Recommendation as Generation: Unifying Personalized Video Generation and Recommendation at Industrial Scale

Yanhua Cheng[*1], Bo Wang[*1], Haotian Zhang[*§2], Xinyuan Gao[*1], Zhihui Yin[1], Ben Xue[1], Yongzhi Li[1], Jieting Xue[1], Ye Ma[1], Minquan Wang[1], Jiahui Li[1], Tianyu Xu[1], Zhiqiang Liu[1], Xiao Lin[1], Shiyang Wen[1], Changcheng Li[1], Liu Liu[2], Quan Chen[1], Peng Jiang[†1], Kun Gai[1]

[1]Kuaishou Technology
Beijing, China

[2]Beihang University
Beijing, China

Abstract

Traditional short-video recommendation systems match user interest to a fixed pool of pre-produced videos, which limits their ability to capture fine-grained and dynamic preferences. We propose **Recommendation-as-Generation** (**RaG**), a new paradigm that generates personalized videos on demand from inferred user interest. Our framework unifies generative recommendation and video generation through shared semantic IDs (SIDs), which disentangle video representation into content semantics and creative style semantics, enabling both fine-grained modeling of user interest and controllable generation of interest-aligned videos. We further develop **Video Generation Agents** (**VGAs**) that are conditioned on inferred SIDs to drive hierarchical planning and refinement for video creation, including visual composition, audio alignment, and artistic effect enhancement. To optimize the framework, we effectively introduce a synergistic cross-domain reward learning mechanism that jointly enforces interest alignment, user feedback, and video quality assessment.

We deploy RaG [1] on an industrial-scale platform with over 400 million daily active users and evaluate it in a revenue-critical advertising scenario. Online A/B tests show up to **1.87%** ad revenue improvement compared to a strong production GRM baseline, demonstrating its effectiveness in driving further revenue gains beyond generative recommendation. Our results highlight a closed-loop generative system as a promising paradigm for integrating personalized video generation into recommendation.

Keywords

Generative Recommendation, Personalized Video Generation, Agents, Semantic Quantization, Reward Learning

1 Introduction

Over the past decade, industrial video recommendation systems have followed a content-first paradigm, where videos are produced offline and recommendation models retrieve and rank items from a fixed pool. Deep learning recommendation models (DLRMs) [3, 5, 34, 36] improve matching accuracy under this setting. More recently, generative recommendation models (GRMs) [7, 27] extend this paradigm by modeling user interest through large-scale autoregressive generation over semantic IDs (SIDs) [16].

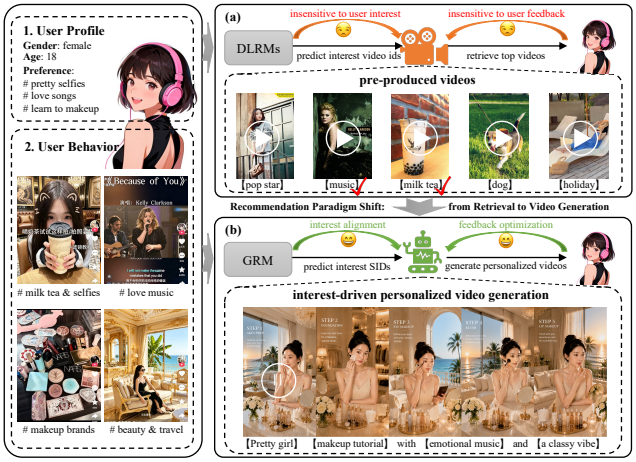

Figure 1: Recommendation paradigm shift. (a) DLRMs retrieve videos from a fixed content pool, leading to suboptimal matches when user interests fall outside the pool; (b) Our paradigm generates personalized videos on demand that both align with the user interests predicted by a GRM and are driven by real user feedback in a closed loop, breaking the fixed-pool limit.

Despite these advances, existing systems remain fundamentally constrained by a static pool of pre-produced videos. Recommendation models can only retrieve the best available content from the existing pool, even when user interests fall outside the pool. This limitation is particularly severe in modern short-video platforms, where user interests are more dynamic, long-tailed, and diverse. As a result, retrieval-based systems are inherently limited in faithfully capturing fine-grained user interest.

Meanwhile, recent breakthroughs in AI-generated content (AIGC) [8, 15, 17, 20, 22] have demonstrated unprecedented capabilities in open-domain video creation. Modern video generation models can produce cinematic-quality visual content with strong semantic controllability, opening up a new opportunity for recommendation:

Can recommendation systems move beyond retrieving existing videos to directly generate personalized videos from inferred user interests?

Answering this question requires addressing two key challenges in recommendation and generation systems.

The first challenge is how to bridge recommendation and generation into a unified framework. Recommendation models are trained on heterogeneous and discrete data, including user profiles, item features, and user behaviors, aiming to predict user interests. In contrast, video generation models operate on multimodal continuous signals, such as text, images, audio, and motion,

[1]Project page: https://recommendation-as-generation.github.io/

[*]Equal contribution.
[§]Work done during an internship at Kuaishou Technology.
[†]Corresponding author.

focusing on generating coherent and high-fidelity videos. Given such fundamental differences in data representation and learning objectives, recommendation and generation are typically developed as two separate tasks, making it difficult to translate predicted user interests into controllable video generation. This separation also blocks user feedback from flowing back into the generation process, limiting the diversity and interest-alignment of the produced content.

The second challenge is how to generate high-quality and interest-aligned videos at industrial scale. Although recent state-of-the-art video generation models [8, 15, 17, 20, 22] achieve strong visual quality, they remain difficult to deploy in large-scale recommendation systems. These models often rely on manual prompting, multi-stage refinement and post-processing with professional tools, resulting in high latency and computational cost to produce a single user-satisfactory video. Personalizing across the diverse and long-tailed interests of hundreds of millions of users further amplifies these costs, making it infeasible to deploy such models directly in production.

To address these challenges, we propose **Recommendation-as-Generation** (**RaG**), a new paradigm that unifies recommendation and personalized video generation in a closed-loop framework, as illustrated in Figure 1. Instead of retrieving from a fixed pool, RaG generates personalized videos directly from inferred user interests.

A key idea of RaG is to use **Disentangled Semantic IDs** (**D-SIDs**) as a unified interface between recommendation and generation. A multimodal large language model encodes each video into two factorized embeddings—one for *content* (entities, topics) and the other for *creative* attributes (style, rhythm, atmosphere). These embeddings are then quantized into discrete *content SIDs* and *creative SIDs*, jointly forming the video's D-SIDs. On the recommendation side, a generative recommendation model (GRM) autoregressively predicts the D-SIDs of user interests. On the generation side, the predicted D-SIDs are decoded into personalized videos, connecting fine-grained interest modeling with controllable video generation.

To realize controllable video generation at scale, RaG develops **Video Generation Agents** (**VGAs**). Compared to monolithic, high-cost diffusion-based or prompt-engineering-heavy pipelines, VGAs adopt a hierarchical planning and refinement framework. Conditioned on user-interest D-SIDs, a fine-tuned **Instruction Model** (**IM**) first translates them into structured generation blueprints. Three role-specialized agents then reason and act over the blueprints, jointly modeling visual composition, audio alignment, and artistic effects. The three agents share a single LLM backbone and are jointly trained end-to-end, differentiated only through prompts and tool access. After the agents complete the pipeline, a bounded reflection loop (capped at two iterations) refines cross-modal consistency, balancing output quality with generation efficiency. The shared backbone further enables KV-cache reuse across agents to substantially accelerate inference. Combined with an SID-indexed cache that amortizes generation cost, VGAs reliably serve recommendation requests for hundreds of millions of users at industrial scale.

To close the optimization loop, RaG introduces **Synergistic Cross-Domain Reward Learning** (**SCRL**). Instead of naive reward aggregation that conflates heterogeneous reward signals, SCRL formulates multi-objective optimization as a constrained policy learning problem: user feedback serves as the primary objective, while interest alignment and video quality act as constraints. Group-decoupled reward normalization (GDPO [10]) is applied per channel to reconcile scale mismatch, followed by a PID-controlled Lagrangian update [19] to stabilize training. Together, SCRL unifies recommendation and video generation into a single closed-loop optimization where user interests, content quality, and real-world feedback co-evolve.

We deploy RaG on a large-scale production platform serving over 400 million daily active users in a revenue-critical advertising scenario. Online A/B testing shows significant improvements in ad revenue, validating the effectiveness of generation-driven personalization for recommendation. To the best of our knowledge, this is the first production-scale system that effectively unifies recommendation and personalized video generation.

Our main contributions are summarized as follows:

- We propose **Recommendation-as-Generation** (**RaG**), a new paradigm that shifts recommendation from retrieving videos within a fixed pool to generating personalized videos directly from inferred user interests. Disentangled Semantic IDs (D-SIDs) serve as the unified latent interface between recommendation and generation, and Synergistic Cross-Domain Reward Learning (SCRL) closes the loop by enforcing interest alignment, user feedback, and video quality assessment.
- We develop industrial-scale **Video Generation Agents** (**VGAs**) with hierarchical planning, collaborative multi-agent execution, and iterative refinement, enabling scalable and high-quality personalized video production.
- Extensive offline experiments and online A/B testing on a production platform demonstrate substantial improvements in ad revenue, validating the effectiveness of large-scale personalized video generation for recommendation.

2 Methodology

2.1 Paradigm Shift: Recommendation as Generation

Conventional recommendation systems [3, 5, 34, 36] retrieve or rank videos from a fixed content pool. Recent generative recommendation models (GRMs) formulate recommendation as autoregressive token prediction [7, 27], but still retrieve videos from the existing content pool set according to the predicted tokens. As a result, these approaches remain limited by content coverage, often yielding suboptimal recommendations when user interests involve novel or long-tail semantics.

To overcome this limitation, we introduce the **Recommendation-as-Generation** (**RaG**) paradigm, which reformulates recommendation as an interest-conditioned video generation problem (Figure 2). Instead of retrieving existing videos, RaG directly generates personalized videos from inferred user interests. One key idea is to unify recommendation and video generation within a shared discrete latent space.

We construct this space using **Disentangled Semantic Video Encoders** (Section 2.2), which map videos into disentangled semantic IDs (D-SIDs). These D-SIDs capture both semantic content and

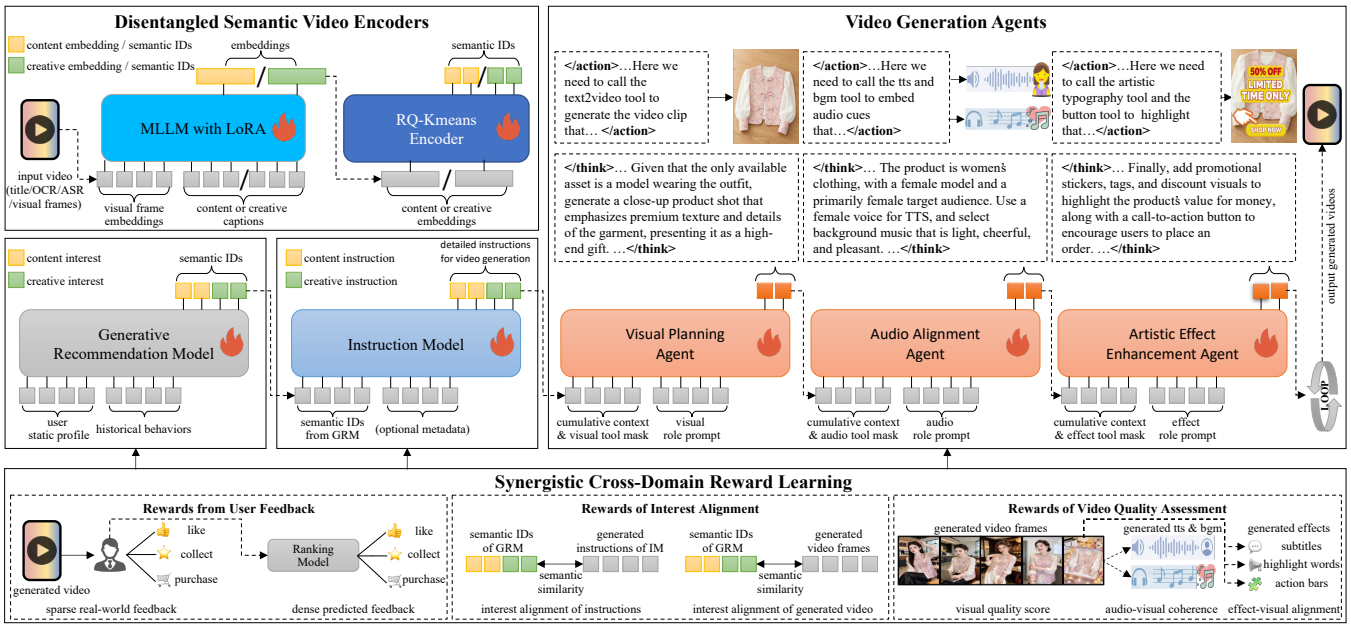

Figure 2: Architecture of the Recommendation-as-Generation (RaG) framework. Videos are encoded into Disentangled Semantic IDs (D-SIDs) that decouple content and creative semantics, forming a shared latent interface for recommendation and generation. The Generative Recommendation Model (GRM) predicts a user's interest D-SIDs from user context. The Instruction Model (IM) then translates these predicted D-SIDs, together with optional metadata, into shot-level production instructions, which are executed by the Video Generation Agents (VGAs) through hierarchical planning and refinement. The full pipeline is jointly optimized under Synergistic Cross-Domain Reward Learning (SCRL).

creative attributes, enabling both fine-grained interest modeling and controllable video generation. Given a video v, the encoder $\mathcal{E}$ produces a sequence of tokens:

$$\text{D–SIDs} = \mathcal{E}(v) = (s^1_{\text{content}}, \ldots, s^L_{\text{content}}, s^1_{\text{creative}}, \ldots, s^L_{\text{creative}}), \quad (1)$$

which jointly represent video semantics and creative structure.

Building on this semantic space, recommendation is recast as generative interest modeling: given a user's profile and interaction history, the Generative Recommendation Model (GRM, Appendix C) autoregressively predicts the sequential D-SIDs representing the user's future interests:

$$p(\text{D–SIDs} \mid \mathbf{c}_{\text{user}}) = \prod_{t=1}^{2L} p(s_t \mid s_{<t}, \mathbf{c}_{\text{user}}), \quad (2)$$

where $\mathbf{c}_{\text{user}}$ denotes the user context.

Unlike prior GRM-based approaches that use predicted D-SIDs as retrieval keys, we treat D-SIDs as *generative interest representations* that can be directly decoded into new content, beyond a fixed pool. The overall pipeline is:

$$\text{D–SIDs} = \mathcal{E}(v) \rightarrow p(\text{D–SIDs} \mid \mathbf{c}_{\text{user}}) \rightarrow \hat{v} = \mathcal{G}(\text{D–SIDs}), \quad (3)$$

where user interests are modeled in the latent semantic space and decoded into personalized videos. However, directly optimizing $\mathcal{G}$ for both generation quality and interest alignment is challenging. We therefore decompose the generation process into a hierarchical framework.

We introduce an **Instruction Model** (Section 2.3) that translates D-SIDs into natural language instructions, providing interpretable and structured guidance for downstream agents. Building on this, we develop **Video Generation Agents** (Section 2.4) that generate videos through collaborative agents, enabling hierarchical planning, multimodal alignment, artistic enhancement, and iterative refinement. Finally, we optimize the entire framework via **Synergistic Cross-Domain Reward Learning** (Section 2.5), jointly capturing user interest alignment, generation quality, and user engagement signals.

2.2 Disentangled Semantic Video Encoders

2.2.1 Multimodal Representation Learning. Building upon Qwen2.5-VL-7B-Instruct [21], we propose an instruction-guided disentangled representation framework that separates semantic content and creative attributes from the same video. For multimodal input processing, we directly reuse Qwen2.5-VL's native visual encoder and text tokenizer.

We first extract its visual token representations using the vision encoder: $H = \mathcal{F}(v), H \in \mathbb{R}^{N \times d}$, where H denotes a sequence of visual tokens capturing spatial-temporal semantics.

To obtain disentangled signals, we leverage our in-house dense captioning model (CapModel) to generate factor-specific textual descriptions:

$$D_m = \text{CapModel}(v, \text{PROMPT}_m), \quad m \in \{\text{content}, \text{creative}\}, \quad (4)$$

where D_{content} describes semantic content (entities, topics), while D_{creative} captures creative attributes (style, rhythm and atmosphere).

The instructions are encoded via the text tokenizer: $Q_m = \mathcal{T}(D_m)$, $Q_m \in \mathbb{R}^{L_m \times d}$, where L_m is the instruction length. We obtain multimodal representations by jointly encoding visual and textual inputs with Qwen2.5-VL-7B-Instruct, and use the last-token hidden state

of the final layer as the pooled multimodal representation:

$$\mathbf{z}_m = \text{Normalize}(\text{VLM}(H, Q_m)), \quad \mathbf{z}_m \in \mathbb{R}^d, \ \|\mathbf{z}_m\|_2 = 1, \tag{5}$$

yielding $\mathbf{z}_{\text{content}}$ and $\mathbf{z}_{\text{creative}}$ as L2-normalized content and creative representations, respectively.

To encourage representation consistency, we employ a contrastive loss for each module:

$$\mathcal{L}_m = -\log \frac{\exp(\text{sim}(\mathbf{z}_m^i, \mathbf{z}_m^j)/\tau)}{\sum_k \exp(\text{sim}(\mathbf{z}_m^i, \mathbf{z}_m^k)/\tau)}, \tag{6}$$

where $\mathbf{z}_m^j$ is the positive pair of $\mathbf{z}_m^i$ within a batch, and k indexes all candidates including the positive.

To reduce cross-factor leakage, we introduce an orthogonality constraint:

$$\mathcal{L}_{\text{orth}} = \|\mathbf{z}_{\text{content}}^\top \mathbf{z}_{\text{creative}}\|_2^2. \tag{7}$$

The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{content}} + \gamma_1 \mathcal{L}_{\text{creative}} + \gamma_2 \mathcal{L}_{\text{orth}}. \tag{8}$$

2.2.2 Discrete Tokenization. To facilitate generative recommendation within the latent space, we discretize the disentangled multimodal representations into semantic IDs.

Specifically, each representation $\mathbf{z}_m$ is independently quantized via Residual Quantization (RQ)-based K-means [11], yielding a quantized embedding $\mathbf{e}_m$ that approximates $\mathbf{z}_m$ as a sum of codebook vectors across L hierarchical layers:

$$\mathbf{e}_m = \sum_{l=1}^{L} \mathbf{c}_m^l(s_m^l) \ \approx \ \mathbf{z}_m, \qquad \mathbf{e}_m \in \mathbb{R}^d, \tag{9}$$

where s_m^l denotes the discrete code index at layer l for modality m, and $\mathbf{c}_m^l(\cdot)$ is the corresponding codebook lookup. Each modality maintains an independent codebook with 8,192 entries per layer. The final D-SIDs are obtained by concatenating the per-modality code sequences: $\textbf{D-SIDs} = \left[\, s_{\text{content}}^{1:L}; \, s_{\text{creative}}^{1:L} \,\right]$.

2.3 Instruction Model

The Instruction Model translates disentangled semantic IDs into shot-level video production instructions. Unlike conventional caption generation, these instructions explicitly specify scene composition, camera motion, temporal pacing, and cinematic style, serving as an intermediate semantic bridge between discrete user interests and controllable video generation.

2.3.1 Supervision Construction. Since no off-the-shelf dataset contains video-instruction pairs at the shot level, we distill supervision from a strong multimodal teacher. For each video v, we first extract its D-SIDs (Section 2.2), and then prompt Gemini2.5 Pro [4] with a carefully designed instruction template $\text{PROMPT}_{\text{inst}}$ to produce the target shot-level script:

$$D_{\text{inst}} \ = \ \text{Gemini}(v, \ \text{PROMPT}_{\text{inst}}) \ = \ (y_1, \, y_2, \, \ldots, \, y_{L_{\text{inst}}}), \tag{10}$$

where D_{inst} is a token sequence of length L_{inst} serving as ground-truth supervision. To accommodate advertising scenarios where the generated video must reflect specific products being promoted, we further introduce an *optional* metadata factor D_{meta} (e.g., product information and marketing topics) as an auxiliary conditioning signal. When unavailable (e.g., for pure organic videos), D_{meta} is simply masked, leaving instruction generation conditioned on D-SIDs alone.

2.3.2 Model and Optimization Objective. We instantiate the Instruction Model with Qwen3-8B [29], which consumes two heterogeneous token sequences—the primary D-SIDs and the auxiliary D_{meta} (masked when unavailable)—mapped into the LLM's input embedding space and concatenated as the prefix.

For D-SIDs, we reconstruct continuous embeddings from the discrete codes via reverse RQ-Kmeans, $\mathbf{e}_{\text{D-SIDs}} = [\mathbf{e}_{\text{content}}; \, \mathbf{e}_{\text{creative}}] \in \mathbb{R}^{2\times d}$, and map them through a learnable projector $\phi(\cdot)$ to $\mathbf{h}_{\text{D-SIDs}} = \phi(\mathbf{e}_{\text{D-SIDs}}) \in \mathbb{R}^{2\times d'}$. For metadata, D_{meta} is tokenized and embedded by the LLM's native text tokenizer $\mathcal{T}$ to $Q_{\text{meta}} = \mathcal{T}(D_{\text{meta}}) \in \mathbb{R}^{L_{\text{meta}}\times d'}$, with L_{meta} denoting the token length. Conditioned on both, the model autoregressively predicts the instruction sequence

$$\hat{D}_{\text{inst}} \ = \ \text{LLM}(\mathbf{h}_{\text{D-SIDs}}, \, Q_{\text{meta}}), \tag{11}$$

and is optimized with a standard next-token prediction loss against the Gemini-distilled supervision:

$$\mathcal{L}_{\text{NTP}} \ = \ -\sum_{t=1}^{L_{\text{inst}}} \log P\big(y_t \,\big|\, y_{<t}, \, \mathbf{h}_{\text{D-SIDs}}, \, Q_{\text{meta}}\big), \tag{12}$$

so that $\hat{D}_{\text{inst}}$ token-wise approximates D_{inst}.

2.3.3 Three-Stage Training. We adopt a three-stage training strategy: in the first stage, the backbone LLM is frozen and only the projector $\phi(\cdot)$ is optimized to align D-SIDs' embeddings with the language space; in the second stage, both the projector and LLM parameters are jointly fine-tuned for improved semantic fidelity and controllable instruction generation; in the third stage, we further enhance the model using reinforcement learning with reward optimization, as described in Section 2.5.

2.4 Video Generation Agents

As illustrated in Figure 2, industrial-scale personalized video generation cannot be effectively handled by a monolithic generator. The one-shot production of visuals, audio, and effects often leads to semantic inconsistency and limited controllability. Moreover, video production exhibits a hierarchical dependency structure where visual planning determines the narrative flow, while audio and effects are conditioned on the visual state.

To address this, we propose Video Generation Agents (VGAs), formulated as a structured multi-agent decision process over an evolving generation state.

Agentic Formulation. We model video generation as a sequential decision process executed by a team of sub-agents. At each step t, the active sub-agent observes a state $\mathcal{S}_t$, selects an action a_t according to its policy π_θ, and transitions to the next state via a deterministic operator $\mathcal{P}$:

$$a_t \sim \pi_\theta(a_t \mid \mathcal{S}_t), \qquad \mathcal{S}_{t+1} = \mathcal{P}(\mathcal{S}_t, a_t). \tag{13}$$

To enable efficient backbone reuse across sub-agents (detailed later), we serialize the state as an ordered prefix followed by stage-dependent tokens:

$$\mathcal{S}_t = \big[\, \underbrace{\hat{D}_{\text{inst}}; \, D_{\text{tool}}}_{\text{shared prefix}}; \ \underbrace{O_{<t}; \, \text{PROMPT}_{\text{role}}}_{\text{stage-dependent}} \,\big], \tag{14}$$

where $\hat{D}_{\text{inst}}$ is the instruction sequence produced by the Instruction Model; D_{tool} is the description of all available tools, including in-house pretrained text-to-video and image-to-video models and external audio synthesis and visual effect APIs; $O_{<t}$ is the running concatenation of all earlier sub-agents' role prompts and their generated outputs; and $\text{PROMPT}_{\text{role}}$ is a short role-specific prompt that activates the current sub-agent. The action a_t corresponds to a modality-specific intent for visual, audio, or effect generation.

VGAs consist of three role-specialized sub-agents, each acting according to its own policy:

$$\pi_\theta(a_t \mid \mathcal{S}_t) = \{\pi_{\text{visual}}, \pi_{\text{audio}}, \pi_{\text{effect}}\}(a_t \mid \mathcal{S}_t), \tag{15}$$

corresponding to visual planning, audio synthesis, and post-production effects, respectively. We next describe each sub-agent in turn.

1. Visual Planning Agent (VPA). At the visual stage, $\text{PROMPT}_{\text{role}}$= $\text{PROMPT}_{\text{visual}}$ and $O_{<t}$ is empty (or carries the previous reflection round's content). The VPA acts as the global controller, producing a clip-level storyboard with scene segments, layout configurations, and temporal boundaries: $\mathcal{I}_{\text{visual}} = \pi_{\text{visual}}(\mathcal{S}_t)$.

2. Audio Alignment Agent (AAA). At the audio stage, $\text{PROMPT}_{\text{role}}$ = $\text{PROMPT}_{\text{audio}}$ and $O_{<t}$ extends with ($\text{PROMPT}_{\text{visual}}, \mathcal{I}_{\text{visual}}$). The AAA generates temporally aligned audio (speech and music) synchronized with scene transitions: $\mathcal{I}_{\text{audio}} = \pi_{\text{audio}}(\mathcal{S}_t)$.

3. Artistic Effect Enhancement Agent (AEEA). At the effect stage, $\text{PROMPT}_{\text{role}}$ = $\text{PROMPT}_{\text{effect}}$ and $O_{<t}$ further extends with ($\text{PROMPT}_{\text{audio}}, \mathcal{I}_{\text{audio}}$). The AEEA performs post-production refinement by adding subtitles, visual effects, transitions, and call-to-action elements: $\mathcal{I}_{\text{effect}} = \pi_{\text{effect}}(\mathcal{S}_t)$.

Hierarchical Generation with Bounded Reflection. The three intent outputs are composed into the final video via a unified generation operator $\mathcal{G}$:

$$\mathcal{V} = \mathcal{G}(\mathcal{I}_{\text{visual}}, \mathcal{I}_{\text{audio}}, \mathcal{I}_{\text{effect}}). \tag{16}$$

To improve cross-modal consistency, VGAs operate within a bounded reflection loop that follows the standard Observe→Think→Act cycle, capped at two iterations to balance output quality with generation efficiency.

Shared Backbone and KV-Cache Reuse. Although VGAs comprise three role-specific sub-agents, they are not three separate models—all share a single Qwen2.5-32B backbone [28] with fully shared parameters. Differentiation arises purely from the state $\mathcal{S}_t$ in Eq. (14): $\text{PROMPT}_{\text{role}}$ activates the target sub-agent, an attention mask over D_{tool} restricts it to the accessible tool subset, and $O_{<t}$ supplies the cumulative upstream context. This serialization also enables straightforward KV-cache reuse: with sub-agents invoked sequentially and $O_{<t}$ growing append-only, every previously generated token stays in the KV cache, leaving each downstream sub-agent to only encode its own $\text{PROMPT}_{\text{role}}$, substantially reducing per-request inference latency.

The agent policies are optimized via synergistic cross-domain reward signals that jointly capture generation quality, interest alignment, and user feedback, as detailed in Section 2.5.

2.5 Synergistic Cross-Domain Reward Learning

2.5.1 Cross-Domain Reward Formulation. To further enhance both recommendation and video generation performance, we formulate a structured cross-domain reward scheme with three synergistic objectives: *video quality*, *interest alignment*, and *user feedback*. Unless otherwise specified, all reward models share the same Transformer-based architecture and are trained on task-specific datasets.

1. Video Quality Rewards.

We evaluate the perceptual and compositional quality of generated videos from three complementary aspects: visual quality, audio-visual coherence, and effect-visual alignment. Formally, we define: $R_{\text{quality}} = R_{\text{visual}} + R_{\text{audio}} + R_{\text{effect}}$, where:

- R_{visual} evaluates visual quality, including aesthetic appeal and spatio-temporal consistency to ensure coherent motion and stable rendering,
- R_{audio} measures alignment between audio and visual content, covering both speech synchronization (TTS) and background music consistency (BGM),
- R_{effect} captures the quality and alignment of visual effects, including subtitles, highlights, and interactive elements such as action bars.

2. Interest Alignment Rewards.

To keep generated content aligned with user interests throughout the pipeline, we apply alignment rewards at multiple stages, anchored on the D-SIDs that encode user interests in a structured latent space: $R_{\text{align}} = R_{\text{instr-align}} + R_{\text{rep-align}}$, where:

- $R_{\text{instr-align}}$ enforces semantic consistency between GRM-generated D-SIDs and the generated instructions,
- $R_{\text{rep-align}}$ measures semantic similarity between GRM-generated D-SIDs and the generated videos.

3. User Feedback Rewards.

To enhance downstream user engagement, we leverage user interaction signals such as clicks and conversions as the core reward for optimization. However, real-world interaction signals are sparse and delayed, making them insufficient for stable and efficient policy optimization.

To mitigate this issue, we augment sparse interaction signals with dense engagement estimates from deployed ranking models. We define the overall reward as: $R_{\text{feedback}} = R_{\text{real}} + R_{\text{pred}}$, where:

- R_{real} denotes sparse but high-fidelity user interaction signals observed from real feedback, including behaviors such as click, like, collect, and purchase,
- R_{pred} denotes dense engagement signals estimated by ranking models, which capture user preference strength beyond explicit interactions.

2.5.2 Constrained Policy Optimization with GDPO. To jointly optimize the heterogeneous, cross-domain rewards introduced above, we formulate reward learning within the RaG framework as a *constrained policy optimization* problem solved by GDPO [10]. The design addresses two practical challenges in multi-reward RL: (i) the scale mismatch and optimization instability caused by heterogeneous reward signals, and (ii) the difficulty of statically balancing competing objectives without sacrificing the dominant goal.

Problem setup. Given an input context x, the policy π_θ samples a candidate set $\mathcal{Y} = \{y_1, \ldots, y_K\} \sim \pi_\theta(\cdot \mid x)$. Each candidate y_i is evaluated by a collection of heterogeneous reward functions covering user feedback $R_{\text{feedback}}(y_i)$, interest alignment $R_{\text{align}}(y_i)$,

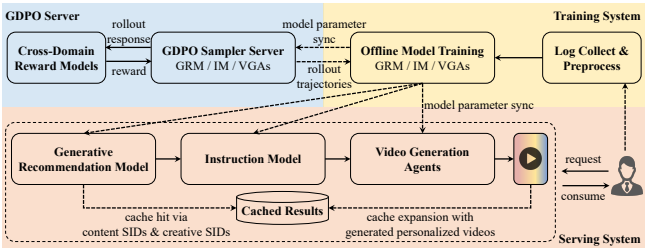

Figure 3: Training and serving architecture of the proposed Recommendation-as-Generation system.

and video quality $R_{\text{quality}}(y_i)$. These rewards differ in scale, density, and reliability, motivating the constrained formulation below.

Constrained reward formulation. We designate user feedback as the primary objective and treat interest alignment and video quality as inequality constraints with target thresholds $\tau_{a(lign)}$ and $\tau_{q(uality)}$. The composite reward for each candidate y_i is defined as

$$R(y_i) = R_{\text{feedback}}(y_i) - \sum_{c\in\{a,q\}} \lambda_c(t)\,\text{ReLU}\big(\tau_c - R_c(y_i)\big), \tag{17}$$

where $\lambda_a(t), \lambda_q(t) \geq 0$ are time-varying Lagrangian multipliers, updated via a PID-controlled rule on constraint violations [19] to avoid the oscillation and overshoot of naive primal–dual updates. To avoid hand-tuned magic numbers, we calibrate each threshold relative to the SFT baseline distribution on a held-out validation set as $\tau_c = \mu_c^{\text{base}} + k_c\,\sigma_c^{\text{base}}$, where the strictness factor k_c encodes the module's role in RaG: VGAs adopt the strictest setting ($k_c = 1.1$ for both τ_a and τ_q) as it directly governs final video generation; IM retains a comparable τ_a ($k_a = 0.8$) to enforce instruction-level alignment; while GRM applies a relaxed τ_a ($k_a = 0.3$), with the video-quality constraint omitted for the latter two modules.

Group-decoupled normalization and advantage. Given the constrained reward in Eq. (17), GDPO further eliminates residual scale mismatch among reward channels via per-reward standardization prior to aggregation, and computes a group-relative advantage over the sampled candidate set $\mathcal{Y}$:

$$A_i = \frac{R(y_i) - \mu(\mathcal{Y})}{\sigma(\mathcal{Y}) + \epsilon}, \tag{18}$$

where $\mu(\mathcal{Y})$ and $\sigma(\mathcal{Y})$ denote the group-level mean and standard deviation of the rewards over $\mathcal{Y}$. This decoupled normalization stabilizes optimization across rewards with disparate magnitudes.

Optimization objective. The policy is updated by maximizing the group-relative advantage anchored to the frozen SFT policy π_{ref}:

$$\mathcal{L}_{\text{GDPO}} = -\mathbb{E}_{(x,y_i)}\left[A_i\,\log\frac{\pi_\theta(y_i \mid x)}{\pi_{\text{ref}}(y_i \mid x)}\right]. \tag{19}$$

For brevity, we omit the importance-sampling ratio clipping and the KL regularization term against π_{ref} that are commonly used to stabilize policy optimization; both are retained in our implementation and follow the standard GDPO formulation [10].

3 Deployment

We deploy RaG in Kuaishou's large-scale advertising system, serving over 400 million users (Figure 3). The system unifies real-time user interest modeling with large-scale personalized video generation under strict latency constraints. Since video generation is orders of magnitude slower than interest inference (Appendix A), we design a decoupled deployment architecture to bridge this efficiency gap while maintaining end-to-end personalization quality. The system consists of three decoupled modules: real-time interest modeling, nearline video generation, and latency-aware serving.

Real-Time Interest Modeling. The Generative Recommendation Model (GRM) is continuously trained on streaming user interaction logs (impression, click, watch time, and conversion) to adapt to non-stationary user behavior, combining streaming supervised updates with periodic GDPO-based optimization.

At real-time inference, GRM performs low-latency autoregressive generation of structured Semantic IDs (SIDs), which encode user interests as semantic targets for downstream content generation.

Nearline Video Generation. The Instruction Model (IM) and Video Generation Agents (VGAs) are trained on large-scale agentic supervision data curated from high-quality videos, and optimized via supervised fine-tuning followed by constrained GDPO to jointly improve generation quality and interest alignment. Both models are periodically updated in full-batch mode to ensure training stability while adapting to evolving user interests and emerging video patterns.

At serving time, conditioned on GRM-generated SIDs, IM and VGAs operate in a nearline pipeline to generate personalized videos. To handle the heavy generation load, VGAs apply *KV-cache reuse* as established in Section 2.4: with sub-agents invoked sequentially over an append-only state, every previously generated token stays in the KV cache, leaving only each sub-agent's own short $\text{PROMPT}_{\text{role}}$ to be encoded per call, substantially reducing per-request inference latency. The outputs are continuously accumulated into a growing personalized video space, enabling coverage expansion while decoupling video generation from real-time serving.

Latency-Aware Serving. To meet real-time consumption requirements in recommendation scenarios, the system adopts a hierarchical serving strategy organized around whether the requested content-level SIDs are covered by the cache.

Case 1: content-SIDs hit. If the matched cache entry also covers the creative-level SIDs, the system returns the previously generated video directly with negligible latency; otherwise, it serves a content-consistent cached video while asynchronously scheduling the missing creative variations, with higher-frequency creatives prioritized in the generation queue.

Case 2: content-SIDs miss. The system serves videos associated with the nearest-neighbor SIDs for immediate consumption, while enqueuing the uncovered SIDs for prioritized future generation.

4 Experiments

4.1 Online A/B Testing

We deploy the proposed Recommendation-as-Generation (RaG) framework in the real-world advertising platform of Kuaishou and conduct large-scale online A/B experiments to evaluate its industrial effectiveness. The experiments mainly focus on two aspects:

Table 1: Online A/B test results. Rev. denotes ad revenue. Results are reported as relative improvements over production baselines.

Method	Rev. (%↑) vs. DLRM baseline	Rev. (%↑) vs. GRM baseline
Production Baseline		
DLRM baseline	–	–
GRM baseline [27]	+3.526%	–
Enhanced GRM		
GRM + Disentangled-SIDs (D-SIDs)	+4.460%	+0.902%
Full System (RaG)		
RaG (GRM + D-SIDs + IM + VGAs + SCRL)	**+5.462%**	**+1.870%**

Table 2: Quality of the Disentangled SIDs. We report both (i) embedding-based semantic retrieval quality and (ii) SID discretization quality. Improvements over the strongest baseline are highlighted. *Impr.*: improvement; *R@k*: Recall@k; *Cpr.*: compression rate; *Col.*: collision rate.

Method	Semantic Retrieval (R@K)			Discretization Quality
	R@1 ↑	*R@5* ↑	*R@10* ↑	*Cpr.* ↓ / *Col.* ↓
VLM2Vec-V2 [12]	0.485	0.690	0.756	–
QARM [11]	0.541	0.812	0.893	1.14 / 18.24%
Qwen2.5-VL-7B [32]	0.769	0.948	0.977	–
Ours (D-SIDs)	**0.896**	**0.985**	**0.994**	**1.02 / 2.62%**
Impr.	**+16.5%**	**+3.9%**	**+1.7%**	**-10.5% / -15.6pp**

(1) the effectiveness of Disentangled Semantic IDs (D-SIDs) for generative recommendation, and **(2)** the additional gains brought by SID-driven personalized video generation.

Table 1 summarizes the online results. Replacing the production DLRM-based pipeline with the Generative Recommendation Model (GRM) yields consistent ad revenue gains, and the proposed Disentangled Semantic IDs (D-SIDs) further lift the improvement from +3.526% to +4.460%, confirming that decoupling content and creative semantics yields a more structured latent space and mitigates interference during autoregressive generation. Nevertheless, both variants remain within the retrieval paradigm, selecting candidates from a fixed pool.

Finally, the full RaG framework—integrating GRM, D-SIDs, the Instruction Model (IM), and Video Generation Agents (VGAs) under Synergistic Cross-Domain Reward Learning (SCRL)—delivers a +5.462% ad revenue gain over the DLRM-based pipeline. Crucially, RaG also outperforms the strong GRM baseline by +1.870%, with this additional lift coming directly from D-SIDs-driven personalized video generation. This marks a paradigm shift from retrieval-based to generation-based recommendation, where user interests actively drive personalized content production rather than merely matching existing candidates.

4.2 Offline Ablation Studies

We ablate the key components of RaG framework—D-SIDs, IM, VGAs, and SCRL optimization—to assess their contributions in terms of semantic representation quality, instruction generation capability, and reward-driven video generation performance.

4.2.1 Quality of Disentangled SIDs. The D-SIDs consist of two core components, i.e., multimodal representation learning and semantic quantization; we systematically analyze their effectiveness in the following experiments.

Table 3: Evaluation of videos between the proposed VGAs vs. the workflow baseline. For Automated Score, we present average and median score.

Metric	Workflow Baseline	VGAs	*Impr.*
Automated Score ↑	62.4 / 62.0	**71.3 / 76.0**	**+14.3% / +22.6%**
Automated Win Rate ↑	28.7%	**70.1%**	**+41.4pp**
User Study Win Rate ↑	34.4%	**52.9%**	**+18.5pp**

Table 4: Reward ablation with corresponding evaluation metrics. For each reward component, we report its dedicated metric on a corresponding evaluation set, comparing the policy trained with that reward (*Ours*) against the no-reward base policy (*Base*).

Video Quality Rewards			Automated Win Rate ↑
R_{visual}	R_{audio}	R_{effect}	Base → Ours
✓			29.3% → **50.7% (+21.4pp)**
	✓		24.0% → **48.0% (+24.0pp)**
		✓	22.7% → **41.3% (+18.6pp)**
✓	✓	✓	37.3% → **56.0% (+18.7pp)**
+ Interest Alignment Rewards			**Interest Alignment Score ↑**
R_{align}			Base → Ours
✓			0.707 → **0.828 (+17.1%)**

Multimodal Representation Learning. We evaluate the proposed instruction-guided representation learning under a product-level retrieval setting to ensure fair comparison of semantic alignment capability. As shown in Table 2, our method consistently outperforms all baselines, achieving 0.896/0.985/0.994 in R@1/5/10. In particular, R@1 improves by +16.5% over the strongest baseline (Qwen2.5-VL-7B), demonstrating stronger semantic discriminability under identical retrieval conditions.

Semantic Quantization. We construct the D-SIDs by applying RQ-KMeans residual quantization separately to the content and creative embeddings, yielding disentangled content SIDs and creative SIDs. For a fair comparison, both D-SIDs and QARM adopt an identical quantization setup with a 4-layer codebook and 8,192 codes per layer. As shown in Table 2, our method achieves superior discretization quality, reducing compression distortion to 1.02 and collision rate to 2.62%. Compared to QARM (1.14/18.24%), this corresponds to a 10.5% reduction in compression error and a 15.6pp lower collision rate, indicating a more compact and collision-resistant semantic space.

4.2.2 Instruction Model Configuration. We evaluate the Instruction Model in terms of decoding fidelity, measured by the cosine similarity between generated instructions and ground-truth summaries using Qwen3-Embedding-8B. Empirically, we observe that both increased training data and model capacity lead to consistent improvements. Specifically, the performance improves from 0.7760 (8B model, 100K samples) to 0.8096 (8B model, 1M samples), and further to 0.8212 with a 32B model trained on 1M samples.

Considering the trade-off between performance and computational efficiency, we adopt the 8B model trained on 1M samples as the default configuration, which achieves competitive decoding fidelity while offering significantly lower deployment cost compared to larger models.

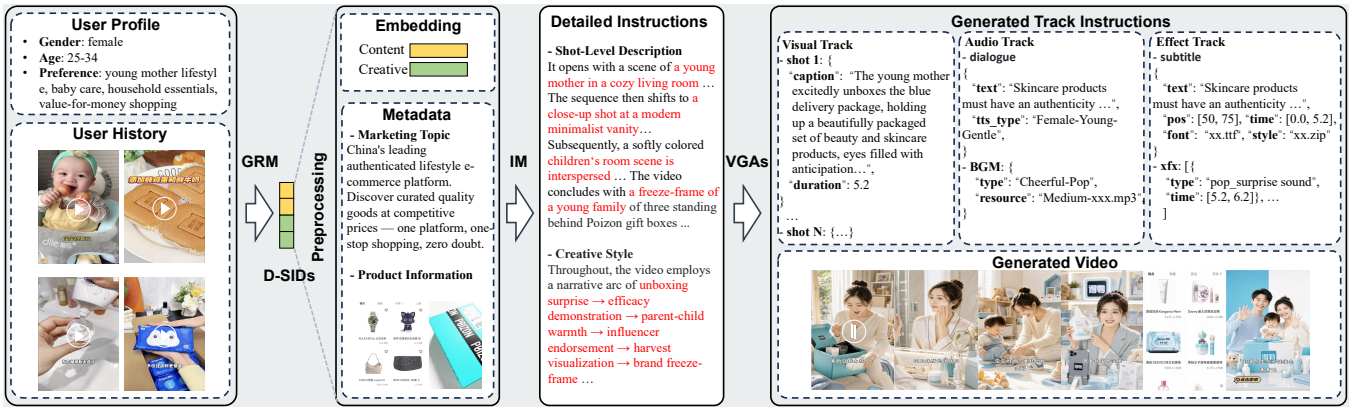

Figure 4: Qualitative example of interest-driven personalized video generation in advertising scenarios.

4.2.3 Performance Analysis of Video Generation. We evaluate the effectiveness of the proposed Video Generation Agents (VGAs) from two perspectives: (i) system-level comparison against conventional production pipelines, and (ii) the contribution of different reward components to optimization. Specifically, we assess three aspects: generation quality, including automated and human preference evaluations; and interest alignment score, measuring consistency between generated videos and target user interests. See Appendix B for details.

System-level Comparison. We compare VGAs with a conventional *workflow baseline*—a hand-crafted pipeline composed of instruction generation, rough-cut (visual clip composition), and fine-cut (TTS synthesis and post-production effects) stages executed in a fixed order. Such rigid execution prevents adaptation to diverse user-specific generation requirements, motivating the agentic design of VGAs.

As shown in Table 3, VGAs consistently outperform the baseline across all metrics. The gains stem from two capabilities: *reasoning*, enabled by a hierarchically structured end-to-end framework that supports coherent cross-modal planning; and *reflection*, which improves output quality through iterative self-correction and re-planning, capped at two iterations to maintain inference latency comparable to the workflow baseline while achieving substantial quality gains.

Reward Contribution Analysis. We analyze the contribution of individual reward components under our synergistic cross-domain rewards. Since user feedback serves as the primary objective and is always retained, we focus the ablation on the two constraint-side rewards: video quality and interest alignment.

As shown in Table 4, each video-quality sub-reward is evaluated on its own dedicated test set, with *Base* denoting the policy trained without any reward optimization. Each sub-reward—visual fidelity, audio alignment, and effect enhancement—independently improves the Automated Win Rate over the base, and jointly optimizing all three yields the strongest result, confirming the necessity of balancing all three perceptual aspects.

Building on the quality rewards, incorporating the interest alignment reward further lifts the Interest Alignment Score (0.707 → 0.828), indicating substantially stronger consistency between generated content and user interest.

Overall, these results show that the quality and alignment rewards play complementary roles—the former safeguards perceptual fidelity while the latter enforces semantic relevance—and their joint optimization, anchored by the primary user-feedback objective, produces a more robust and user-aligned video generation policy.

4.2.4 Qualitative Analysis of Personalized Video Generation. Figure 4 illustrates the end-to-end pipeline of our RaG framework, where user interests are directly transformed into video generation.

Given a representative user profile (female, 25–34) interested in young-mother lifestyle, baby care, household essentials, and value-oriented shopping, the Generative Recommendation Model (GRM) first infers Disentangled SIDs (D-SIDs) that jointly capture content and creative preferences. These D-SIDs are mapped into structured embeddings and, in this advertising scenario, further enriched with the optional metadata factor D_{meta} encoding product information and marketing topics. Conditioned on these representations, the Instruction Model (IM) produces shot-level production instructions, which are subsequently executed by the Video Generation Agents (VGAs) to coordinate visual, audio, and effect generation. The resulting video achieves high quality with strong alignment to user interests.

5 Related Work

Retrieval-Based Recommendation. Traditional recommendation systems [3, 5, 9, 13, 34, 36] follow a retrieve-and-rank paradigm over a fixed pool of pre-produced items.

Recent advances explore generative recommendation by modeling item IDs as discrete tokens and formulating recommendation as next-token prediction over Semantic IDs [24, 31, 33]. To meet industrial latency constraints, efficient architectures have been further proposed for large-scale SID prediction [7, 27, 35].

However, these methods still rely on retrieving from a static content pool conditioned on predicted tokens, leaving recommendation and content creation fundamentally decoupled and preventing end-to-end optimization.

Personalized AI-Generated Content. Recent progress in generative models has motivated a shift from retrieving pre-produced content to generating personalized content conditioned on user

preferences. Early efforts explore preference-guided LLM generation or conditioning diffusion models on user signals for image synthesis [18, 23, 30].

These ideas have been extended to richer modalities, including personalized advertising text generation [2] and dialogue-based preference elicitation for visual content generation [25]. More recently, NextAds [26] studies personalized video advertising by conditioning generation on observed user preferences, but focuses primarily on the generation module without modeling an end-to-end pipeline from user interest representation to controllable video production, and does not consider industrial deployment efficiency and cost.

Overall, existing approaches mostly treat user interest modeling and controllable content generation as separate tasks, leaving room for a unified framework that jointly optimizes both within a closed-loop industrial system.

6 Conclusion

We propose **Recommendation-as-Generation** (**RaG**), a unified paradigm that shifts recommendation toward generation-driven personalization. RaG bridges user interest modeling and controllable video generation through Disentangled Semantic IDs as a shared interface, scalable Video Generation Agents for industrial deployment, and Synergistic Cross-Domain Reward Learning for closed-loop optimization. Online A/B tests show that RaG consistently improves ad revenue over strong commercial baselines, demonstrating the effectiveness of the proposed paradigm in real-world industrial settings.

RaG currently serves nearline rather than in real time, with VGAs being the dominant latency bottleneck. Future work will fold the Instruction Model into VGAs for a tighter generation path, and further accelerate VGAs through stronger model distillation and inference optimization, moving toward on-the-fly personalized generation.

References

[1] Anthropic. Introducing Claude 4.5 Sonnet. https://www.anthropic.com/news/claude-4-5-sonnet, September 2025. Anthropic announcement.

[2] Junyi Chen, Lu Chi, Siliang Xu, Shiwei Ran, Bingyue Peng, and Zehuan Yuan. Hllm-creator: Hierarchical llm-based personalized creative generation. *arXiv preprint arXiv:2508.18118*, 2025.

[3] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[5] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[6] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[7] Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*, 2025.

[8] Google DeepMind. Veo 3: Our most capable video generation model. https://deepmind.google/models/veo/, May 2025. Google DeepMind product announcement.

[9] Jian Jia, Yipei Wang, Yan Li, Honggang Chen, Xuehan Bai, Zhaocheng Liu, Jian Liang, Quan Chen, Han Li, Peng Jiang, et al. Learn: knowledge adaptation from large language model to recommendation for practical industrial application. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11861–11869, 2025.

[10] Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

[11] Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, Changqing Qiu, Jiaqi Zhang, Xu Zhang, Zhiheng Yan, Jingming Zhang, Simin Zhang, Mingxing Wen, Zhaojie Liu, and Guorui Zhou. QARM: quantitative alignment multi-modal recommendation at kuaishou. In *CIKM*, pages 5915–5922. ACM, 2025.

[12] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

[13] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*, 2019.

[14] OpenAI. GPT-5.1 Instant and GPT-5.1 Thinking system card addendum. https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf, November 2025. OpenAI technical report addendum.

[15] OpenAI. Sora 2 system card. https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf, 2025.

[16] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

[17] Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

[18] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 3833–3843, 2024.

[19] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9133–9143. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

[20] Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

[21] Qwen Team. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[22] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[23] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516*, 2023.

[24] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2400–2409, 2024.

[25] Xianquan Wang, Zhaocheng Du, Huibo Xu, Shukang Yin, Yupeng Han, Jieming Zhu, Kai Zhang, and Qi Liu. Personalized visual content generation in conversational systems. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[26] Yiyan Xu, Ruoxuan Xia, Wuqiang Zheng, Fengbin Zhu, Wenjie Wang, and Fuli Feng. Nextads: Towards next-generation personalized video advertising. *arXiv preprint arXiv:2603.02137*, 2026.

[27] Ben Xue, Dan Liu, Lixiang Wang, Mingjie Sun, Peng Wang, Pengfei Zhang, Shaoyun Shi, Tianyu Xu, Yunhao Sha, Zhiqiang Liu, et al. Generative recommendation for large-scale advertising. *arXiv preprint arXiv:2602.22732*, 2026.

[28] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[29] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[30] Hao Yang, Jianxin Yuan, Shuai Yang, Linhe Xu, Shuo Yuan, and Yifan Zeng. A new creative generation pipeline for click-through rate with stable diffusion model. In *Companion Proceedings of the ACM Web Conference 2024*, pages 180–189, 2024.

[31] Jun Yin, Zhengxin Zeng, Mingzheng Li, Hao Yan, Chaozhuo Li, Weihao Han, Jianjin Zhang, Ruochen Liu, Hao Sun, Weiwei Deng, et al. Unleash llms potential for sequential recommendation by coordinating dual dynamic index mechanism. In *Proceedings of the ACM on Web Conference 2025*, pages 216–227, 2025.

[32] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[33] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1435–1448. IEEE, 2024.

[34] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1059–1068, 2018.

[35] Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, Pengfei Zheng, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Ruiming Tang, Shiyao Wang, Shujie Yang, Tao Wu, Wuchao Li, Xinchen Luo, Xingmei Wang, Yi Su, Yunfan Wu, Zexuan Cheng, Zhanyu Liu, Zixing Zhang, Bin Zhang, Boxuan Wang, Chaoyi Ma, Chengru Song, Chenhui Wang, Chenglong Chu, Di Wang, Dongxue Meng, Dunju Zang, Fan Yang, Fangyu Zhang, Feng Jiang, Fuxing Zhang, Gang Wang, Guowang Zhang, Han Li, Honghui Bao, Hongyang Cao, Jiaming Huang, Jiapeng Chen, Jiaqiang Liu, Jinghui Jia, Kun Gai, Lantao Hu, Liang Zeng, Qiang Wang, Qidong Zhou, Rongzhou Zhang, Shengzhe Wang, Shihui He, Shuang Yang, Siyang Mao, Sui Huang, Tiantian He, Tingting Gao, Wei Yuan, Xiao Liang, Xiaoxiao Xu, Xugang Liu, Yan Wang, Yang Zhou, Yi Wang, Yiwu Liu, Yue Song, Yufei Zhang, Yunfeng Zhao, Zhixin Ling, and Ziming Li. Onerec-v2 technical report, 2025. URL https://arxiv.org/abs/2508.20900.

[36] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1079–1088, 2018.

Table 5: Inference efficiency comparison across modules of RaG system.

Component	D-SIDs (Nearline)	GRM (Online)	IM (Nearline)	VGAs (Nearline)
Latency	~4s	~100ms	~2.5s	~180s

A Runtime Analysis of RaG Modules

We deploy the proposed RaG system in Kuaishou's large-scale advertising and recommendation infrastructure, serving over 400 million users. To better understand system efficiency, we conduct a runtime analysis of each core component under both online and nearline deployment settings. Specifically, the Generative Recommendation Model (GRM) operates in an online serving regime for real-time recommendation, while Disentangled Semantic IDs (D-SIDs), the Instruction Model (IM), and the Video Generation Agents (VGAs) are executed in a nearline pipeline due to their higher computational cost and generation latency. All latency numbers reported in Table 5 are measured under live production traffic, reflecting the actual serving conditions.

B Evaluation Metrics for Video Quality and Interest Alignment

To provide a fair and unbiased evaluation of the proposed Video Generation Agents (VGAs), we adopt an evaluation protocol that is fully decoupled from the reward functions used during training. Specifically, we assess generated videos from three complementary perspectives: *instruction-level interest alignment score*, *automated multi-dimensional quality evaluation*, and *human preference assessment*. All evaluation scores are computed using external judges or human annotators, ensuring that the reported results reflect generalization quality rather than optimization-specific reward fitting.

Interest Alignment Score. We first evaluate whether generated videos faithfully follow their corresponding video production instructions derived from interest SIDs. To mitigate single-judge bias, we employ an ensemble of three state-of-the-art multimodal evaluators—GPT-5.1 [14], Gemini-2.5 Pro [4], and Claude-4.5 Sonnet [1]—each independently scoring the same benchmark of 1,000 multi-category video instances under the identical protocol defined in Box B.1. Each generated video is rated along five dimensions—semantic consistency, attribute accuracy, thematic alignment, completeness, and narrative coherence—producing a continuous alignment score in $[0, 1]$. We report the per-instance average across the three judges as the final Interest Alignment Score.

Automated Quality Evaluation. Beyond semantic alignment, we further evaluate the overall production quality of generated videos using the same three-judge ensemble—GPT-5.1 [14], Gemini-2.5 Pro [4], and Claude-4.5 Sonnet [1]—under the identical protocol defined in Box B.2. Each judge independently rates every video along four aspects: (1) instruction attractiveness, measuring hook quality, pacing, and call-to-action effectiveness; (2) BGM compatibility, evaluating music-tone consistency and beat synchronization; (3) SFX and sticker design quality, assessing visual effects and subtitle design; and (4) instruction–visual alignment, measuring the consistency between visual progression and instruction semantics. Per-instance scores are averaged across the three judges, yielding two metrics: a normalized **Automated Score** in $[0, 1]$ and an **Automated Win Rate** under the Good-Same-Bad (GSB) setting.

Human Preference Assessment. To further validate real-world perceptual quality and user preference alignment, we conduct a human evaluation study with 20 annotators from diverse professional backgrounds—including algorithm engineers, product managers, and advertising clients—covering both algorithm-side and business-side perspectives to reduce single-role bias. Each annotator performs 50 pairwise comparisons between generated videos and baseline results under a Good-Same-Bad (GSB) protocol, yielding 1,000 pairwise judgments in total. All comparisons are presented in a blind, randomized order, with each video pair independently evaluated by at least three annotators to mitigate individual subjectivity; we report the majority-vote outcome as the **User Study Win Rate**.

Box B.1: Interest Alignment Prompt

Role

You are an expert evaluator assessing the alignment between the video production instructions and the corresponding generated video. Focus only on semantic and creative consistency, while ignoring production quality (e.g., resolution, smoothness, or visual artifacts).

Inputs

- **Instructions**: {instructions}
- **Video**: {video}

Task

Evaluate the video along the following five dimensions, each scored in $[0, 1]$.

[A1] Content Fidelity

- Subjects, actions, and scenes match the instruction.

[A2] Attribute Accuracy

- Visual attributes and spatial-temporal relationships are correctly represented.

[A3] Intent & Theme Alignment

- Creative intent, mood, and stylistic cues align with the instruction.

[A4] Completeness

- All key elements are included without hallucinated content.

[A5] Narrative Coherence

- The temporal progression and story flow remain coherent.

Scoring Scale

- 0.9–1.0: Perfect alignment
- 0.7–0.9: Minor deviations
- 0.5–0.7: Moderate deviations
- 0.3–0.5: Major missing elements
- 0.0–0.3: Largely unrelated

Output Format

```
{
  "content_fidelity": {"score": 0.0},
  "attribute_accuracy": {"score": 0.0},
  "intent_theme_alignment": {"score": 0.0},
  "completeness": {"score": 0.0},
  "narrative_coherence": {"score": 0.0},
  "overall_alignment_score": 0.0
}
```

Rules

Scores should be continuous in $[0, 1]$. The overall score is a holistic judgment rather than the arithmetic mean of sub-scores. Do not consider production quality unless explicitly required by the instruction.

Box B.2: Video Quality Assessment Prompt

Role
You are a professional short-video advertising evaluator assessing video quality from editing and audio-visual perspectives.

Task
Evaluate the video across four dimensions with a total score of 100 points.

[D1] Instruction Attractiveness (25)
- Hook quality: pain-point, suspense, benefit-first, or contrast design.
- Pacing and structure: logical progression without redundant segments.
- CTA effectiveness: clarity, urgency, and consistency with opening intent.

[D2] BGM Compatibility (25)
- Mood and tempo alignment with video content.
- Synchronization between cuts and music beats.
- Balanced audio volume and speech clarity.

[D3] SFX & Sticker Design (25)
- Effectiveness of transition, emphasis, ambient, and emotional SFX.
- Consistency and readability of subtitles, tags, arrows, and motion effects.

[D4] Instruction–Visual Alignment (25)
- Consistency between instruction keywords and visual content.
- Narrative flow and temporal coherence.
- Absence of visual-information gaps or dead-air segments.

Output Format

```
{
  "instruction_attractiveness": {"total": 0},
  "bgm_compatibility": {"total": 0},
  "sfx_sticker_design": {"total": 0},
  "instruction_visual_alignment": {"total": 0},
  "final_summary": {
    "total_score": 0,
    "grade": "S/A/B/C/D"
  }
}
```

Rules
The total score ranges from 0 to 100. Grades are defined as: S (90+), A (75+), B (60+), C (45+), and D (<45). All evaluations should be evidence-based and supported by specific visual or temporal observations.

C Details of Generative Recommendation Model

Our Generative Recommendation Model (GRM) follows an architecture similar to GR4AD [27].

In training, the model takes two inputs: (1) user context $\mathbf{C}$, consisting of static profile features $\mathcal{F}_{\text{prof}}$ (e.g., age, gender, region, device type) and multi-granularity behavior sequences $\mathcal{F}_{\text{seq}}$, where each interaction is encoded via a sparse embedding table into latent interest tokens; and (2) prefix SID sequence $(BOS, s^1_{\text{content}}, s^2_{\text{content}}, s^1_{\text{creative}}, s^2_{\text{creative}})$ derived from the target item. We formulate GRM as an autoregressive sequence modeling problem, predicting the full SID sequence conditioned on the prefix, where the generated SIDs serve as a discrete representation of user interests, optimized via token-level cross-entropy loss, followed by reinforcement learning fine-tuning with constrained GDPO to further improve user feedback and interest alignment.

Concretely, each SID token is retrieved from a sparse embedding table and projected into a 768-dimensional latent space, then processed by a 7-layer Transformer decoder (LazyDecoder) with hidden size 768, FFN size 3072, 12 attention heads, and vocabulary size 8192. FlashAttention [6] is adopted for efficient computation. We train the model on 8 GPUs with batch size 8192 using Adam ($lr = 1 \times 10^{-4}$). During inference, beam search (beam size 512) is used, achieving 130 QPS throughput.