of predictability whose modules explain additional predictability beyond preceding scales rather than act as independent predictors. **Context Prior Modeling** captures population or temporal structure available before item observation. **Multimodal Evidence Refinement** uses current-item evidence to predict the unresolved target, optionally conditioning on context. **Historical Analogy Refinement** estimates remaining local structure by aggregating OOF residuals from similar historical instances. The resulting prefixes form the coarse-to-fine pathway in Figure 2, jointly producing the endpoint prediction and attributing its gains to the scale at which they emerge.

Section 3 grounds this perspective by formalizing prediction-time admissible information as a filtration. Under squared loss, the corresponding conditional predictors have orthogonal scale increments, providing the statistical basis for unresolved-target learning.

Our contributions are:

- **Information-scale perspective.** We introduce a hierarchy of predictability defined by the scope over which each regularity generalizes, and ground it through a filtration whose scale increments are orthogonal martingale differences under squared loss.
- **Scale-structured modeling.** We propose SSPM as a methodological realization of this perspective, approximating the scale-wise decomposition through cross-fitted residual learning and a historical bank indexed by OOF residuals.
- **Predictability attribution.** We introduce held-out scale attribution to identify where predictive gains originate and reveal how the dominant predictive scale varies across datasets.

2 Related Work

Popularity dynamics and temporal context. Popularity forecasting has long treated popularity as an evolving process. Multi-scale temporal decomposition and diffusion models capture how

popularity develops over time (Wu et al., 2016; Cheng et al., 2024b), while successive Social Media Prediction challenges have established multimodal benchmarks and evaluation settings (Wu et al., 2019; 2023; 2024; 2025). In SSPM, temporal variables enter the Context scale only when they are available under the declared prediction protocol.

Current-item evidence. Cold-start popularity models combine content with author, social, and contextual variables. Early work studies visual popularity cues (Khosla et al., 2014), while micro-video and social-media models extend this direction through transductive multimodal learning, attention, feature engineering, and latent-variable prediction (Chen et al., 2016; Xu et al., 2020; Lai et al., 2020; Hsu et al., 2023; Zhang et al., 2023; Xie et al., 2021; Mao et al., 2023; Ye et al., 2025). Modern item encoders exploit transferable image, video, and state-space representations, including CLIP, I3D, SlowFast, TSM, TimeSformer, and ViViT (Radford et al., 2021; Carreira & Zisserman, 2017; Feichtenhofer et al., 2018; Lin et al., 2018; Bertasius et al., 2021; Arnab et al., 2021; Gu & Dao, 2024; Li et al., 2024a;b). These approaches typically improve representations within a full-target predictor. In contrast, SSPM assigns prediction-time item variables to Multimodal Evidence Refinement and trains the module on the unresolved target left by the Context prefix.

Retrieval and historical memory. Recent retrieval methods use multimodal references, hypergraph memory, semantic priors, in-context demonstrations, calibrated alignment, or temporal reranking (Zhong et al., 2024; Cheng et al., 2024a; 2025a;b; Heo & Lee, 2026; Wang et al., 2026; Li et al., 2026b;a). This direction builds on retrieval-augmented generation, dense passage retrieval, and late-interaction retrieval (Lewis et al., 2020; Karpukhin et al., 2020; Khattab & Zaharia, 2020), with scalable memory access supported by product quantization, anisotropic vector quantization, and graph-based or hybrid nearest-neighbor search (Jégou et al., 2011; Guo et al., 2020; Malkov & Yashunin, 2020; Yang et al., 2026). Existing retrieval-based methods may use retrieved examples, outcomes, or demonstrations within a full-target predictor. SSPM likewise retrieves similar historical instances, but Historical Analogy Refinement aggregates their OOF residuals rather than their full outcomes or independent full-target predictions, making the historical contribution separable from population context and transferable item evidence.

Local estimation and cross-fitting. Kernel and nearest-neighbor regression provide classical foundations for estimating local response structure (Stone, 1977; Altman, 1992). Cross-validated predictions support stacked generalization and Super Learner ensembles (Wolpert, 1992; van der Laan et al., 2007), while cross-fitting decouples learned nuisance components from their evaluation observations in semiparametric estimation (Chernozhukov et al., 2018). SSPM uses cross-fitting to give unresolved targets an out-of-sample interpretation before they are assigned to a finer information scale. It therefore focuses on how heterogeneous evidence should be ordered, residualized, and attributed.

3 FORMULATION

3.1 INFORMATION SCALE

Definition 1 (Information scale)**.** *An information scale is a prediction-time information set whose associated regularities are assumed to generalize over a specified statistical scope. It is therefore distinct from modality or representation resolution: visual, textual, or metadata variables may belong to the same scale when they share the same prediction-time scope.*

Let $Y \in L^2$ denote popularity. We represent increasingly specific admissible information by the filtration

$$\mathcal{F}_{-1} \subseteq \mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_L,$$

where each $\mathcal{F}_\ell$ is the sigma-field generated by information available at prediction time. The trivial sigma-field $\mathcal{F}_{-1}$ defines the No-info reference; it is a baseline, not a fourth SSPM scale. The three informative scales are $\mathcal{F}_0$, $\mathcal{F}_1$, and $\mathcal{F}_2$, which respectively contain population or temporal context, current-item evidence, and admissible historical references. The order is declared as part of the prediction task. The resulting decomposition is unique relative to that filtration; reordering the scales defines a different decomposition.

3.2 SCALE-WISE PREDICTABILITY DECOMPOSITION

Under squared loss, the Bayes predictor and risk at scale ℓ are

$$m_\ell^* = \mathbb{E}[Y \mid \mathcal{F}_\ell], \qquad R_\ell^* = \mathbb{E}\big[(Y - m_\ell^*)^2\big]. \tag{1}$$

Predictability supported by $\mathcal{F}_\ell$ relative to No-info is $P^*(\mathcal{F}_\ell) := R_{-1}^* - R_\ell^*$.

Define the scale increment $\delta_\ell^* = m_\ell^* - m_{\ell-1}^*$. The finest predictor admits the telescoping representation

$$m_L^* = m_{-1}^* + \sum_{\ell=0}^{L} \delta_\ell^*. \tag{2}$$

Proposition 1 (Orthogonal scale increments)**.** *The conditional predictors $\{m_\ell^*\}_{\ell=-1}^{L}$ form a Doob martingale with respect to $\{\mathcal{F}_\ell\}$. Consequently,*

$$\mathbb{E}[\delta_\ell^* \mid \mathcal{F}_{\ell-1}] = 0, \quad \mathbb{E}[\delta_j^* \delta_\ell^*] = 0 \ (j < \ell), \quad R_{-1}^* - R_L^* = \sum_{\ell=0}^{L} \mathbb{E}[(\delta_\ell^*)^2]. \tag{3}$$

In particular, $R_{\ell-1}^* - R_\ell^* = \mathbb{E}[(\delta_\ell^*)^2] \geq 0$.

The proof follows from the tower property of conditional expectation and the standard L_2 projection identity.

The proposition formalizes the learning principle behind scale-structured refinement. Bayes-optimal increments do not relearn variation represented at earlier scales; they form orthogonal martingale differences. The mathematical identities are standard properties of conditional expectation. Our formulation uses them to connect a declared prediction-time filtration to scale-specific targets, estimation, and attribution.

Empirical contribution. For a fixed held-out set, let $\widehat{R}_\ell$ denote the squared risk of the prefix through scale ℓ, and define its predictive contribution as $\widehat{A}_\ell := \widehat{R}_{\ell-1} - \widehat{R}_\ell$. This quantity is measured in risk units. MAE and Spearman correlation are reported as complementary diagnostics rather than terms in the squared-risk decomposition.

4 METHOD

4.1 OVERVIEW: PROGRESSIVE PREDICTABILITY DECOMPOSITION

Progressive decomposition. Different information scales reveal different portions of predictable variation: context, current-item evidence, and historical recurrence operate at coarse, medium, and fine scales. SSPM therefore follows the progressive path in Figure 2: Context Prior Modeling establishes a coarse baseline, Multimodal Evidence Refinement explains its item-level residual, and Historical Analogy Refinement corrects recurrent residual structure. Each stage preserves the preceding prefix and is trained to target what remains unexplained.

Unresolved-target refinement. For query q, let C_q denote coarse-scale context, X_q medium-scale current-item evidence, and H_q the fine-scale historical state available at prediction time. A Context Prior estimator T, Multimodal Evidence refiner G, and Historical Analogy refiner M construct the nested prefixes

$$\widehat{m}_q^{(0)} = T(C_q), \quad \widehat{m}_q^{(1)} = \widehat{m}_q^{(0)} + G(C_q, X_q), \quad \widehat{y}_q = \widehat{m}_q^{(2)} = \widehat{m}_q^{(1)} + M(C_q, X_q, H_q). \tag{4}$$

The three stages are the nested prefixes $\widehat{y}_q^{(1)} \equiv \widehat{m}_q^{(0)}$, $\widehat{y}_q^{(2)} \equiv \widehat{m}_q^{(1)}$, and $\widehat{y}_q^{(3)} \equiv \widehat{m}_q^{(2)} = \widehat{y}_q$, rather than independent predictors. The sigma-fields in Section 3 define the population targets, while Equation 4 approximates their Bayes increments using finite-dimensional inputs (C_q, X_q, H_q).

4.2 CONTEXT PRIOR MODELING

Coarse-scale prior. Context Prior Modeling captures broad temporal and population regularities as coarse prediction-time structure, defining the baseline $\widehat{y}_q^{(1)}$. Depending on the benchmark, C_q may

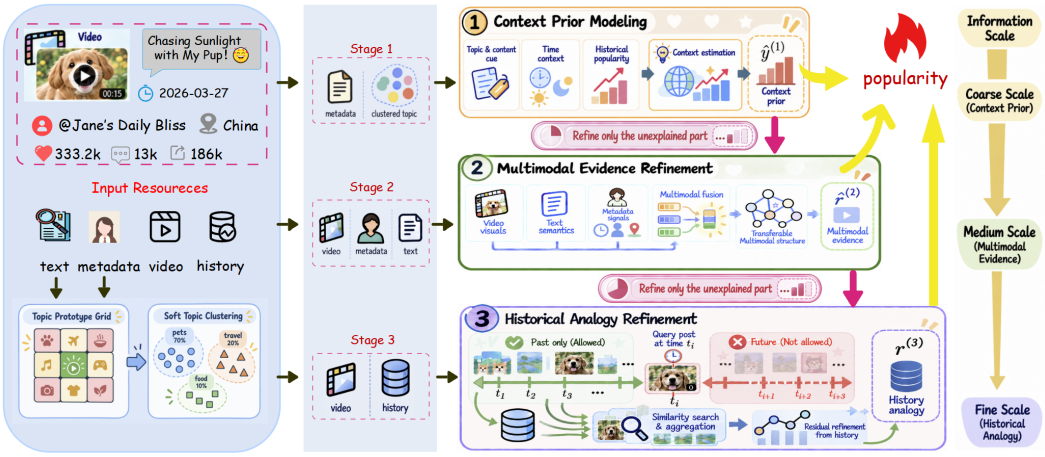

Figure 2: Overview of Scale-Structured Predictability Modeling (SSPM). SSPM organizes popularity prediction as a progressive refinement process over different information scales. It first estimates coarse-scale context priors, then refines unexplained residuals with medium-scale multimodal evidence, and finally captures fine-scale recurrent patterns through historical analogy. Each stage targets the remaining uncertainty beyond previous stages, forming a structured pathway from coarse- to fine-scale predictability.

include population statistics, calendar variables, coarse topic or community context, and aggregate statistics available before prediction; item semantics belong to X_q, and sample-level references to H_q.

OOF residual. When no informative query-varying context exists, T reduces to a training-population prior. Its cross-fitted prediction yields

$$r^{(1)}_{i,\mathrm{OOF}} = y_i - \widehat{m}^{(0)}_{i,\mathrm{OOF}}.$$

We call this cross-fitted unresolved target the *OOF residual*; it is the target for Stage 2.

4.3 Multimodal Evidence Refinement

Item-level evidence. Multimodal Evidence Refinement asks what additional variation becomes predictable once the current item is observed. It targets the residual rather than the original label, because another full-target predictor could relearn the Context prior and obscure the source of improvement. The medium-scale input X_i contains visual content, text, and prediction-time item or author metadata; it excludes post-publication behavior unavailable at prediction time. SSPM fits

$$\widehat{\Delta}^{(1)}_i = G(C_i, X_i), \qquad \widehat{m}^{(1)}_i = \widehat{m}^{(0)}_i + \widehat{\Delta}^{(1)}_i$$

against $r^{(1)}_{i,\mathrm{OOF}}$. Conditioning G on both C_i and X_i permits context–item interactions when supported by the backbone; a restricted $G(X_i)$ is an additive approximation.

Residual objective. The encoder may be linear, tree-based, or neural; the defining choice is the residual target. In the population formulation, the new component is $m^*_1 - m^*_0$; finite-sample cross-fitting approximates it without relearning the coarse prior. The prefix $\widehat{y}^{(2)}_i \equiv \widehat{m}^{(1)}_i$ is trained on the cross-fitted Context residual, approximating the additional variation explained by multimodal evidence. Fine-scale recurrence may still remain, motivating historical analogy after this prefix is established.

4.4 Historical Analogy Refinement

Local recurrence. Historical recurrence is modeled last because it is the finest, most local scale. Retrieving outcomes earlier could reintroduce structure already available to the first two stages.

Historical Analogy Refinement instead searches an admissible neighborhood for similar historical instances and aggregates their Context-to-Multimodal residuals to estimate recurrent variation.

Residual memory. Cross-fitting the complete prefix produces the Historical Residual Bank

$$r^{(2)}_{i,\mathrm{OOF}} = y_i - \widehat{m}^{(1)}_{i,\mathrm{OOF}}, \qquad \mathcal{M} = \{(\mathbf{z}_i, r^{(2)}_{i,\mathrm{OOF}}, t_i)\}_{i=1}^{N},$$

where $\mathbf{z}_i$ is the retrieval representation and t_i records temporal information when available. For query q, H_q determines an admissible reference pool $\mathcal{R}(H_q)$. The K most similar references $\mathcal{N}_K(q) \subseteq \mathcal{R}(H_q)$ yield

$$\widehat{\Delta}^{(2)}_q = \alpha \sum_{i \in \mathcal{N}_K(q)} w_{qi} \widetilde{r}^{(2)}_{i,\mathrm{OOF}}, \qquad \widehat{m}^{(2)}_q = \widehat{m}^{(1)}_q + \widehat{\Delta}^{(2)}_q. \tag{5}$$

Here w_{qi} are normalized similarity weights, $\widetilde{r}^{(2)}_{i,\mathrm{OOF}}$ permits validation-selected clipping, and α controls correction strength. Prediction and retrieval representations may differ because transferable item prediction and local residual recurrence are distinct estimation problems.

Residual correction. Aggregating the retrieved neighbors' residuals rather than their labels prevents population- and item-level structure already explained by earlier scales from being counted again. Each OOF residual records what the Context-to-Multimodal prefix failed to explain for a historical instance, so the retrieved neighbors supply a local correction rather than a second full-target predictor. The resulting $\widehat{y}^{(3)}_q \equiv \widehat{m}^{(2)}_q$ is final. Strict temporal protocols require $t_i < t_q$; when query-level timestamps are unavailable, retrieval is restricted to the chronologically earlier training partition defined by the benchmark protocol. The correction is zero when no admissible reference exists.

4.5 Cross-Fitted Optimization

Fold-local construction. Unresolved targets must reflect unseen-observation errors. For each outer holdout V_b, T is cross-fitted within the admissible training portion $\mathcal{D}_b$ to construct fold-local $r^{(1)}_{j,\mathrm{OOF}|b}$ targets. The refiner G_b is trained only on these targets, then $T_b + G_b$ is evaluated on V_b to obtain $r^{(2)}_{i,\mathrm{OOF}}$; thus no label from V_b enters G_b's training targets.

Generalization target. Cross-fitting serves a statistical role beyond leakage prevention. An in-sample residual may primarily reflect interpolation by the preceding estimator and need not characterize unresolved structure for a new observation. Its cross-fitted counterpart approximates the error encountered by the learned prefix on an unseen item, giving the next scale an out-of-sample target. Standard K-fold splits are used when exchangeability is assumed; strict temporal protocols use forward blocks so that every fitted component and historical reference comes from the query's past.

Algorithm 1 Cross-Fitted Training and Inference of SSPM

Require: Training data $\mathcal{D}$, declared inputs (C, X, H), admissible fold policy, retrieval rule

for each held-out fold or forward block V_b **do**
  Within $\mathcal{D}_b$, cross-fit target-dependent transformations and T to form $r^{(1)}_{j,\mathrm{OOF}|b}$ for $j \in \mathcal{D}_b$
  Refit T_b on $\mathcal{D}_b$; fit G_b only to $\{((C_j, X_j), r^{(1)}_{j,\mathrm{OOF}|b}) : j \in \mathcal{D}_b\}$
  Evaluate T_b and $T_b + G_b$ on $i \in V_b$; form and store $(\mathbf{z}_i, r^{(2)}_{i,\mathrm{OOF}}, t_i)$
end for
Cross-fit T on the full training partition to form $r^{(1)}_{i,\mathrm{OOF}}$; refit T and fit final G only to these targets; index $\mathcal{M}$
for each query q **do**
  Compute $\widehat{m}^{(0)}_q$, $\widehat{m}^{(1)}_q$, and $\widehat{m}^{(2)}_q$ using Eqs. 4 and 5
  Return $\widehat{y}_q = \widehat{m}^{(2)}_q$ and retain all prefixes for scale attribution
end for

Deployment. Final training cross-fits T over the full partition to form $r^{(1)}_{\mathrm{OOF}}$, then refits T and fits G to these targets; the retained bank and prefixes support Section 5.4 attribution. Unlike boosting over a shared information set or stacking full-target experts, SSPM introduces a new prediction-time scope at each transition and learns the preceding prefix's held-out error. Cross-fitting requires outer prefix fits plus the inner fits used to construct fold-local residual targets, while retrieval stores one representation and one residual per reference and permits approximate indexing.

Table 1: Popularity-prediction performance against the baselines reported by STAP. Best results are bold, second-best results are underlined, and the final row gives the relative gain over STAP.

Method	Type	SMPD			MicroLens			Informs		
		nMSE ↓	MAE ↓	SRC ↑	nMSE ↓	MAE ↓	SRC ↑	nMSE ↓	MAE ↓	SRC ↑
SVR	Feature	0.8102	1.9559	0.4455	0.7829	1.0899	0.4796	0.7243	0.0997	0.5720
HyFea	Feature	0.8932	2.0711	0.3416	0.7481	1.0846	0.4974	0.8185	0.1042	0.4823
MFTM	Feature	0.8822	2.0182	0.3734	0.7826	1.1111	0.4662	0.6803	0.0966	0.6357
CLSTM	Multimodal	0.8477	1.9628	0.4187	1.0832	1.1830	0.3332	0.8610	0.1098	0.4987
TMALL	Multimodal	1.2025	2.3752	0.2137	0.7567	1.0830	0.4871	6.9960	0.3115	0.0674
MASSL	Multimodal	1.1334	2.3037	0.3788	1.1090	1.3206	0.0120	1.2386	0.1343	0.0596
CBAN	Multimodal	0.8777	2.0011	0.3951	0.9906	1.2542	0.1797	0.8482	0.1093	0.5064
HMMVED	Multimodal	0.9254	2.0455	0.4331	1.1912	1.1326	0.4311	1.0762	0.1260	0.0564
MQMC	Multimodal	6.1899	1.9343	0.4274	2.4157	1.2369	0.4002	0.8524	0.1116	0.3941
MMRA	Retrieval	0.9144	2.0345	0.3849	0.7512	1.1887	0.4951	0.8391	0.1084	0.5052
ICPF	Retrieval	0.7611	1.8358	0.4980	0.7305	1.1088	0.5102	0.9551	0.1153	0.4985
STAP	Retrieval	<u>0.4887</u>	<u>1.5148</u>	<u>0.6704</u>	<u>0.7148</u>	<u>1.0482</u>	<u>0.5346</u>	<u>0.5806</u>	<u>0.0876</u>	<u>0.6460</u>
SSPM	Scale-wise	**0.4839**	**1.4466**	**0.6764**	**0.6509**	**1.0059**	**0.5866**	**0.4658**	**0.0787**	**0.7372**
Improv.	over STAP	0.98% ↓	4.50% ↓	0.89% ↑	8.94% ↓	4.04% ↓	9.73% ↑	19.77% ↓	10.16% ↓	14.12% ↑

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets, protocols, and information scales. We evaluate SSPM on SMPD (Wu et al., 2019), MicroLens (Ni et al., 2025), and Informs (Lu et al., 2025) using temporally ordered splits: 8:1:1 for SMPD and Informs and 7:1:2 for MicroLens. Historical references satisfy query-time eligibility when timestamps are available and are otherwise restricted to the training partition. Following SSPM, signals are assigned to Context Prior, Multimodal Evidence, and Historical Analogy according to the scope over which they are available at prediction time.

Evaluation metrics. We use nMSE as the primary metric because it directly matches the squared-risk decomposition in Proposition 1; MAE and Spearman rank correlation (SRC) provide complementary error and ranking measurements.

5.2 MAIN RESULTS

Overall. Table 1 compares SSPM with feature-engineering, multimodal-fusion, and retrieval baselines (Ghosh et al., 2016; Lai et al., 2020; Cheung & Lam, 2022; Zhang et al., 2023; Du et al., 2023; Hsu et al., 2023; Xie et al., 2021; Zhong et al., 2024; Cheng et al., 2025b; Wang et al., 2026). SSPM achieves the best result on every dataset and metric, improving all nine measurements over the strongest baseline, STAP. The advantage is not confined to one criterion: lower nMSE and MAE indicate more accurate popularity magnitudes, while higher SRC confirms better ranking. The agreement across metrics shows that scale-wise refinement improves calibration and ordering simultaneously rather than exchanging one objective for another.

Gains. Relative to STAP, SSPM changes nMSE/MAE/SRC by $-0.98\%/-4.50\%/+0.89\%$ on SMPD, $-8.94\%/-4.04\%/+9.73\%$ on MicroLens, and $-19.77\%/-10.16\%/+14.12\%$ on Informs. SMPD shows a modest nMSE and SRC advantage but a clearer MAE reduction, suggesting that its main benefit is correcting prediction magnitude. MicroLens exhibits balanced improvements in error and ranking, whereas Informs yields the largest gains on all three metrics. Thus, the endpoint improvements are consistently positive but vary substantially in size, motivating the analyses below of which design principles and information scales produce them.

5.3 DESIGN PRINCIPLE ANALYSIS

Principles. We test whether Scale-Structured Refinement (SSR) must learn unresolved targets stage by stage and whether Historical Analogy Refinement (HAR) must aggregate cross-fitted residual information rather than retrieved labels. Table 2 compares matched variants on MicroLens and

Table 2: Matched ablation study of SSR and HAR using identical prediction-time inputs, model capacity, retrieval features, neighbor sets, and tuning budgets. Parentheses report direction-aware changes relative to the full SSPM model (green: Improv.; red: degradation).

Module	Variant	MicroLens			SMPD		
		nMSE ↓	MAE ↓	SRC ↑	nMSE ↓	MAE ↓	SRC ↑
SSPM	Full	**0.6509**	**1.0059**	**0.5866**	**0.4839**	**1.4466**	**0.6764**
SSR	One-shot	0.6669 (-2.46%)	1.0238 (-1.78%)	0.5762 (-1.77%)	0.4872 (-0.68%)	1.4605 (-0.96%)	0.6702 (-0.92%)
	Full-target	0.6712 (-3.12%)	1.0273 (-2.13%)	0.5725 (-2.40%)	0.4903 (-1.32%)	1.4826 (-2.49%)	0.6696 (-1.01%)
	w/o OOF	0.6657 (-2.27%)	1.0233 (-1.73%)	0.5771 (-1.62%)	0.5425 (-12.11%)	1.4854 (-2.68%)	0.6697 (-0.99%)
	w/o context	0.6712 (-3.12%)	1.0270 (-2.10%)	0.5714 (-2.59%)	0.4925 (-1.78%)	1.4915 (-3.10%)	0.6698 (-0.98%)
HAR	w/o history	0.6657 (-2.27%)	1.0233 (-1.73%)	0.5771 (-1.62%)	0.4872 (-0.68%)	1.4605 (-0.96%)	0.6702 (-0.92%)
	Label retrieval	1.0209 (-56.84%)	1.2632 (-25.58%)	0.5787 (-1.35%)	0.9541 (-97.17%)	2.0699 (-43.09%)	0.6808 (+0.65%)
	In-sample bank	0.6657 (-2.27%)	1.0233 (-1.73%)	0.5772 (-1.60%)	0.4837 (+0.04%)	1.4516 (-0.35%)	0.6744 (-0.30%)
	Uniform Top-K	0.6608 (-1.52%)	1.0191 (-1.31%)	0.5802 (-1.09%)	0.4845 (-0.12%)	1.4509 (-0.30%)	0.6727 (-0.55%)
	Random neighbor	0.6662 (-2.35%)	1.0238 (-1.78%)	0.5771 (-1.62%)	0.4913 (-1.53%)	1.4631 (-1.14%)	0.6636 (-1.89%)

SMPD under identical inputs, capacity, retrieval representations, neighbor sets, and tuning budgets. The SSR controls isolate progressive targets, OOF construction, and context conditioning; the HAR controls separately test the retrieved quantity, residual-bank construction, aggregation weights, and neighbor relevance.

Refinement. Every SSR variant underperforms the full model across both datasets and all metrics. On MicroLens, one-shot and full-target learning raise nMSE from 0.6509 to 0.6669 and 0.6712, showing that adding the same predictors is insufficient when later stages can relearn already-explained structure. Removing context causes a similar 3.12% nMSE degradation. On SMPD, removing OOF construction produces the largest SSR nMSE degradation (12.11%), highlighting the importance of learning from held-out unresolved error rather than an in-sample residual.

Retrieval. HAR shows an even sharper distinction between retrieving residual structure and retrieving outcomes. Direct-label retrieval worsens nMSE by 56.84% on MicroLens and 97.17% on SMPD because neighbor labels reintroduce variation already represented by the prefix. Uniform aggregation and random neighbors also underperform, demonstrating that neighborhood relevance and similarity weighting both matter. Although the in-sample bank marginally improves SMPD nMSE by 0.04%, it degrades the other five measurements. Together, these controls attribute the gain to stage-specific residual targets, cross-fitted memory, and similarity-aware historical correction rather than additional capacity or retrieval alone.

5.4 Information Scale Analysis

Scale-wise attribution. Rather than stopping at endpoint accuracy, we trace the held-out nMSE across nested prediction prefixes. Each dataset therefore induces a *predictability profile*, with the contribution of scale ℓ measured by $\widehat{A}_\ell = \widehat{R}_{\ell-1} - \widehat{R}_\ell$ and normalized by the total reduction when reporting relative shares. This construction links each attributed share to an observed change in predictive risk.

Dataset-specific profiles. Multimodal Evidence accounts for 85.6% of the measured reduction on SMPD and 95.8% on MicroLens, while Context Prior accounts for 93.2% on Informs. Figure 3(b) independently supports these shares: the sharpest prefix-wise nMSE drop occurs at the corresponding dataset-specific transition. Historical Analogy supplies a smaller final correction in all three cases. Thus, similar endpoint objectives can depend on substantially different predictive sources, and the value of an information scale cannot be inferred from aggregate accuracy alone. The contrast also separates information availability from predictive usefulness: a scale may be present in every benchmark yet explain little held-out risk when its regularities do not transfer reliably.

5.5 Qualitative Examples

Instance-level view. Figure 4 presents six SMPD examples comparing MMRA, ICPF, STAP, and SSPM with ground truth. For signed Context, Multimodal, and Historical contributions c, m, h (with h including α), we report normalized absolute shares $100(|c|, |m|, |h|)/S$, where $S = |c| +$

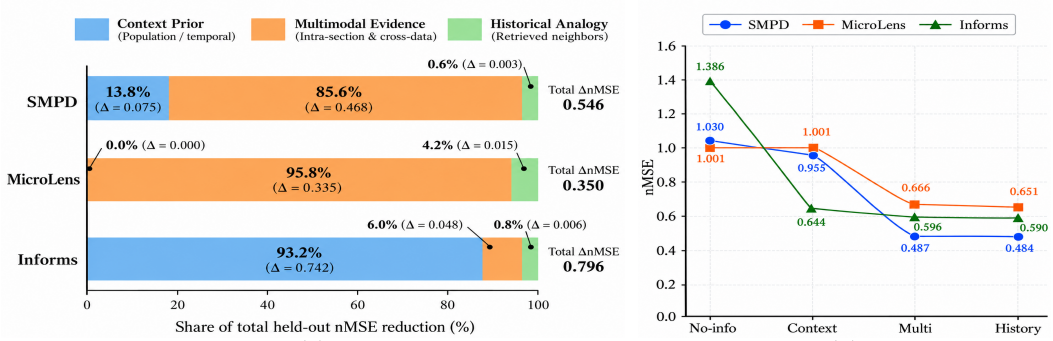

Figure 3: Information-scale analysis: (a) nMSE reduction by scale; (b) prefix-wise nMSE.

$|m|+|h|$, while the prediction remains $c+m+h$. These quantities describe per-instance contribution magnitudes rather than the dataset-level risk reductions in Figure 3. SSPM is closest or nearly closest to ground truth in Videos 1–4 and remains competitive in Videos 5–6.

Example-wise variation. Context Prior has the largest displayed share in all six cases (67.5%–89.9%). Multimodal Evidence rises to 27.5%–30.8% for the scenic landscape and dental demonstration in Videos 3–4, where content cues correct deviations from the coarse prior. Historical Analogy contributes 1.4%–5.9%, with its largest shares in Videos 5–6. The examples therefore complement the dataset-level profile: the prediction pathway is fixed, but the relative contribution of each scale changes with the item. This variation explains how a dataset-level dominant scale can coexist with content-dependent corrections without changing the ordered coarse-to-fine refinement procedure.

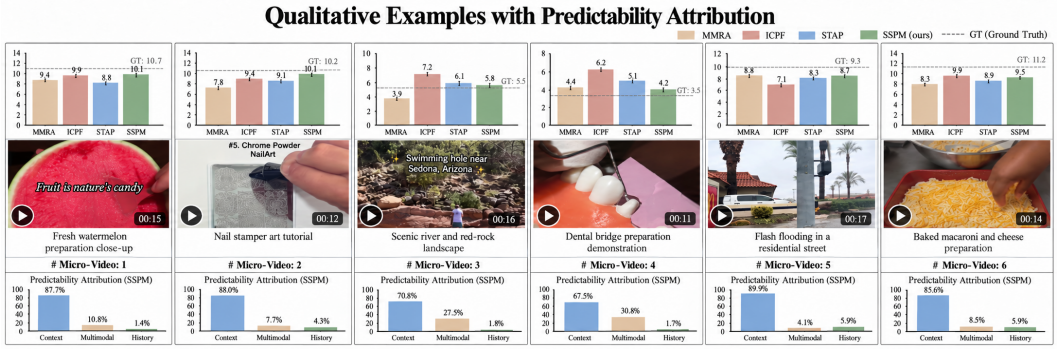

Figure 4: Qualitative examples from SMPD with instance-level attribution. For each micro-video, the upper panel compares predictions with the ground truth, while the lower panel reports the normalized absolute contribution shares of Context Prior, Multimodal Evidence, and Historical Analogy.

6 Conclusion

We introduced information scale as a perspective for understanding where predictive regularities generalize and instantiated this view through SSPM, which progressively targets the variation unresolved by preceding scales using Context Prior Modeling, Multimodal Evidence Refinement, and Historical Analogy Refinement. Experiments on three benchmarks show that predictive gains arise from different scales: item-level evidence contributes most on SMPD and MicroLens, while Context dominates on Informs, with Historical refinement providing an additional fine-scale correction across all three datasets. Qualitative examples further reveal that these contributions vary across individual items. Overall, SSPM reframes popularity prediction from a monolithic endpoint task into a scale-structured decomposition of predictability, showing that the value of predictive evidence depends on the statistical scope over which its regularities generalize.

Reproducibility Statement

We document the information assigned to each scale, dataset-specific protocols, feature construction, split definitions, fold-local preprocessing, out-of-fold unresolved targets, historical reference rules, hyperparameter search spaces, scale-wise attribution, and pre-specified test configurations. We will release an anonymized code package for review.

AI Use Statement

Large language models were used to assist with literature discovery, research ideation, experimental workflow planning, code drafting and debugging, language editing, and drafting parts of the manuscript. All research decisions, experimental design, implementation choices, data analysis, verification of results, and final manuscript content were reviewed and validated by the authors. AI tools were not treated as sources of scientific evidence, and factual and scholarly claims were independently verified by the authors.

References

Naomi S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. doi: 10.1080/00031305.1992.10475879.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6816–6826, 2021. doi: 10.1109/ICCV48922.2021.00676.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *CoRR*, abs/2102.05095, 2021. URL https://arxiv.org/abs/2102.05095.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pp. 898–907, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450336031. doi: 10.1145/2964284.2964314. URL https://doi.org/10.1145/2964284.2964314.

Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 445–455, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3672041. URL https://doi.org/10.1145/3637528.3672041.

Zhangtao Cheng, Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Philip S. Yu. Information cascade popularity prediction via probabilistic diffusion. *IEEE Trans. on Knowl. and Data Eng.*, 36(12):8541–8555, December 2024b. ISSN 1041-4347. doi: 10.1109/TKDE.2024.3465241. URL https://doi.org/10.1109/TKDE.2024.3465241.

Zhangtao Cheng, Jian Lang, Ting Zhong, and Fan Zhou. Seeing the unseen in micro-video popularity prediction: Self-correlation retrieval for missing modality generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 142–152, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709308. URL https://doi.org/10.1145/3690624.3709308.

Zhangtao Cheng, Jiao Li, Jian Lang, Ting Zhong, and Fan Zhou. In-context prompt-augmented micro-video popularity prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11527–11535, 2025b. doi: 10.1609/aaai.v39i11.33254.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Tsun-hin Cheung and Kin-man Lam. Crossmodal bipolar attention for multimodal classification on social media. *Neurocomput.*, 514(C):1–12, December 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.09.140. URL https://doi.org/10.1016/j.neucom.2022.09.140.

Yali Du, Yinwei Wei, Wei Ji, Fan Liu, Xin Luo, and Liqiang Nie. Multi-queue momentum contrast for microvideo-product retrieval. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pp. 1003–1011, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570405. URL https://doi.org/10.1145/3539597.3570405.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018. URL http://arxiv.org/abs/1812.03982.

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. Contextual lstm (clstm) models for large scale nlp tasks. 2016. URL https://arxiv.org/abs/1602.06291.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. 2024. URL https://arxiv.org/abs/2312.00752.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Ryang Heo and Dongha Lee. Will it go viral? grounding micro-video popularity prediction on the open web, 2026. URL https://arxiv.org/abs/2605.18653.

Chih-Chung Hsu, Chia-Ming Lee, Xiu-Yu Hou, and Chi-Han Tsai. Gradient boost tree network based on extensive feature analysis for popularity prediction of social posts. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 9451–9455, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612843. URL https://doi.org/10.1145/3581783.3612843.

Herve Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. doi: 10.1109/TPAMI.2010.57.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pp. 867–876, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442. doi: 10.1145/2566486.2567996. URL https://doi.org/10.1145/2566486.2567996.

Xin Lai, Yihong Zhang, and Wei Zhang. Hyfea: Winning solution to social media popularity prediction for multimedia grand challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 4565–4569, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3416273. URL https://doi.org/10.1145/3394171.3416273.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXVI*, pp. 237–255, Berlin, Heidelberg, 2024a. Springer-Verlag. ISBN 978-3-031-73346-8. doi: 10.1007/978-3-031-73347-5_14. URL https://doi.org/10.1007/978-3-031-73347-5_14.

Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multi-modal fusion with coupled state space model. *arXiv preprint arXiv:2405.18014*, 2024b.

Zixu Li, Yupeng Hu, Zhiwei Chen, Zhiheng Fu, Xiaowei Zhu, Weili Guan, and Liqiang Nie. TempRet: Temporal enhancement and two-stage reranking for CVPR 2026 EPIC-KITCHENS-100 multi-instance retrieval challenge. *arXiv preprint arXiv:2605.24470*, 2026a. doi: 10.48550/arXiv.2605.24470.

Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. ReTrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23373–23381, 2026b. doi: 10.1609/aaai.v40i28.39507.

Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *CoRR*, abs/1811.08383, 2018. URL http://arxiv.org/abs/1811.08383.

Jiacheng Lu, Mingyuan Xiao, Weijian Wang, Yuxin Du, Zhengze Wu, and Cheng Hua. Multi-modal and metadata capture model for micro video popularity prediction, 2025.

Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836, April 2020. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2889473. URL https://doi.org/10.1109/TPAMI.2018.2889473.

Shijian Mao, Wudong Xi, Lei Yu, Gaotian Lü, Xingxing Xing, Xingchen Zhou, and Wei Wan. Enhanced catboost with stacking features for social media prediction. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 9430–9435, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612839. URL https://doi.org/10.1145/3581783.3612839.

Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. A content-driven micro-video recommendation dataset at scale. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 6486–6491, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720406. doi: 10.1145/3746252.3761655. URL https://doi.org/10.1145/3746252.3761655.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977. doi: 10.1214/aos/1176343886.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. doi: 10.2202/1544-6115.1309.

Dali Wang, Yunyao Zhang, Junqing Yu, Yi-Ping Phoebe Chen, Chen Xu, and Zikai Song. Seeing further and wider: Joint spatio-temporal enlargement for micro-video popularity prediction. *arXiv preprint arXiv:2604.20311*, 2026.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. Unfolding temporal dynamics: predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 272–278. AAAI Press, 2016.

Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. Smp challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pp. 2667–2671, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3356084. URL https://doi.org/10.1145/3343031.3356084.

Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. Smp challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 9651–9655, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3613853. URL https://doi.org/10.1145/3581783.3613853.

Bo Wu, Peiye Liu, Qiushi Huang, Zhaoyang Zeng, Jia Wang, Bei Liu, Jiebo Luo, and Wen-Huang Cheng. Smp challenge summary: Social media prediction challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 11442–11444, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3688996. URL https://doi.org/10.1145/3664647.3688996.

Bo Wu, Peiye Liu, Qiushi Huang, Zhaoyang Zeng, Jia Wang, Bei Liu, Jiebo Luo, and Wen-Huang Cheng. Smpv: Social media prediction for videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 14055–14057, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3763757. URL https://doi.org/10.1145/3746027.3763757.

Jiayi Xie, Yaochen Zhu, and Zhenzhong Chen. Micro-video popularity prediction via multimodal variational information bottleneck. *IEEE Transactions on Multimedia*, 25:24–37, 2021.

Kele Xu, Zhimin Lin, Jianqiao Zhao, Peicang Shi, Wei Deng, and Huaimin Wang. Multimodal deep learning for social media popularity prediction with attention mechanism. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 4580–4584, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3416274. URL https://doi.org/10.1145/3394171.3416274.

Qianyun Yang, Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, and Liqiang Nie. STABLE: Efficient hybrid nearest neighbor search via magnitude-uniformity and cardinality-robustness. *arXiv preprint arXiv:2604.01617*, 2026.

Liliang Ye, Yunyao Zhang, Yafeng Wu, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. Mvp: Winning solution to smp challenge 2025 video track. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 14079–14085, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3763761. URL https://doi.org/10.1145/3746027.3763761.

Zhuoran Zhang, Shibiao Xu, Li Guo, and Wenke Lian. Multi-modal variational auto-encoder model for micro-video popularity prediction. In *Proceedings of the 8th International Conference on Communication and Information Processing*, ICCIP '22, pp. 9–16, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450397100. doi: 10.1145/3571662.3571664. URL https://doi.org/10.1145/3571662.3571664.

Ting Zhong, Jian Lang, Yifan Zhang, Zhangtao Cheng, Kunpeng Zhang, and Fan Zhou. Predicting micro-video popularity via multi-modal retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 2579–2583, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657929. URL https://doi.org/10.1145/3626772.3657929.

APPENDIX

A Implementation and Protocol Details

This appendix specifies the prediction-time information, dataset protocols, cross-fitting construction, and retrieval constraints used by SSPM. Throughout, *OOF residual* denotes the held-out prefix error used as the next stage's unresolved target.

A.1 Scale-Wise Information Assignment

Table 3 summarizes each stage's scope and target. SSPM follows one prediction pathway over three informative scales; intermediate prefixes are retained only for attribution.

Table 3: Prediction-time information assignment and training targets across scales.

Scale	Assigned information	Scope	Training target
Context Prior	Calendar, population, topic, or community statistics available before current-item observation	Population context	y_i (coarse baseline)
Multimodal Evidence	Current video, text, and prediction-time item or author metadata	Current item	$r^{(1)}_{i,\mathrm{OOF}}$
Historical Analogy	Admissible historical representations and the OOF residual bank	Local historical neighborhood	$r^{(2)}_{i,\mathrm{OOF}}$

A.2 Dataset Protocols

All partitions preserve temporal order: SMPD and Informs use 8:1:1 train/validation/test splits, while MicroLens uses 7:1:2.

SMPD. Context Prior Modeling uses holiday, time-of-day, and population-shared variables available before content observation. Multimodal Evidence Refinement combines the 175-dimensional base representation, UAE-Large-V1 text projected to 128 dimensions, static metadata, and 34 query-side prior-history summaries. All are prediction-time, pre-publication aggregates without sample-level or target-derived values. Historical Analogy Refinement aggregates cross-fitted residuals from designated benchmark references; target-derived query variables are excluded.

Informs-Strict. Context Prior Modeling uses chronological history and train-only target normalization. Multimodal Evidence Refinement uses visual features with train-only PCA64 and Ridge regression. Historical Analogy Refinement uses raw-CLIP cosine retrieval, $K = 20$, and $\alpha = 0.35$. Cross-fitting is forward-blocked, and every label-bearing reference precedes its query.

MicroLens. With no query-varying pre-item context, Context Prior Modeling reduces to the training-population prior. Multimodal Evidence Refinement uses 32 CLIP ViT-L/14 frames, mean 768-dimensional features, train-only PCA128, seven metadata features, and a three-seed ExtraTrees ensemble. Historical Analogy Refinement uses $0.9s_{\mathrm{CLIP}} + 0.1s_{\mathrm{meta}}$, $K = 40$, temperature 0.03, $\alpha = 0.5$, and Q01–Q99 clipping.

A.3 Cross-Fitting Construction

Cross-fitting is not merely leakage control: it constructs the out-of-sample unresolved target for each refinement. For outer holdout V_b, target-dependent transformations and T are cross-fitted within admissible training data $\mathcal{D}_b$ to form $r^{(1)}_{j,\mathrm{OOF}|b}$. The refiner G_b is fitted only to these targets, and $T_b + G_b$ is evaluated on V_b to produce $r^{(2)}_{i,\mathrm{OOF}}$; no label in V_b enters G_b's training targets. Final

training cross-fits T over the complete partition, refits T, and fits G to the resulting $r^{(1)}_{i,\mathrm{OOF}}$. The bank retains $r^{(2)}_{i,\mathrm{OOF}}$, the held-out error of the complete Context-to-Multimodal prefix.

A.4 Historical Retrieval Constraints and Protocol Scope

Under query-level temporal protocols, reference i is eligible only when $t_i < t_q$; future instances are excluded from retrieval and residual aggregation. When query-level timestamps are unavailable, candidates are restricted to the chronologically earlier training partition, and validation or test items never enter the reference bank. Informs-Strict additionally uses forward-only fitting and retrieval. MicroLens predictions were fixed before test-label access, and the audit found no future-label use. Two historical Informs official-test labels had been manually inspected before evaluation; STAP-Compatible remains a separate all-split numerical reference. Scale contributions are predictive risk differences under the declared filtration, not causal effects or mutual information. Additional scales require a predeclared sigma-field, OOF residual, and held-out prefix evaluation.

B Additional Mathematical Details

For the conditional projection argument in Section 3, the tower property gives $\mathbb{E}[m^*_\ell \mid \mathcal{F}_{\ell-1}] = m^*_{\ell-1}$, and hence $\mathbb{E}[\delta^*_\ell \mid \mathcal{F}_{\ell-1}] = 0$. For $j < \ell$, δ^*_j is $\mathcal{F}_{\ell-1}$-measurable, so

$$\mathbb{E}[\delta^*_j \delta^*_\ell] = \mathbb{E}[\delta^*_j \mathbb{E}[\delta^*_\ell \mid \mathcal{F}_{\ell-1}]] = 0.$$

The risk identity follows from the orthogonal expansion in Eq. 2 and the standard L_2 projection identity.

C Additional Analysis Details

These diagnostics and matched controls quantify the held-out prefixes and design principles in Section 5.

C.1 Local Historical Validity

MicroLens test gain. The prediction configuration was fixed before test labels were read. On 3,948 test items, Multimodal-prefix MSE is 1.609849 and Historical-prefix MSE is 1.573943, giving $\widehat{A}_2 = 0.035905$ with paired-bootstrap 95% CI $[0.022339, 0.049733]$. Historical refinement improves squared error for 2,106 samples and worsens it for 1,842, yielding a positive aggregate contribution.

Informs similarity-conditioned gain. Table 4 reports Historical contribution as a function of the most similar strictly past reference. Similarity is one minus cosine distance; intervals use 5,000 bootstrap resamples of the 318 validation queries.

Table 4: Informs Historical contribution by strictly past neighbor similarity.

Similarity bin	$\widehat{A}_2$	95% CI	Queries
Q1	0.000668	$[-0.000033,\ 0.001454]$	64
Q2	0.000353	$[-0.000365,\ 0.000988]$	63
Q3	0.000054	$[-0.000661,\ 0.000812]$	64
Q4	-0.000347	$[-0.000977,\ 0.000265]$	63
Q5	**0.001819**	$[\mathbf{0.000507},\ \mathbf{0.003686}]$	64
Overall	0.000513	$[0.000096,\ 0.000980]$	318

C.2 Residual Recurrence

For Informs, 318 validation queries are paired with 6,360 strictly past OOF references. Table 5 reports residual-sign agreement. Its non-monotone pattern and Q5 value of 0.4748 show that simple sign clustering does not account for Table 4's gain.

Table 5: Residual-sign agreement by pairwise similarity on Informs.

Similarity bin	Sign agreement	Pairs
Q1	0.5173	1,272
Q2	0.5472	1,272
Q3	**0.5676**	1,272
Q4	0.5204	1,272
Q5	0.4748	1,272

C.3 Module Ablation

Table 6 removes SMPD feature modules one at a time under the same endpoint evaluation. The matched controls in Appendix C.4 instead alter residual targets and memory construction.

Table 6: SMPD module leave-one-out results on the test set (15 seeds).

Variant	MAE ↓	nMSE ↓	SRC ↑
Full	1.4555	0.4901	0.6690
w/o Text	1.5023	0.5173	0.6443
w/o User	1.4632	0.4895	0.6690
w/o prior-history summaries	1.4943	0.4846	0.6782
w/o Time features	1.4677	0.4887	0.6681
Base only	1.4851	0.4938	0.6643
Full + Historical Analogy	**1.4449**	0.4870	0.6744

C.4 Matched Design-Principle Variants

Let $m_q^{(1)} = m_q^{(0)} + G_r(C_q, X_q)$ denote the residual-trained Multimodal prefix, $G_y(C_q, X_q)$ an equally sized full-target Multimodal predictor, and $\mathcal{K}_q(v_i) = \sum_i w_{qi} v_i$ the fixed retrieval operator. The matched controls are

$$\begin{aligned}
\widehat{y}_q^{\text{one}} &= F(C_q, X_q, H_q), \\
\widehat{y}_q^{\text{full}} &= G_y(C_q, X_q) + \alpha \mathcal{K}_q\Big(y_i - \widehat{G}_{y,\text{OOF}}(C_i, X_i)\Big), \\
\widehat{y}_q^{\text{label}} &= (1-\alpha) m_q^{(1)} + \alpha \mathcal{K}_q(y_i), \\
\widehat{y}_q^{\text{in}} &= m_q^{(1)} + \alpha \mathcal{K}_q\Big(y_i - \widehat{m}_{i,\text{fit}}^{(1)}\Big), \\
\widehat{y}_q^{\text{SSPM}} &= m_q^{(1)} + \alpha \mathcal{K}_q\Big(y_i - \widehat{m}_{i,\text{OOF}}^{(1)}\Big).
\end{aligned}$$

The variants share inputs, capacity, retrieval representations, neighbors, and tuning budgets, isolating target, aggregation, and memory choices.

C.5 Controlled Synthetic Verification

We generate $Y = f_C(C) + f_I(X) + f_H(Z) + \epsilon$ with independent component variances $0.30/0.50/0.20$ and noise variance 0.10. Context and Multimodal stages use linear estimators; Historical refinement applies K-nearest-neighbor smoothing to five-fold cross-fitted residuals. We use 6,000 training and 2,000 test samples for each of ten seeds.

Table 7: Synthetic attribution recovery (ten-seed mean ± standard deviation); SSPM matches every oracle share within 0.13 points and remains near the one-shot endpoint.

Model	Context share	Multimodal share	Historical share	Endpoint MSE
Empirical oracle	$30.19 \pm 1.62\%$	$49.76 \pm 1.72\%$	$20.05 \pm 0.86\%$	–
SSPM	$30.23 \pm 1.59\%$	$49.85 \pm 1.71\%$	$19.92 \pm 0.82\%$	0.1007 ± 0.0039
One-shot oracle basis	–	–	–	0.0987 ± 0.0037