

Efficient Multimodal Generative Recommendation with Latent Narrative Reasoning

Chenxing Wang^{1*}, Nantao Zheng^{1*}, Hao Miao², Juyuan Wang¹, Xinke Jiang,
Yuchen Fang^{1†}, Aolin Li¹, Haijun Wu¹

¹Weixin Group, Tencent

²The Hong Kong Polytechnic University

magnuswang@tencent.com, tmaczheng@tencent.com, hao-miao@outlook.com, jovwang@tencent.com,
thinkerjiang@foxmail.com, fyclmiss@gmail.com, churenli@tencent.com, ethanhjwu@tencent.com

Abstract

Generative recommendation reformulates item prediction as semantic identifier generation, yet episodic content introduces a fundamentally different setting where the target is determined by narrative evolution rather than user preference. This task requires models to understand multimodal storyline progression while addressing the efficiency challenges caused by redundant visual contexts and costly explicit reasoning generation. We propose **NarraLite**, an efficient multimodal generative recommendation framework that jointly compresses perception and reasoning. Specifically, Progressive Spectral Compression selectively distills long visual contexts into compact narrative-relevant evidence, preserving transition-critical information while reducing redundant visual computation. Latent Narrative Reasoning introduces context-routed latent reasoning tokens and aligns their contextualized representations with future continuation semantics, enabling implicit narrative inference without autoregressively decoding textual rationales. We further establish a user-agnostic multimodal benchmark for short-form drama continuation across UGC, PGC, and OOD settings. Extensive experiments demonstrate that NarraLite consistently improves continuation accuracy, narrative coherence, and robustness over existing approaches, while achieving a favorable accuracy–efficiency trade-off.

1 Introduction

Generative recommendation reformulates item prediction as sequence generation by representing items with discrete semantic identifiers (SIDs) and directly generating target identifiers (Li et al. 2024; Wang et al. 2024). However, existing methods mainly focus on user preference modeling, while episodic content requires predicting the segment that naturally follows an ongoing storyline. As shown in Figure 1, traditional recommendation retrieves content based on user interests (Chen et al. 2025; Deng et al. 2025), whereas narrative-aware recommendation requires understanding intrinsic event transitions and predicting semantically and temporally coherent continuations. This task further demands

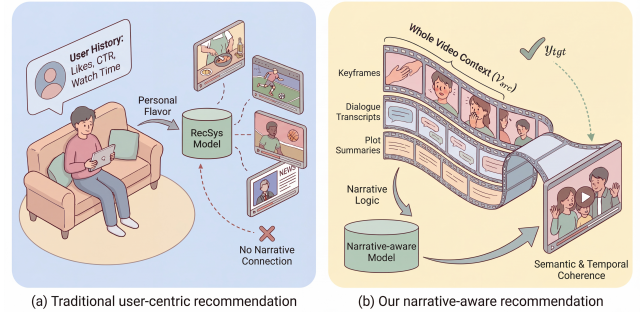


Figure 1: Paradigm shift from user-centric recommendation to narrative-aware recommendation.

joint modeling of visual plot evidence and textual descriptions, as either modality alone may miss critical narrative cues.

Applying multimodal generative models to this setting, however, faces two critical efficiency challenges (Zhang et al. 2025a; Cao et al. 2026; He et al. 2024; Zhan et al. 2026). First, long visual contexts contain substantial redundancy: consecutive frames often repeat similar scenes, characters, and backgrounds, while only a small subset of events is essential for determining future storyline transitions. Processing all visual tokens therefore introduces considerable computational overhead, whereas aggressive compression may remove subtle but decisive narrative cues. Second, explicit chain-of-thought (CoT) reasoning (Yue et al. 2025a) introduces additional autoregressive decoding costs by requiring the model to generate lengthy intermediate rationales before predicting the final SID. Although such reasoning can improve structured inference, the generated explanations are unnecessary for recommendation and increase latency. Efficient narrative recommendation therefore requires retaining transition-critical visual evidence while enabling structured reasoning without explicitly generating reasoning chains.

To address these challenges, we propose **NarraLite**, an efficient multimodal generative recommendation framework that jointly compresses perception and reasoning. NarraLite introduces two complementary components. **Progressive**

*These authors contributed equally.

†Corresponding author.

Spectral Compression (PSC) selectively distills long visual contexts into compact narrative-relevant tokens. Instead of uniformly processing all visual observations, PSC identifies visual evidence that provides complementary information to textual plot descriptions, reducing redundant multimodal computation while preserving important storyline cues. **Latent Narrative Reasoning (LNR)** models the implicit transition from the observed narrative state to its continuation through compact latent reasoning states. Rather than autoregressively generating textual rationales, LNR performs internal narrative inference and directly conditions SID generation, reducing reasoning overhead while maintaining structured prediction capability.

To facilitate evaluation of this task, we introduce a user-agnostic multimodal short-form drama continuation benchmark with aligned story contexts and ground-truth successor segments across UGC, PGC, and OOD settings. Extensive experiments demonstrate that NarraLite achieves superior continuation accuracy and a more favorable accuracy–efficiency trade-off compared with retrieval-based, sequential, multimodal, and generative recommendation baselines.

Our contributions are summarized as follows:

- We formulate efficient multimodal narrative-aware generative recommendation, where the goal is to predict a coherent story continuation rather than user-preference-driven item relevance, and identify two key efficiency challenges caused by redundant visual perception and explicit reasoning generation.
- We propose NarraLite, which introduces Progressive Spectral Compression for retaining compact and text-complementary visual evidence, and Latent Narrative Reasoning for modeling implicit story transitions without generating explicit reasoning chains.
- We establish a user-agnostic multimodal continuation benchmark and extensive experiments demonstrate that NarraLite consistently improves recommendation accuracy, narrative coherence, and inference efficiency.

2 Related Work

2.1 Multimodal Recommendation

Multimodal recommendation has evolved from feature-level fusion toward adaptive alignment and foundation-model-based representation learning. Early studies focus on integrating heterogeneous item modalities: MTSTRec (Hong et al. 2025) introduces modality-shared tokens to coordinate temporal textual, visual, and price information, while FindRec (Wang et al. 2025) further improves multimodal interaction modeling through adaptive routing and efficient sequence modeling. Recent works study robustness and efficiency under imperfect multimodal inputs. MoDiCF (Yang et al. 2026) and SSR (Li et al. 2025) respectively explore generative modality recovery and frequency-aware multimodal denoising to improve representation quality. With the emergence of multimodal LLM, LaViC (Jeon et al. 2025) and PRIME (Yue et al. 2025b) further investigate visual token compression and LMM-based recommendation pipelines. Despite their progress, existing multimodal recommenders

mainly target user preference modeling and item matching. Their multimodal representations are optimized for identifying relevant items rather than preserving sparse visual evidence and inferring cross-segment event transitions.

2.2 CoT Reasoning for Recommendation

Recent studies introduce reasoning mechanisms into recommendation by using large language models to improve preference understanding. RecZero (Kong et al. 2026) and R2ec (You et al. 2026) explore explicit reasoning chains with reinforcement learning or joint reasoning–prediction optimization, demonstrating the potential of structured reasoning for recommendation. However, explicit CoT introduces additional autoregressive decoding costs, and recent analyses such as MME-CoT (Jiang et al. 2025) reveal that excessive reasoning may bring unnecessary inference overhead in multimodal scenarios. To improve efficiency, subsequent works attempt to internalize reasoning into latent representations. SIREN (Ding et al. 2026) distills explicit preference reasoning into hidden states, while LatentR3 (Zhang et al. 2025b) and IntuRec (Liu et al. 2026) further explore latent reasoning tokens for direct recommendation generation. Nevertheless, existing reasoning-based recommenders primarily model latent user preferences from interaction histories or candidate items. They neither ground latent states in multimodal narrative evidence nor capture the transition from observed events to future story states.

3 Preliminaries

3.1 Problem Formulation

Let $\mathcal{Y} = \{y_j\}_{j=1}^{|\mathcal{Y}|}$ denote the candidate video segments. We define the training set as narrative continuation pairs:

$$\mathcal{D} = \{(C_i, y_i^+)\}_{i=1}^N, \quad y_i^+ \in \mathcal{Y}, \quad (1)$$

where C_i denotes the observed narrative context and y_i^+ is the ground-truth continuation. Each context consists of temporally ordered multimodal segments:

$$C_i = \{(\mathcal{F}_{i,s}, \mathcal{X}_{i,s})\}_{s=1}^{S_i}, \quad (2)$$

where $\mathcal{F}_{i,s}$ and $\mathcal{X}_{i,s}$ represent visual keyframes and textual information, respectively. Different from conventional recommendation, the target segment is determined by intrinsic storyline progression rather than user preference.

We formulate continuation prediction as semantic identifier generation. A fixed multimodal encoder $\phi(\cdot)$ and residual quantizer $\mathcal{Q}(\cdot)$ encode each candidate segment as:

$$\text{SID}(y_j) = \mathcal{Q}(\phi(y_j)) = [c_j^1, \dots, c_j^L]. \quad (3)$$

The model autoregressively generates the SID of the ground-truth continuation:

$$p_{\Theta}(\text{SID}(y_i^+) | C_i) = \prod_{\ell=1}^L p_{\Theta}(c_i^{\ell} | C_i, c_i^{<\ell}). \quad (4)$$

During inference, trie-constrained decoding produces valid SIDs, which are deterministically mapped back to their corresponding segments.

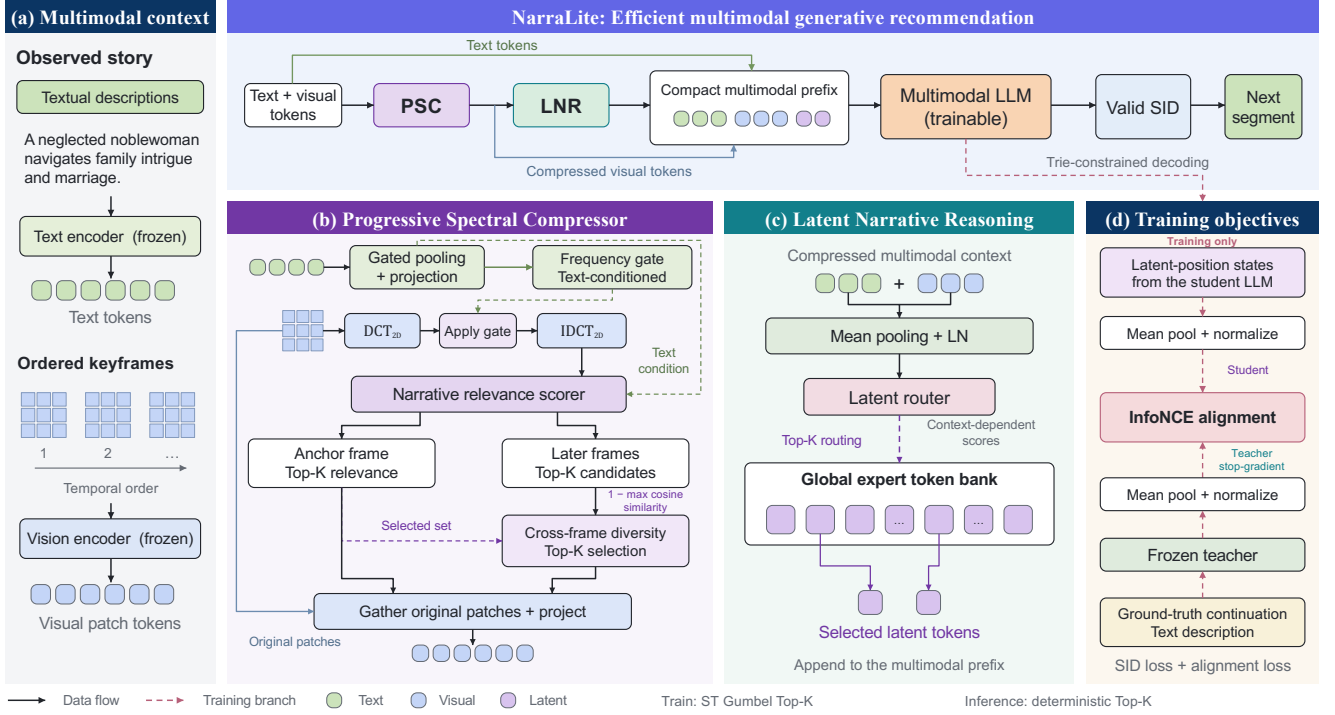


Figure 2: Overview of NarraLite.

4 Method

4.1 Overview

We propose **NarraLite**, an efficient multimodal generative recommendation framework for narrative continuation. As illustrated in Figure 2, NarraLite consists of two major components: **Progressive Spectral Compressor (PSC)** and **Latent Narrative Reasoning (LNR)**.

PSC aims to reduce redundant visual computation while preserving transition-critical visual evidence. Instead of uniformly sampling or compressing visual tokens through soft aggregation, PSC performs text-conditioned spectral modulation followed by relevance-aware and diversity-aware token selection. LNR introduces a mixture-of-expert latent token bank to capture future narrative tendencies without explicitly generating textual reasoning chains.

Given the narrative context $\mathcal{C}_i$, we first aggregate all visual and textual observations:

$$\begin{aligned}\mathcal{X}_i &= \text{Concat}(\mathcal{X}_{i,1}, \dots, \mathcal{X}_{i,S_i}), \\ \mathcal{F}_i &= \text{Concat}(\mathcal{F}_{i,1}, \dots, \mathcal{F}_{i,S_i}).\end{aligned}\quad (5)$$

A frozen text encoder and vision encoder produce:

$$\begin{aligned}\mathbf{T}_i &= f_{\text{txt}}(\mathcal{X}_i) \in \mathbb{R}^{M_i \times d}, \\ \mathbf{V}_i &= f_{\text{vis}}(\mathcal{F}_i) \in \mathbb{R}^{N_i \times H \times W \times d_v},\end{aligned}\quad (6)$$

where the visual context contains $P_i = N_i H W$ patch tokens.

PSC compresses visual features into K_v tokens:

$$\mathbf{V}_i^c = \text{PSC}(\mathbf{V}_i, \mathbf{T}_i) \in \mathbb{R}^{K_v \times d}, \quad K_v \ll P_i. \quad (7)$$

The compressed visual tokens and textual tokens are further processed by LNR to obtain K_r latent narrative tokens:

$$\mathbf{R}_i = \text{LNR}(\mathbf{T}_i, \mathbf{V}_i^c) \in \mathbb{R}^{K_r \times d}. \quad (8)$$

Finally, the multimodal prefix is constructed as:

$$\mathbf{Z}_i = [\mathbf{T}_i; \mathbf{V}_i^c; \mathbf{R}_i], \quad (9)$$

which is fed into the multimodal LLM for SID generation.

4.2 Progressive Spectral Compressor

Long video contexts contain redundant spatial and temporal observations, while directly removing tokens may discard subtle but important events. PSC therefore performs narrative-aware visual compression through three steps: text-conditioned spectral modulation, patch relevance scoring, and cross-frame diversity selection.

Text-Conditioned Spectral Modulation We first obtain a narrative condition from textual tokens. For each text token $\mathbf{t}_{i,m}$:

$$a_{i,m} = \sigma(\mathbf{w}_p^\top \mathbf{t}_{i,m} + b_p), \quad (10)$$

and compute gated pooling:

$$\bar{\mathbf{t}}_i = \frac{\sum_m a_{i,m} \mathbf{t}_{i,m}}{\sum_m a_{i,m} + \epsilon}. \quad (11)$$

The textual condition is projected into the visual space:

$$\mathbf{q}_i = \text{LN}(\mathbf{W}_t \bar{\mathbf{t}}_i + \mathbf{b}_t). \quad (12)$$

For each frame, PSC transforms visual patches into frequency space:

$$\hat{\mathbf{V}}_{i,n} = \text{DCT}_{2D}(\mathbf{V}_{i,n}). \quad (13)$$

A sample-specific frequency gate is generated:

$$\mathbf{G}_i = \sigma(\text{MLP}_f(\mathbf{q}_i)), \quad (14)$$

and applied before inverse transformation:

$$\tilde{\mathbf{V}}_{i,n} = \text{IDCT}_{2D}(\mathbf{G}_i \odot \hat{\mathbf{V}}_{i,n}). \quad (15)$$

This operation allows textual narrative information to adaptively preserve either coarse scene structure or fine-grained visual details.

Narrative-Aware Token Selection After spectral modulation, PSC assigns each patch a relevance score:

$$s_{i,n,p} = \mathbf{w}_s^\top \text{GELU}(\mathbf{W}_s(\tilde{\mathbf{v}}_{i,n,p} \odot \mathbf{q}_i) + \mathbf{b}_s). \quad (16)$$

For the anchor frame, PSC directly selects the top- K_a patches:

$$\mathcal{I}_{i,1} = \text{STGumbelTopK}(\mathbf{s}_{i,1}, K_a). \quad (17)$$

For subsequent frames, PSC first constructs a candidate pool:

$$\mathcal{P}_{i,n} = \text{STGumbelTopK}(\mathbf{s}_{i,n}, K_c), \quad K_c > K_f. \quad (18)$$

To avoid redundant selections across frames, we compute diversity scores:

$$d_{i,n,p} = 1 - \max_{q \in \mathcal{I}_i} \cos(\mathbf{v}_{i,n,p}, \mathbf{v}_{i,q}). \quad (19)$$

The final tokens are selected according to:

$$\mathcal{I}_{i,n} = \text{STGumbelTopK}(\{d_{i,n,p}\}, K_f). \quad (20)$$

The selected complete patch embeddings are concatenated and projected:

$$\mathbf{V}_i^c = \text{Proj}_v(\text{Select}(\mathbf{V}_i)) \in \mathbb{R}^{K_v \times d}. \quad (21)$$

4.3 Latent Narrative Reasoning

The compressed multimodal context provides evidence of the observed storyline, but the model still needs to infer the latent direction toward the future continuation. Explicit chain-of-thought reasoning introduces additional autoregressive cost, while directly generating context dependent latent tokens may simply duplicate information already captured by the multimodal LLM.

We therefore introduce a mixture-of-expert latent token bank, where the semantic space of future narrative tendencies is learned globally and the router dynamically activates suitable latent tokens for each instance.

Context-Routed Expert Token Bank We first summarize the observed narrative context:

$$\mathbf{h}_i^{ctx} = \text{LN}(\text{MeanPool}([\mathbf{T}_i; \mathbf{V}_i^c])) \in \mathbb{R}^d. \quad (22)$$

A lightweight router predicts the relevance of each expert:

$$\rho_i = \mathbf{W}_r \mathbf{h}_i^{ctx} + \mathbf{b}_r \in \mathbb{R}^{M_e}, \quad (23)$$

where M_e is the number of latent experts.

The expert token bank is defined as:

$$\mathbf{E} = [\boldsymbol{\theta}_1; \boldsymbol{\theta}_2; \dots; \boldsymbol{\theta}_{M_e}] \in \mathbb{R}^{M_e \times d}, \quad (24)$$

where each $\boldsymbol{\theta}_j$ is a learnable continuous token. Unlike context-dependent latent generation, the expert tokens are independent of the current input, while the router determines which future narrative patterns should be activated.

During training, we select K_r experts with straight-through Gumbel Top- K routing:

$$\mathbf{A}_i^r = \text{STGumbelTopK}(\rho_i, K_r), \quad (25)$$

and obtain the latent narrative tokens:

$$\mathbf{R}_i = \mathbf{A}_i^r \mathbf{E} \in \mathbb{R}^{K_r \times d}. \quad (26)$$

At inference time, the stochastic routing is replaced with deterministic Top- K_r selection. The selected latent tokens are appended after the multimodal observations:

$$\mathbf{Z}_i = [\mathbf{T}_i; \mathbf{V}_i^c; \mathbf{R}_i]. \quad (27)$$

The multimodal LLM contextualizes these latent tokens together with the observed narrative context, allowing them to encode sample-specific future transition information.

Target-Guided Latent Alignment The SID generation objective alone does not explicitly enforce that the latent states represent the semantic direction of the future continuation. We therefore introduce a training-only alignment objective between latent hidden states and the ground-truth continuation semantics.

During training, the multimodal LLM processes the compact prefix together with the target SID:

$$\mathbf{H}_i^{stu} = \mathcal{M}_\Theta([\mathbf{Z}_i; \text{Emb}_\Theta(\text{SID}(y_i^+))]). \quad (28)$$

We extract the final hidden states corresponding to the latent positions:

$$\mathbf{H}_i^r = [\mathbf{h}_{i,1}^r; \dots; \mathbf{h}_{i,K_r}^r]. \quad (29)$$

The student latent representation is obtained by:

$$\mathbf{z}_i^s = \text{Normalize}(\text{MeanPool}(\mathbf{H}_i^r)). \quad (30)$$

Meanwhile, a frozen teacher model encodes the textual description of the ground-truth continuation:

$$\mathbf{H}_i^t = \overline{\mathcal{M}}(\mathcal{X}_i^+), \quad (31)$$

and produces:

$$\mathbf{z}_i^t = \text{Normalize}(\text{stopgrad}(\text{MeanPool}(\mathbf{H}_i^t))). \quad (32)$$

Given a mini-batch of size B , we optimize an InfoNCE objective:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp((\mathbf{z}_i^s)^\top \mathbf{z}_i^t / \tau_c)}{\sum_{j=1}^B \exp((\mathbf{z}_i^s)^\top \mathbf{z}_j^t / \tau_c)}, \quad (33)$$

where τ_c is the temperature coefficient.

This alignment encourages latent tokens to capture reusable future narrative directions without requiring explicit reasoning annotations.

4.4 Generative Learning and Inference

Given the compact multimodal prefix $\mathbf{Z}_i$, the multimodal LLM generates the semantic identifier of the ground-truth continuation:

$$p_{\Theta}(\text{SID}(y_i^+) | \mathcal{C}_i) = \prod_{\ell=1}^L p_{\Theta}(c_i^{\ell} | \mathbf{Z}_i, c_i^{<\ell}). \quad (34)$$

The SID generation loss is:

$$\mathcal{L}_{\text{SID}} = -\frac{1}{B} \sum_{i=1}^B \sum_{\ell=1}^L \log p_{\Theta}(c_i^{\ell} | \mathbf{Z}_i, c_i^{<\ell}). \quad (35)$$

The overall training objective combines SID generation and latent narrative alignment:

$$\mathcal{L} = \mathcal{L}_{\text{SID}} + \lambda \mathcal{L}_{\text{NCE}}, \quad (36)$$

where λ controls the contribution of latent supervision.

During inference, the teacher branch and alignment objective are removed. PSC and LNR use deterministic Top- K selection, and the multimodal LLM generates valid SIDs through trie-constrained decoding:

$$\widehat{\text{SID}}_i = \arg \max_{c \in \mathcal{S}_{\text{SID}}} p_{\Theta}(c | \mathbf{Z}_i). \quad (37)$$

The predicted identifier is finally mapped back to its corresponding continuation segment:

$$\hat{y}_i = \text{SID}^{-1}(\widehat{\text{SID}}_i). \quad (38)$$

Compared with processing all visual tokens, NarraLite only introduces $M_i + K_v + K_r$ prefix tokens instead of $M_i + P_i$. Moreover, latent tokens are processed in parallel during LLM inference rather than being autoregressively generated as explicit reasoning chains, reducing both visual computation and reasoning latency.

Table 1: The overall statistics of our dataset.

Split	Train	Eval	UGC	PGC	OOD
Pairs	241,545	14,231	11,175	1,613	1,443
Texts	268	247	294	76	78
Duration	474	431	533	58	60
Segments	8.7	8.1	9.6	2.6	2.7
Continuations	2.2	2.1	2.4	1	1

5 Experimental Setup

5.1 Dataset

We evaluate our method on a user-agnostic multimodal narrative continuation benchmark that we construct for short-form dramas. In this benchmark, each sample contains ordered visual frames, textual descriptions, and the corresponding continuation segment, enabling evaluation of narrative-aware recommendation without user preference bias. As shown in Table 1, our dataset comprises over 250,000 video pairs across five splits: Train, Eval, and three diverse test sets designed to evaluate narrative generalization: User-Generated Content (UGC), Professional Generated Content (PGC), and

Out-of-Distribution (OOD). The OOD split isolates the concluding episodes of PGC dramas to test model performance on unseen narrative endings. While PGC and OOD segments exhibit a concise, linear progression—averaging 60 seconds with unique successors—the UGC subset introduces significant complexity through higher narrative density (294 works on average) and greater continuation multiplicity (avg. 2.4). Details of the benchmark can be found in Appendix A.

5.2 Metrics

We evaluate exact continuation retrieval using standard H@K, which measures whether the ground-truth segment y_i^+ appears in the top- K generated candidates.

To evaluate whether the model captures intrinsic narrative coherence rather than relying on superficial visual similarity, we introduce **Overlap-Penalized Semantic Coherence (OPSC)**. Given the top- K predicted segments $\{\hat{y}_{i,1}, \dots, \hat{y}_{i,K}\}$ and the ground-truth continuation y_i^+ , OPSC measures their semantic alignment while penalizing excessive visual overlap:

$$\text{OPSC} = \frac{1}{K} \sum_{k=1}^K \text{sim}(\phi(\hat{y}_{i,k}), \phi(y_i^+)) \cdot \exp(-\beta R(\hat{y}_{i,k}, y_i^+)), \quad (39)$$

where $\phi(\cdot)$ denotes the multimodal segment encoder defined in Section 3.1, $\text{sim}(\cdot, \cdot)$ measures semantic similarity, and $R(\hat{y}_{i,k}, y_i^+)$ represents the visual overlap ratio between the predicted and ground-truth segments. The hyperparameter β controls the strength of the overlap penalty. Higher OPSC indicates that the predicted continuation is semantically consistent with the ground truth while avoiding trivial visual duplication. Together, H@K and OPSC evaluate both exact continuation accuracy and intrinsic narrative progression.

5.3 Baselines

We compare NarraLite with representative methods from three categories. **Retrieval-based methods:** including BM25 and Vector Retrieval, which measure continuation relevance based on textual or multimodal semantic similarity. **Sequential recommendation models:** including SAS-Rec (Kang and McAuley 2018), GRU4Rec (Shehzad and Jannach 2025), and BERT4Rec (Sun et al. 2019), which model sequential dependencies from historical interactions and are adapted to the narrative continuation setting. **Generative recommendation models:** including TIGER (Rajput et al. 2023), LETTER (Wang et al. 2024), EAREC (Yang et al. 2025), and LatentR3 (Zhang et al. 2025b). TIGER and LETTER represent the standard paradigm of generating discrete semantic identifiers for target items, while EAREC and LatentR3 further introduces visual and latent reasoning tokens to improve recommendation generation without explicit reasoning chains. The details can be found in Appendix B.

5.4 Implementation Details

We implement NarraLite based on the pretrained multimodal large language model Qwen-3.5-2B (Qwen Team 2026). Following prior generative recommendation frameworks, we construct semantic identifiers using RQ-KMeans (Xu

Table 2: Performance comparison on narrative-aware recommendation. The best and second-best results among non-ablation methods are highlighted in bold and underline, respectively.

Method	Eval				UGC				PGC				OOD			
	H@1	H@3	H@5	H@10	H@1	H@3	H@5	H@10	H@1	H@3	H@5	H@10	H@1	H@3	H@5	H@10
<i>Retrieval</i>																
BM25	0.0715	0.1285	0.1585	0.2022	0.0827	0.1429	0.1754	0.2241	0.0254	0.0468	0.0561	0.0827	0.0173	0.0335	0.0462	0.0618
Dense	0.0860	0.1655	0.2053	0.2569	0.0879	0.1694	0.2107	0.2638	0.0704	0.1287	0.1621	0.2186	0.0788	0.1475	0.1812	0.2127
<i>Sequential</i>																
SASRec	0.0595	0.0984	0.1153	0.1382	0.0637	0.1041	0.1226	0.1472	0.0508	0.0869	0.1017	0.1261	0.0419	0.0718	0.0874	0.1085
GRU4Rec	0.0538	0.0874	0.1064	0.1312	0.0559	0.0918	0.1105	0.1351	0.0459	0.0754	0.0916	0.1167	0.0378	0.0662	0.0801	0.1024
BERT4Rec	0.0570	0.0987	0.1189	0.1405	0.0613	0.1016	0.1221	0.1463	0.0498	0.0847	0.1036	0.1288	0.0426	0.0729	0.0879	0.1079
<i>Multimodal</i>																
EARec	<u>0.1963</u>	0.2760	0.3103	0.3483	<u>0.2084</u>	0.2898	0.3275	0.3681	0.1715	0.2449	0.2804	0.3145	0.1368	0.1938	0.2257	0.2596
<i>Generative</i>																
TIGER	0.1888	0.2768	0.3108	0.3458	0.2019	0.2945	0.3338	0.3757	0.1654	0.2408	0.2795	0.3140	0.1286	0.1901	0.2244	0.2578
LETTER	0.1798	0.2689	0.3032	0.3379	0.1917	0.2817	0.3205	0.3619	0.1576	0.2314	0.2687	0.3019	0.1208	0.1815	0.2148	0.2487
LatentR3	0.1908	<u>0.2773</u>	<u>0.3155</u>	<u>0.3583</u>	0.2031	<u>0.3018</u>	<u>0.3407</u>	<u>0.3785</u>	<u>0.1752</u>	<u>0.2569</u>	<u>0.2975</u>	<u>0.3318</u>	<u>0.1497</u>	<u>0.2186</u>	<u>0.2574</u>	<u>0.2910</u>
Ours																
NarraLite	0.1988	0.3068	0.3480	0.3933	0.2175	0.3352	0.3810	0.4218	0.1861	0.2735	0.3192	0.3570	0.1598	0.2328	0.2724	0.3090
<i>Ablation</i>																
w/o LNR	0.2028	0.2915	0.3245	0.3625	0.2148	0.3198	0.3582	0.3950	0.1751	0.2540	0.2935	0.3250	0.1408	0.2036	0.2398	0.2700
w/o PSC	0.1985	0.2963	0.3373	0.3783	0.2042	0.3057	0.3458	0.3820	0.1787	0.2615	0.3004	0.3350	0.1531	0.2207	0.2591	0.2980

Table 3: Overlap-Penalized Semantic Coherence (OPSC) results on narrative-aware recommendation.

Method	Eval	UGC	PGC	OOD
<i>Multimodal</i>				
EARec	0.4392	0.4817	0.3716	0.3489
<i>Generative</i>				
TIGER	0.5286	0.5954	0.4768	0.4512
LETTER	0.5143	0.5819	0.4597	0.4338
LatentR3	<u>0.5587</u>	<u>0.6078</u>	<u>0.5162</u>	<u>0.5051</u>
Ours				
NarraLite	0.5904	0.6367	0.5579	0.5506

et al. 2025). Specifically, a three-level residual quantization scheme with codebook sizes of (512, 256, 256) is applied to multimodal segment representations, producing hierarchical SIDs for all candidate segments.

All experiments are conducted on 8 NVIDIA H20 GPUs. During inference, we employ trie-constrained beam search with 10 generated candidates for each query. Unless otherwise specified, NarraLite retains 30 compressed visual tokens ($K_v = 30$) and 2 latent narrative tokens ($K_r = 2$), which are selected according to the hyperparameter analysis.

6 Experimental Results

6.1 Main Results

Table 2 summarizes the overall performance comparison. We have the following observations.

Generative recommendation is superior to retrieval-based approaches. Generative models consistently outperform retrieval methods across all splits. For example, NarraLite achieves an H@10 of 0.3933 on Eval, significantly exceeding Dense Retrieval (0.2569), demonstrating that direct SID generation can capture high-level narrative transitions beyond surface-level similarity matching. This advantage is further maintained under OOD scenarios, where retrieval methods suffer from limited generalization due to their dependence on explicit semantic overlap.

Latent reasoning and multimodal evidence are complementary. LatentR3 achieves stronger performance than previous generative baselines, especially on OOD, validating the effectiveness of latent reasoning for modeling future narrative states. NarraLite further improves over LatentR3 by jointly leveraging compressed visual evidence and latent narrative reasoning, achieving the best performance across all evaluation settings. These results demonstrate that accurate continuation requires both multimodal understanding of observed events and implicit modeling of future transitions.

NarraLite achieves the best overall accuracy. NarraLite obtains the highest H@1, H@3, H@5, and H@10 on all splits, achieving an H@10 of 0.3933 on Eval and 0.3090 on OOD. The consistent improvements verify that Progressive Spectral Compression effectively preserves narrative-relevant visual information, while Latent Narrative Reasoning enables efficient prediction of future storyline states without explicit reasoning chains.

6.2 Narrative Coherence Analysis

Table 3 reports Overlap-Penalized Semantic Coherence (OPSC), which evaluates whether predicted segments remain

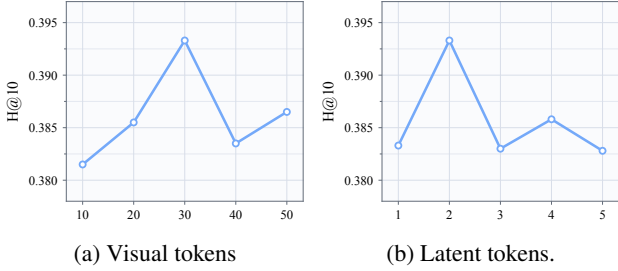


Figure 3: Hyperparameter study.

semantically consistent with the ground-truth continuation after penalizing excessive visual overlap. Generative recommenders consistently outperform the multimodal matching baseline, indicating that SID generation captures narrative relationships beyond static cross-modal similarity. Among the generative baselines, LatentR3 achieves stronger coherence, particularly under structured and out-of-distribution storylines, highlighting the benefit of latent reasoning for modeling future narrative states. NarraLite further obtains the best OPSC across all evaluation settings, suggesting that narrative-relevant visual compression and latent transition reasoning are complementary. Thus, even when the exact target is not ranked first, NarraLite tends to produce continuations that remain logically aligned with the intended storyline rather than relying on superficial visual repetition.

6.3 Ablation Study

We conduct ablation studies to investigate the contribution of the two key components in NarraLite: multimodal perception and latent narrative reasoning. As shown in Table 2, both variants consistently degrade performance across evaluation settings, confirming the necessity of each component.

Effect of latent narrative reasoning. Removing the latent reasoning module (w/o LCOT) leads to noticeable performance drops, particularly on the OOD split. This indicates that explicitly modeling the transition from observed events to future narrative states is important for generalizing beyond seen story patterns. Without latent reasoning, the model mainly relies on direct context matching, making it harder to infer implicit storyline progressions.

Effect of multimodal perception. Removing multimodal inputs (w/o MM) also decreases performance, especially on UGC scenarios where visual evidence is often essential for disambiguating characters, actions, and scene changes. This demonstrates that textual descriptions alone cannot fully capture the visual plot required for narrative continuation.

6.4 Hyperparameter Analysis

We analyze the sensitivity of NarraLite to the number of retained visual tokens K_v and latent reasoning tokens K_r on the Eval split, while keeping all other settings fixed. As shown in Figure 3, increasing K_v initially improves the performance, which peaks at $K_v = 30$. Retaining fewer tokens may discard plot-critical visual evidence, whereas using more tokens reintroduces redundant information and weakens the benefits of visual compression. For latent reasoning, the best result is

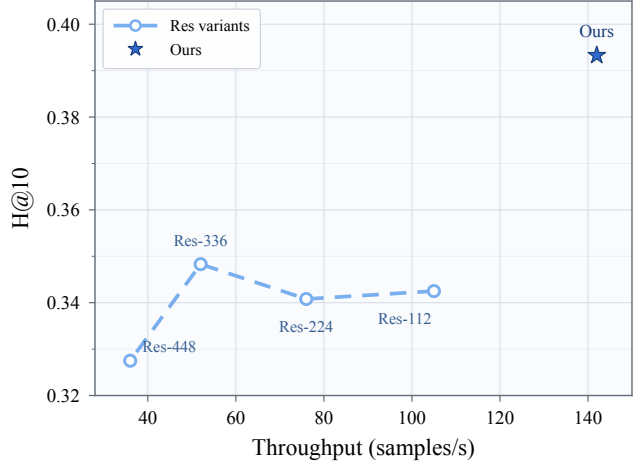


Figure 4: Accuracy-throughput trade-off on the Eval split.

obtained with $K_r = 2$. A single latent token provides limited capacity, while additional tokens may introduce redundant or conflicting reasoning signals.

6.5 Efficiency Analysis

Figure 4 compares the accuracy-throughput trade-off between NarraLite and full-frame variants using different input resolutions. Each *Res- x* variant retains all visual patches extracted from frames resized to resolution x . Reducing the resolution generally improves throughput but may remove fine-grained visual evidence, whereas increasing the resolution introduces substantially more visual tokens without consistently improving continuation accuracy. This indicates that global resolution reduction provides an inefficient trade-off between visual detail and computation.

In contrast, NarraLite lies in the upper-right region of the accuracy-throughput space and strictly dominates all resolution variants. Instead of uniformly degrading the visual input, Progressive Spectral Compression selectively preserves plot-relevant patches while discarding redundant content before multimodal language modeling. These results demonstrate that selective visual-token compression provides a more favorable efficiency-accuracy trade-off than processing complete frames at either low or high resolutions.

7 Conclusion

We studied efficient narrative-aware generative recommendation for episodic content and proposed **NarraLite**. Progressive Spectral Compression selectively preserves compact, text-complementary visual evidence, while Latent Narrative Reasoning models future storyline transitions through context-routed expert tokens without explicitly decoding reasoning chains. Together with a user-agnostic multimodal benchmark, our experiments demonstrate consistent improvements in continuation accuracy, out-of-distribution robustness, and inference efficiency. These results highlight the complementary importance of selective multimodal perception and latent transition reasoning for accurate and efficient episodic recommendation.

References

- Cao, J.; Zhang, Q.; Jia, P.; Zhao, X.; Lan, B.; Zhang, X.; Wei, X.; Chen, S.; Li, L.; Liu, X.; et al. 2026. Fastdrivevla: Efficient end-to-end driving via plug-and-play reconstruction-based token pruning. In *Proceedings of AAAI*, 2571–2579.
- Chen, B.; Guo, X.; Wang, S.; Liang, Z.; Lv, Y.; Ma, Y.; Xiao, X.; Xue, B.; Zhang, X.; Yang, Y.; et al. 2025. One-search: A preliminary exploration of the unified end-to-end generative framework for e-commerce search. *arXiv preprint arXiv:2509.03236*.
- Deng, J.; Wang, S.; Cai, K.; Ren, L.; Hu, Q.; Ding, W.; Luo, Q.; and Zhou, G. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*.
- Ding, Z.; Li, J.; Mu, S.; and Shi, J. 2026. Token-Efficient Long-Term Interest Sketching and Internalized Reasoning for LLM-based Recommendation. In *Proceedings of ICLR*.
- He, L.; Li, Z.; Cai, X.; and Wang, P. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of AAAI*, 18180–18187.
- Hong, M.-Y.; Hsu, Y.-J.; Chiang, M.-C.; and Lin, C. 2025. MTSTRec: multimodal time-aligned shared token recommender. In *Proceedings of ICML*.
- Jeon, H.; Koide, S.; Wang, Y.; He, Z.; and McAuley, J. 2025. Adapting Large Vision-Language Models to Visually-Aware Conversational Recommendation. In *Proceedings of SIGKDD*, 1037–1048.
- Jiang, D.; Zhang, R.; Guo, Z.; Li, Y.; Qi, Y.; Chen, X.; Wang, L.; Jin, J.; Guo, C.; Yan, S.; et al. 2025. MME-CoT: Benchmarking Chain-of-Thought in Large Multimodal Models for Reasoning Quality, Robustness, and Efficiency. In *Proceedings of ICML*, 27793–27830.
- Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *Proceedings of ICDM*, 197–206.
- Kong, X.; Jiang, J.; Liu, B.; Xu, Z.; Zhu, H.; Xu, J.; Zheng, B.; Wu, J.; and Wang, X. 2026. Think before Recommendation: Autonomous Reasoning-enhanced Recommender. In *Proceedings of NeurIPS*, 141209–141232.
- Li, J.; Wang, S.; Zhang, Q.; Yu, S.; and Chen, F. 2025. Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. In *Proceedings of WWW*, 2787–2798.
- Li, L.; Zhang, Y.; Liu, D.; and Chen, L. 2024. Large language models for generative recommendation: A survey and visionary discussions. In *Proceedings of LREC-COLING*, 10146–10159.
- Liu, C.; Bai, Y.; Zhao, X.; Zhang, Y.; Wang, Q.; Feng, F.; and Rong, W. 2026. Intuition-Guided Latent Reasoning for LLM-Based Recommendation. *arXiv preprint arXiv:2606.27684*.
- Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents. <https://qwen.ai/blog?id=qwen3.5>. Accessed: 2026-02-16.
- Rajput, S.; Mehta, N.; Singh, A.; Keshavan, R. H.; Vu, T.; Heldt, L.; Hong, L.; Tay, Y.; Tran, V. Q.; Samost, J.; et al. 2023. Recommender systems with generative retrieval. In *Proceedings of NeurIPS*.
- Shehzad, F.; and Jannach, D. 2025. Revisiting the Performance of Graph Neural Networks for Session-based Recommendation. In *Proceedings of RecSys*, 842–846.
- Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of CIKM*, 1441–1450.
- Wang, M.; Xiao, Y.; Wang, B.; Zhang, S.; Ye, S.; Wang, W.; Yin, H.; Guo, R.; and Xu, Z. 2025. FindRec: Stein-Guided Entropic Flow for Multi-Modal Sequential Recommendation. In *Proceedings of SIGKDD*, 3008–3018.
- Wang, W.; Bao, H.; Lin, X.; Zhang, J.; Li, Y.; Feng, F.; Ng, S.-K.; and Chua, T.-S. 2024. Learnable item tokenization for generative recommendation. In *Proceedings of CIKM*, 2400–2409.
- Xu, Y.; Zhang, M.; Li, C.; Liao, Z.; Xing, H.; Deng, H.; Hu, J.; Zhang, Y.; Zeng, X.; and Zhang, J. 2025. Mmq: Multi-modal mixture-of-quantization tokenization for semantic id generation and user behavioral adaptation. *arXiv preprint arXiv:2508.15281*.
- Yang, S.; Ma, W.; Guo, Z.; Zhang, M.; Wu, H.; Zhai, J.; Zhang, C.; and Yang, Y. 2025. Explainable Multi-Modality Alignment for Transferable Recommendation. In *Proceedings of WWW*, 2076–2084.
- Yang, W.; Zhong, R.; Chen, Y.; Lu, C.; and Jiang, P. 2026. Structured Spectral Reasoning for Frequency-Adaptive Multimodal Recommendation. In *Proceedings of NeurIPS*, 28122–28143.
- You, R.; Li, Y.; Lin, X.; Zhang, X.; Wang, W.; Li, W.; and Nie, L. 2026. R2ec: Towards Large Recommender Models with Reasoning. In *Proceedings of NeurIPS*, 62376–62405.
- Yue, W.; Yin, Y.; Zhang, X.; Shi, B.; Liang, T.; and Wan, J. 2025a. Cot4rec: Revealing user preferences through chain of thought for recommender systems. In *Proceedings of AAAI*, 13142–13151.
- Yue, Z.; Zeng, H.; Wang, Y.; McAuley, J.; and Wang, D. 2025b. Preference-Optimized Retrieval and Ranking for Efficient Multimodal Recommendation. In *Proceedings of SIGKDD*, 3692–3703.
- Zhan, Y.-L.; Tang, X.; Wan, H.; Li, J.; Wen, J.; and Sun, H. 2026. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. In *Proceedings of AAAI*, 12358–12366.
- Zhang, Q.; Cheng, A.; Lu, M.; Zhang, R.; Zhuo, Z.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2025a. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of ICCV*, 20857–20867.
- Zhang, Y.; Xu, W.; Zhao, X.; Wang, W.; Feng, F.; He, X.; and Chua, T.-S. 2025b. Reinforced latent reasoning for llm-based recommendation. *arXiv preprint arXiv:2505.19092*.