RIGC: A RECOMMENDATION-INFORMED FOUNDATION MODEL FOR PERSONALIZED MULTIMODAL CONTENT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In real-world recommender systems, generating personalized multimodal content remains largely unexplored, and the readily available recipe is a cascaded pipeline: a recommendation model compresses user behavior into an intermediate representation, on which a separate frozen generator then conditions. Such a bottleneck interface is fundamentally lossy, as the user is invisible to the generator once compressed, and personalization reduces to rendering a static description. Closing this gap calls for an end-to-end structure where recommendation and content generation share one model, one representation. We therefore present **RIGC** (**R**ecommendation-**I**nformed **G**enerated **C**ontent), a foundation model realizing this design within a single mixture-of-transformers backbone, where user information persists in the shared attention context and flows losslessly into generation. Such unification is non-trivial, as the backbone must reconcile heterogeneous objectives spanning item retrieval, text generation, and image synthesis. To this end, we construct a large-scale industrial multimodal dataset pairing user behavior with item content, and design a multi-stage alignment curriculum that progressively instills recommendation capability into a native multimodal generative model. Aligning with real user preferences further poses a cross-modal credit assignment problem, as a single feedback signal entangles all modalities, for which we introduce Nested-GRPO to sample hierarchical rollouts per modality and credit each by its marginal advantage. Experiments on public and industrial benchmarks show that RIGC surpasses dedicated recommenders in retrieval accuracy and consistently outperforms cascaded and unified baselines in generation fidelity, quality, and personalization, with Nested-GRPO yielding further gains in conversion and user specificity. These results indicate that recommendation and generation are mutually reinforcing over a shared representation, positioning unified generative recommendation as a foundation for personalized content delivery.

1 INTRODUCTION

Modern recommender systems (Kang & McAuley, 2018; Rajput et al., 2023; Zhou et al., 2025b) personalize by selection, not creation: all content is produced in advance for average audiences, and the system merely decides which to show. Such a paradigm is particularly limiting in advertising, where user engagement rests not only on the item itself, but equally on how it is presented. A growing body of work on personalized content generation (Shen et al., 2024; Zhou et al., 2026; Cheng et al., 2026) therefore asks a natural next question: can the text and imagery presenting an item be generated for each user, guided by what their behavior reveals?

Existing answers to this question converge on a cascaded recipe, powered by rapid progress in LLM-based generative recommendation (Wang et al., 2024): a pretrained LLM (OpenAI, 2023; DeepMind, 2023; Yang et al., 2025) infers the user's interests from their interaction history, summarizes them as a natural-language description, and hands this prompt to an off-the-shelf text-to-image model (Esser et al., 2024). Although each stage reuses a strong pretrained component, their interface constitutes an information bottleneck: the generator observes only the prompt, never the interaction sequence itself. Such pipelines are lossy by construction, as once compressed, the user is invisible

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

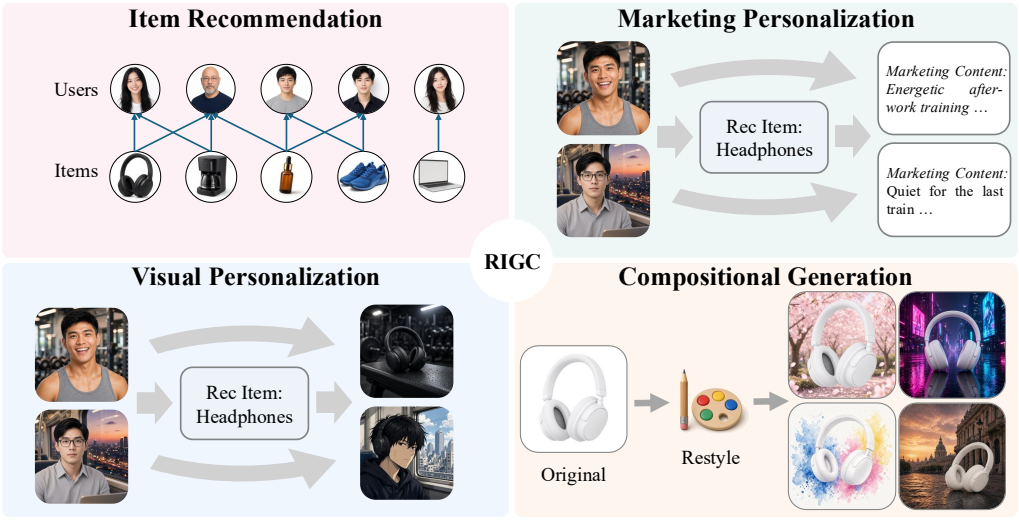


Figure 1: **RIGC unifies item recommendation with personalized multimodal generation.** Given a user's interaction history, RIGC first recommends the next item (top-left). For the same recommended item, it then generates personalized marketing text and visual content according to the user's information: marketing text is adapted to the user's lifestyle (top-right; e.g., energetic after-work training vs. a quiet last-train commute), and the keyframe is rendered in a scene and aesthetic matching the user's visual taste (bottom-left). The native visual head further restyles the recommended product into diverse compositional scenes while preserving item identity (bottom-right).

to the generator, and personalization reduces to rendering a static description. Since the deficiency lies in the cascaded structure rather than in any single component, the remedy must be architectural. Indeed, an analogous transition has already occurred in multimodal learning, where unified multimodal models (UMMs) (Chen et al., 2025; Cui et al., 2025; Deng et al., 2025; Diao et al., 2026) replaced cascaded understanding and generation with a single backbone over a shared representation, such that generation is conditioned on the model's full understanding rather than on what a prompt can convey.

Inspired by this transition, we unify recommendation and content generation in a single model, where both condition on the same user representation encoded from the raw interaction sequence and static profile. We instantiate this idea as **RIGC** (**R**ecommendation-**I**nformed **G**enerated **C**ontent), to our knowledge the first foundation model to unify recommendation and native multimodal generation. In one interleaved trajectory, RIGC decodes the next item and its marketing text, then renders a personalized keyframe through native flow matching.

Training such a unified model is non-trivial, as the backbone must reconcile heterogeneous objectives without eroding its pretrained multimodal knowledge. We therefore construct a large-scale industrial multimodal dataset pairing real user behavior with item content, and design a five-stage curriculum that progressively instills recommendation into the generative backbone. To further align RIGC with real user preferences, we apply reinforcement learning with a user-feedback simulator trained on interaction logs as the reward. This exposes a credit assignment problem unique to unified generation: vanilla GRPO (Shao et al., 2024) shares one advantage across item, text, and keyframe, so a strong item choice is penalized for a weak keyframe. We address this with Nested-GRPO, which samples the modalities in nested fashion and credits each by its marginal contribution.

The resulting model exhibits a spectrum of capabilities, as illustrated in Figure 1. From a user's behavior alone, RIGC selects the next item, composes marketing text adapted to the user's lifestyle, and renders a keyframe grounded in that item yet staged for the user's taste, so that the same product is presented in entirely different scenes to different users. Moreover, as RIGC is trained from a native multimodal model, it inherits general visual editing capabilities, restyling the keyframe across diverse genres with item identity preserved. Our experiments verify each capability in turn: RIGC

Table 1: **From cascaded pipelines to unified multimodal foundation modeling for personalized content generation.** ✓/○/✗: fully/partially/not supported. *Rec-Models* (recommendation models) predict *which* item to show but cannot synthesize its presentation; *Gen-Models* (multimodal generation models) synthesize high-fidelity content but carry no notion of user preference. Compositional generation denotes restyling a recommended item into diverse scenes while preserving identity. To our knowledge, RIGC is the first system closing this gap with native multimodal output, producing recommendation and generation jointly rather than via a cascaded pipeline.

	Rec-Models	Gen-Models	PMG	NextAds	NaviGen	Uni-AdGen	RaG	RIGC
Training Data								
Real-world User Feedback	✓	✗	✗	○	○	○	✓	✓
Industrial-Scale Data	○	✗	✗	✗	✗	✓	✓	✓
Model Capability								
Item Recommendation	✓	✗	✓	✗	✓	✗	✓	✓
Native Multimodal Generation	✗	✓	✗	✗	✗	✓	✗	✓
Compositional Generation	✗	✓	✗	○	✗	✓	✓	✓

surpasses dedicated recommenders in retrieval accuracy, outperforms cascaded and unified baselines in generation fidelity and personalization. Our contributions are summarized as follows:

• **Unified Model.** We present RIGC, to our knowledge the first foundation model to unify user modeling, item-level recommendation, and native multimodal generation within a single backbone, removing the information bottleneck of cascaded pipelines.

• **Data and Training Paradigm.** We construct a large-scale industrial multimodal dataset pairing real user behavior with item content, and devise a five-stage curriculum that reconciles cross-entropy and flow-matching objectives over a shared representation, followed by Nested-GRPO, which credits each modality by its marginal advantage under an offline proxy of online advertisement revenue.

• **Validation.** RIGC surpasses dedicated recommenders in retrieval accuracy, while its generated content consistently outperforms cascaded and unified baselines in fidelity and personalization, indicating that recommendation and generation are mutually reinforcing over a shared representation.

2 Related Work

2.1 Generative Recommendation and Personalized Content Generation

Generative recommendation reformulates retrieval as sequence generation: TIGER (Rajput et al., 2023), OneRec (Zhou et al., 2025b;c;a) and LC-Rec (Zheng et al., 2024) tokenize items into semantic identifiers (SIDs) and decode the next one directly, establishing the SID formulation we build upon. These models nonetheless remain bounded by a fixed inventory: they decide which item to surface, while its presentation is authored in advance for an average audience.

Personalized content generation lifts this bound by synthesizing the presentation itself, yet existing attempts each cover only part of the pipeline (Table 1). PMG (Shen et al., 2024) and NextAds (Xu et al., 2026b) follow the cascaded recipe: user preferences are summarized into prompts that steer frozen downstream generators. NaviGen (Zhou et al., 2026) moves further by coupling recommendation with native text generation, but still outsources imagery to an external text-to-image model. RaG (Cheng et al., 2026) translates predicted SIDs into a renderable script, which downstream text-to-image and text-to-video generators, together with an editing agent, then execute. Uni-AdGen (Xu et al., 2026a) suffers from the opposite limitation: it unifies image and text generation with industrial-scale data, yet conditions on a pre-selected item and thus performs no recommendation. RIGC instead internalizes the entire pipeline: recommendation and native multimodal generation share one backbone, requiring neither an external renderer nor a task-specific generator.

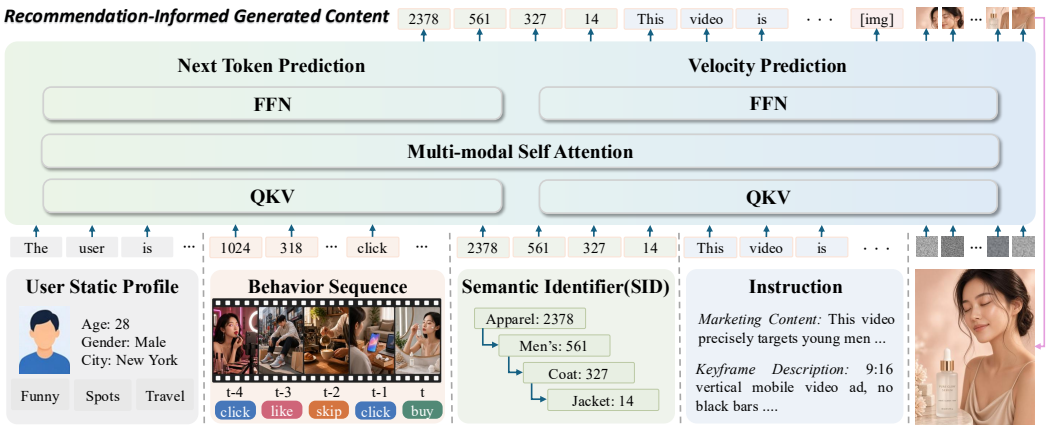


Figure 2: **Overview of RIGC.** A unified multimodal backbone serializes the user's static profile and interaction history into a single token stream, where each historical entry comprises the item's semantic identifiers (SIDs), user feedback, temporal features, and item metadata. Conditioned on this context, the model generates in one interleaved trajectory: (i) the four-level SID of the next item, (ii) a textual instruction comprising marketing content and a keyframe description, and (iii) a keyframe image from the native visual head. Later outputs attend to the full preceding context, so the instruction and keyframe condition on both the user context and the newly generated item, rather than on a detached prompt.

2.2 Unified Multimodal Models

Unified multimodal models (UMMs) replace pipelines of separate encoders and renderers with a single backbone in which understanding and generation share one representation. Purely autoregressive models such as Emu3.5 (Cui et al., 2025) and Janus-Pro (Chen et al., 2025) decode images over discrete visual tokens, which keeps the objective uniform but bounds fidelity by the codebook. The prevailing alternative pairs a mixture-of-transformers backbone with continuous visual generation: BAGEL (Deng et al., 2025) operates in a latent space, whereas SenseNova-U1 (Diao et al., 2026) performs flow matching directly in pixel space. Both families are nonetheless preference-agnostic, conditioning only on an instruction and having no notion of who receives the output, since they are trained without any real-world user feedback. RIGC treats a UMM as substrate rather than endpoint: we adopt the latter design and extend its shared representation with user behavior and SIDs, pursuing unification one level up by joining recommendation to understanding and generation.

3 Method

3.1 Problem Formulation and Task Protocol

User interest representation. Recommendation performs prediction based on user interest, which is evolving with time. For a user u, the user representation input x_u^τ at prediction time τ is jointly represented by static profile u_{pf} (e.g. gender, city...) and a history interaction sequence $\mathcal{H}_u^{<\tau}$:

$$x_u^\tau = (u_{\text{pf}}; \mathcal{H}_u^{<\tau}). \qquad \mathcal{H}_u^{<\tau} = (e_{\tau_1}, \ldots, e_{\tau_n}), \tau_1 \le \cdots \le \tau_n < \tau, \tag{1}$$

where each e_{τ_i} records an item (e.g. shoes, headphones...), the user's action (e.g. click, like, buy...), temporal features, and available item metadata.

Personalized multimodal content generation. The goal of RIGC is to not only predict the SID s of the user's next item, but also generate its marketing text c, description d and the visual keyframe I. Since recommendation tasks and multimodal generation tasks are inherently heterogeneous, we address this problem via a unified mixture-of-transformer structure as illustrated in Figure 2. Appendix E provides examples pairing user profiles and histories with the generated SID, marketing text, keyframe description, and keyframe.

3.2 PROGRESSIVE TRAINING PROCEDURE

To leverage the capabilities of off-the-shelf multimodal generative large model(Diao et al., 2026), we adopt a five-stage supervised finetuning (SFT) pipeline to align it with recommendation and personalized generation tasks.

Stage 1: Semantic interface initialization. To make the whole user representation x_u^τ as a tokenized context input, each item is represented by a SID $s = (z_1, \ldots, z_J)$, following the item-as-token procedure of generative recommendation (Rajput et al., 2023), where each code z_j at level j can represent a certain type of semantic in a coarse-to-fine manner (e.g Apparel→Mens→Coat→Jacket). These extra itematic codes serve as an interface to facilitate joint recommendation and multimodal-generation by enriching the vocabulary of existed text and visual tokens. We append 32,770 rows to both the input embedding and output projection. This preparation supplies a shared input–output interface for items while retaining the pretrained parameters.

Stage 2: Multimodal semantic alignment. We associate this interface with creative semantics through five tasks: $s \to c$, $s \to (c, d)$, $c \to s$, $(c, d) \to s$, and $(s, I) \to d$. The forward and inverse views connect item selection to verbalizable content, while the image-conditioned view grounds descriptions in the corresponding keyframe. In Stage 2A, only the new vocabulary rows are updated, embedding SIDs in the pretrained representation space. Stage 2B then adapts part of understanding blocks $\theta_{\mathcal{U}}$, where the rest generation parameters $\theta_{\mathcal{G}}$ remain fixed. We mix alignment tasks with auxiliary replay in Stage 2B to retain supervision for instruction following and perception. For discrete outputs, let q denote the input and b its target response. The autoregressive objective is

$$\mathcal{L}_{\mathrm{AR}} = -\frac{1}{\sum_{i\in\mathcal{T}} \alpha_i} \sum_{i\in\mathcal{T}} \alpha_i \log \pi_{\theta_{\mathcal{U}}}(b_i \mid q, b_{<i}), \tag{2}$$

where $\mathcal{T}$ contains valid assistant-response positions across the batch and α_i accounts for response length and stage-specific SID weighting. Input tokens are excluded from the loss.

Stage 3: History-conditioned recommendation. We next train $\pi_{\theta_{\mathcal{U}}}(s \mid x_u^\tau)$ on chronological user histories, turning semantic item representations into behavior-conditioned decisions. Each response contains the next SID, supervised by Eq. 2; the history obeys Section 3.1's temporal boundary. We expand adaptation to the final twelve understanding blocks, including their attention parameters, together with the projector, final normalization, and new vocabulary rows. Recommendation, alignment, and auxiliary replay examples are mixed to maintain the connection between behavioral prediction and item semantics.

Stage 4: Conditional keyframe generation. The next stage teaches the native visual branch to render the semantic conditions established above. We train all generation parameters $\theta_{\mathcal{G}}$ on five conditioning views over s, c, d and general text–image pairs, keeping the understanding branch and vocabulary fixed. For training pairs $(I, q) \sim \mathcal{D}$ with normalized target image I and condition q, we sample flow time t on $(0, 1)$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{1})$. Using the noisy image $\widetilde{I}_t = tI + (1-t)\sigma_R\epsilon$ with resolution-dependent noise scale σ_R, we minimize the flow-matching (Lipman et al., 2023) loss:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{I,q,t,\epsilon}\left[\left\| v_{\theta_{\mathcal{G}}}(\widetilde{I}_t, t, q, \sigma_R) - v^\star \right\|_2^2\right], \qquad v^\star = \frac{I - \widetilde{I}_t}{1-t}. \tag{3}$$

Here $v^\star$ is the target velocity, while $v_\theta = (\widehat{I}_\theta - \widetilde{I}_t)/(1-t)$ converts the model's clean-image prediction $\widehat{I}_\theta(\widetilde{I}_t, t, q, \sigma_R)$ to velocity.

Stage 5: Joint recommendation and generation. Finally, we unfreeze the full model and optimize the complete trajectory $s \to c \to d \to I$. Under teacher forcing, preceding target tokens condition subsequent predictions, and the image branch accesses the history and the supervised textual prefix. Text attention is causal; image patches within a generation block attend bidirectionally, while preceding text cannot attend to the noisy image. Thus, the target fields remain response supervision under the history-only protocol. The joint objective is

$$\mathcal{L}_{\mathrm{joint}} = 0.1\,\mathcal{L}_{\mathrm{AR}} + \mathcal{L}_{\mathrm{FM}}. \tag{4}$$

We mix complete business trajectories 1:1 with auxiliary understanding, generation, editing, and instruction data. Joint optimization lets image supervision shape the same context representation that supports recommendation and creative planning.

3.3 User Preference Modeling

To align the multimodal policy with real-world user preferences beyond imitation of logged behavior, we further optimize RIGC with reinforcement learning (RL).

User feedback simulator. We partition user preferences by interaction depth into three ordered interest levels: low (fast skip), medium (shallow interactions, e.g. click, like), and high (deep interactions, e.g. favorite, buy). The levels are ordered by business value because deep conversion events are the principal driver of recommendation system.

In order to score the generated multimodal content as the RL reward, we train a simulator (Zhao et al., 2023; Ie et al., 2019; Zhang et al., 2025; Liu et al., 2026) to predict the probability of these three user preferences given a user u and $\langle$item, text, keyframe$\rangle$ content triplet exposed to him (details in Appendix C). The simulator predicts a distribution p over the three interest levels y, which we convert into a scalar reward:

$$r = \sum_{y} w_y * p_y, y \in \{\text{low}, \text{medium}, \text{high}\}. \tag{5}$$

We set $w_{\text{high}} > w_{\text{medium}} > w_{\text{low}}$ so that shifting probability mass from a lower interest level to a higher one strictly increases r, which performs as a continuous offline proxy of online revenue.

Nested-GRPO. Vanilla GRPO (Shao et al., 2024) broadcasts a single group-relative advantage uniformly across every modality of a rollout. This sharing of advantage matches the case where r is determined by a single factor, as in reasoning tasks (by right or wrong) or conventional text-to-image tasks (by image quality). However, in real-world recommendation systems, r is determined by complicated factors. For instance, a user may click on an item merely because of his/her interest in the product category (e.g., headphones), whereas in other cases, the click behavior is driven by the user's preference for the stylistic presentation of the image itself. This can be reduced to the credit assignment problem, and details are given in Appendix D.

Therefore, we propose Nested-GRPO as illusrated in Algorithm 1. To isolate each modality's marginal contribution, we hierarchically sample groups of $K \times M \times L$ rollouts by conditional generation. Experiments show that this design better disentangles user interests across modalities, thereby facilitating preciser personalized generation.

Algorithm 1 Nested GRPO

Require: policy π_θ, user x_u^τ, reward model R_ϕ, group sizes K, M, L

for $k = 1, \ldots, K$ **do** ▷ hierarchical sampling
  sample sid $o_k^{\text{sid}} \sim \pi_\theta(\cdot \mid x_u^\tau)$
  for $m = 1, \ldots, M$ **do**
    sample text $o_{k,m}^{\text{txt}} \sim \pi_\theta(\cdot \mid x_u^\tau,\, o_k^{\text{sid}})$
    for $l = 1, \ldots, L$ **do**
      sample image $o_{k,m,l}^{\text{img}} \sim \pi_\theta(\cdot \mid x_u^\tau,\, o_k^{\text{sid}},\, o_{k,m}^{\text{txt}}), \quad r_{k,m,l} \leftarrow R_\phi(o_k^{\text{sid}},\, o_{k,m}^{\text{txt}},\, o_{k,m,l}^{\text{img}})$
    end for
  end for
end for

$V_k^{\text{sid}} \leftarrow \text{mean}(r_{k,\cdot,\cdot}), \quad V_{k,m}^{\text{txt}} \leftarrow \text{mean}(r_{k,m,\cdot})$ ▷ hierarchical baselines

$A_k^{\text{sid}} \leftarrow \big(V_k^{\text{sid}} - \text{mean}(\{V_k^{\text{sid}}\}_{k=1}^{K})\big)/\text{std}(\{V_k^{\text{sid}}\}_{k=1}^{K})$

$A_{k,m}^{\text{txt}} \leftarrow \big(V_{k,m}^{\text{txt}} - \text{mean}(\{V_{k,m}^{\text{txt}}\}_{m=1}^{M})\big)/\text{std}(\{V_{k,m}^{\text{txt}}\}_{m=1}^{M})$

$A_{k,m,l}^{\text{img}} \leftarrow \big(r_{k,m,l} - \text{mean}(\{r_{k,m,l}\}_{l=1}^{L})\big)/\text{std}(\{r_{k,m,l}\}_{l=1}^{L})$

Assign $A^{\text{sid}}, A^{\text{txt}}, A^{\text{img}}$ to sid / text / image tokens; update π_θ with GRPO objective

3.4 Training Data Construction

To train recommendation and creative generation jointly, we construct IndusAds-MM, a large-scale industrial multimodal advertising dataset derived from real-world advertising logs that pairs ad-exposure histories. The source logs for IndusAds-MM contain 200,000 users, 72,552,922 logged item occurrences, and 115,637,662 actions within those occurrences. Metadata and annotation

records cover 4,233,391 advertisements. For dataset splitting, we partition users into training, validation, and testing sets containing 155,000, 5,000, and 40,000 users, respectively. We adopt Qwen3-VL (Bai et al., 2025) to derive marketing understanding c from video frames, and first-frame description d from the keyframe. Raw pixels are used for image generation training. More details are shown in Appendix A.

4 EXPERIMENTS

We evaluate RIGC from three complementary perspectives: (1) multimodal personalized generation quality, covering instruction and keyframe fidelity as well as compositional restyling that preserves item identity; (2) recommendation accuracy as a standalone retriever; (3) the effect of Nested-GRPO for personalized content generation.

Table 2: **Multimodal personalized generation quality on public and industrial datasets.** Gen-Models: trained solely for generation; Unified: trained for recommendation and generation. BLEU-4/ROUGE-L measure instruction overlap; FID measures image distribution fidelity. CTS compares generated images with target descriptions using CLIP; CIS/DIS compare them with target images using CLIP/DINO. Joint metrics assess instruction–keyframe pairs: Pref. is the baseline's pairwise win rate minus RIGC's, expressed as a fraction (negative favors RIGC); Attr. is three-way user-attribution accuracy (chance 33.3%).

Method	**Type**	**A. Instruction**		**B. KeyFrame**				**C. Joint**	
		BLEU-4↑	ROUGE-L↑	FID↓	CTS↑	CIS↑	DIS↑	Pref.↑	Attr.↑
Amazon Office Products									
PMG	Unified	0.03	10.28	80.23	0.1335	0.5624	0.1064	-0.07	0.720
BAGEL	Gen-Models	0.10	12.61	48.66	0.1356	0.5641	0.1144	-0.17	0.750
SenseNova-U1	Gen-Models	0.40	14.38	70.98	0.1333	0.5514	0.1015	-0.06	**0.780**
Uni-AdGen	Gen-Models	7.18	25.63	85.64	0.1446	0.5656	0.0904	-0.35	0.680
NaviGen	Unified	6.38	24.85	56.10	0.1155	0.5025	0.0879	-0.04	0.770
RIGC (ours)	Unified	**7.79**	**27.02**	**47.79**	**0.1508**	**0.5809**	**0.1631**	—	**0.780**
IndusAds-MM									
PMG	Unified	0.01	4.68	100.04	0.1278	0.4259	0.0695	-0.73	0.625
BAGEL	Gen-Models	0.33	7.74	120.93	0.1504	0.4769	0.0755	-0.10	0.730
SenseNova-U1	Gen-Models	0.53	7.49	74.56	0.1553	0.4797	0.0822	-0.19	0.640
Uni-AdGen	Gen-Models	18.80	31.50	90.87	0.1746	0.5235	0.1132	-0.66	0.415
NaviGen	Unified	19.35	31.29	190.45	0.1805	0.5161	0.1393	-0.17	0.440
RIGC (ours)	Unified	**48.21**	**53.08**	**32.52**	**0.1840**	**0.6961**	**0.4062**	—	**0.755**

4.1 MULTIMODAL PERSONALIZED GENERATION QUALITY

Selecting the right item is insufficient if its creative depicts another product or ignores the user's history. We evaluate fidelity to the target creative and alignment with user preferences. We not only conduct experiments on our industrial dataset IndusAds-MM, but also evaluate on public benchmarks Amazon Office Product (Hou et al., 2024) by finetuning RIGC. Baselines include Uni-AdGen (Xu et al., 2026a), PMG (Shen et al., 2024), BAGEL (Deng et al., 2025), SenseNova-U1 (Diao et al., 2026), and NaviGen (Zhou et al., 2026). BAGEL and SenseNova-U1 assess the transfer of pretrained multimodal capabilities to personalized generation, inferring interests directly from user profiles and interaction histories. We use Qwen3.8-27B (Qwen Team, 2026) as the judge for pairwise preference (Pref.) and three-way user attribution (Attr.) of generated instruction–keyframe pairs.

Results. Table 2 shows that RIGC leads both domains on instruction overlap, keyframe fidelity, and target-item similarity. The IndusAds-MM gains are pronounced: BLEU-4 reaches 48.21 versus NaviGen's 19.35, and FID falls to 32.52 versus SenseNova-U1's 74.56. All baseline preference margins favor RIGC, which also achieves the highest user attribution on IndusAds-MM (0.755). The smaller gains on Amazon Office (e.g., FID 47.79 versus BAGEL's 48.66) may reflect its more limited training data, which offer fewer examples for learning the correspondence between user histories, items, and creatives.

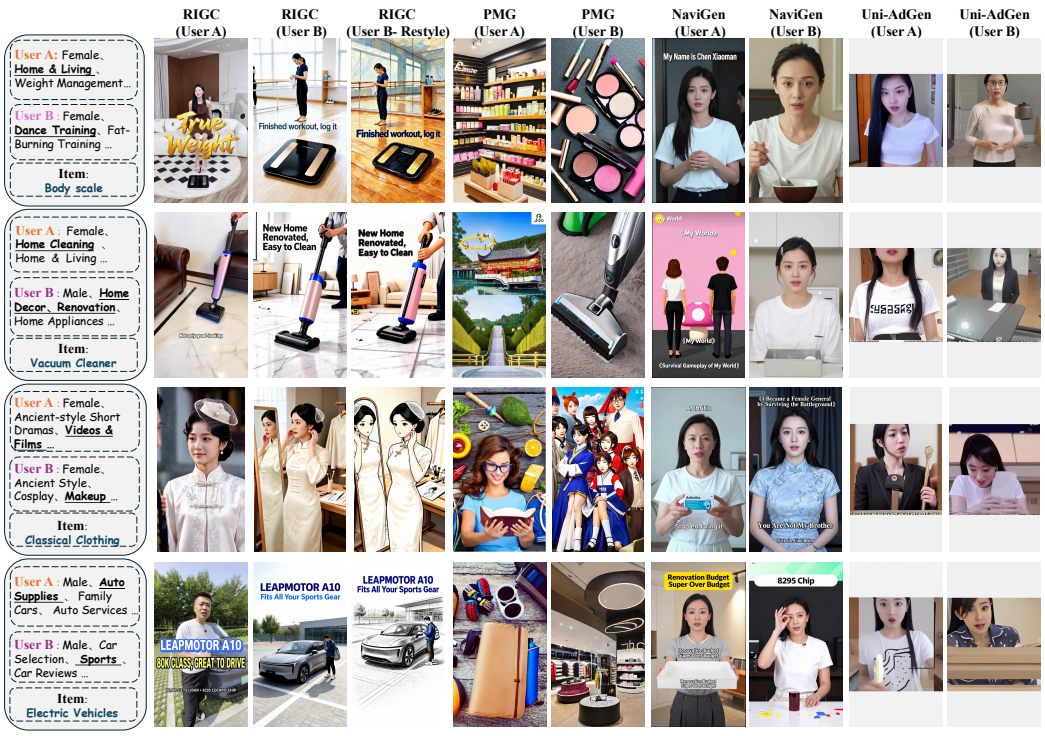


Figure 3: **Item selection, personalization, and restyling.** Each row lists a target item at left and compares creatives generated from two users' histories; the target item is not provided as an input. RIGC's Restyle column changes User B's image style while preserving the item: `oil_painting`, `claymation`, `anime`, and `pencil_sketch` from top to bottom.

BAGEL and SenseNova-U1 exhibit low lexical overlap but competitive joint preference and user attribution. In this transfer setting, their outputs express interests inferred from user context in wording that can differ from the logged instructions, explaining how personalization can coexist with low BLEU-4/ROUGE-L. NaviGen, meanwhile, approaches RIGC on IndusAds-MM CTS (0.1805 versus 0.1840) but trails in CIS (0.5161 versus 0.6961). This gap is consistent with a text interface conveying broad semantics while losing visual detail; RIGC's native visual branch retains access to both user history and the predicted item.

Qualitative analysis. Figure 3 examines whether this alignment holds in individual outputs. Given user histories, RIGC depicts the target item in the body-scale, vacuum-cleaner, classical-clothing, and electric-vehicle examples; several baseline outputs instead show generic presenters, storefronts, or unrelated products. These cases illustrate item consistency, while Table 3 measures recommendation accuracy across the test sets. For the same target within a row, RIGC also changes the presentation with the user: the scale appears in a home weight-management scene for User A and a post-workout scene for User B; the vehicle shifts from a family-oriented pitch to a sportier one. Finally, User B's item is restyled as an oil painting, claymation, anime, and pencil sketch from top to bottom, with the item remaining recognizable. The figure thus illustrates item selection, user-specific generation, and identity-preserving style control in one trajectory. These comparisons are complemented by input–output case in Appendix E and a broader gallery of subjects, compositions, and styles in Appendix F.

4.2 Recommendation Performance

A generated creative is useful only if the SID decoder selects the right item. To test whether joint multimodal training preserves retrieval, we compare RIGC with recommendation-only SASRec (Kang & McAuley, 2018), TIGER (Rajput et al., 2023), and OneRec (Zhou et al., 2025b), and the unified-text NaviGen (Zhou et al., 2026), on IndusAds-MM and Amazon Beauty, Sports, and Toys

Table 3: **Recommendation performance across industrial and public benchmarks.** Rec-Models: trained solely for recommendation; Unified: trained for recommendation and generation. We report Recall@10 (R@10) and NDCG@10 (N@10).

		IndusAds-MM		*Amazon Beauty*		*Amazon Sports*		*Amazon Toys*	
Method	**Type**	R@10↑	N@10↑	R@10↑	N@10↑	R@10↑	N@10↑	R@10↑	N@10↑
SASRec	Rec-Models	-	-	0.0639	0.0289	0.0389	0.0178	0.0658	0.0294
TIGER	Rec-Models	-	-	0.0628	0.0346	0.0331	0.0182	0.0527	0.0307
OneRec	Rec-Models	0.1010	0.0655	0.0883	0.0526	0.0548	0.0314	0.0931	0.0568
NaviGen	Unified	0.0430	0.0312	0.0838	0.0510	0.0497	0.0278	0.0865	0.0510
RIGC (ours)	Unified	**0.1300**	**0.0972**	**0.0937**	**0.0557**	**0.0553**	**0.0318**	**0.0939**	**0.0570**

(Hou et al., 2024). Recall@10 measures whether the next item enters the top-10 list; NDCG@10 also rewards a higher rank.

Table 3 shows that RIGC ranks first on every reported metric. Its largest gain over OneRec is on IndusAds-MM (R@10 0.1300 versus 0.1010; N@10 0.0972 versus 0.0655), while the gains on the three Amazon domains are consistent. NaviGen reaches only 0.0430 R@10 on IndusAds-MM, indicating that combining recommendation with text generation alone does not ensure strong retrieval. The larger IndusAds-MM margin is consistent with its aligned item, instruction, and keyframe supervision enriching the shared representation, as well as the large-scale industrial training data.

4.3 Ablation of Nested-GRPO

We resume from the same SFT checkpoint and compare Nested-GRPO (K=M=4, L=8) with vanilla GRPO (group size 128) under the same simulator. As illusrated in Table 4, SFT already yields a usable conversion signal (Reward 0.9772, DeepConv. 0.2234) and attributes a generation to its user well above chance (0.7553 vs. 33.3%). Vanilla GRPO barely moves either number (Reward 0.9791, DeepConv. 0.2283), and its preference against SFT stays at chance (0.5023). A shared advantage cannot credit SID, instruction, and keyframe separately, so the same reward produces almost no extra conversion uplift.

Table 4: **Ablation on the RL algorithm.** Nested-GRPO vs. vanilla GRPO from the same SFT checkpoint. **Reward**: test-set simulator reward r (Eq. 5, $w_{\text{low}} = 0$, $w_{\text{medium}} = 1$,$w_{\text{high}} = 2$); **DeepConv.**: mean deep-conversion probability p_{high}; **Pref.**: pairwise win rate vs. SFT (chance 50%; — for SFT); **Attr.**: 3-way user attribution (chance 33.3%).

Alignment Method	**Reward↑**	**DeepConv.↑**	**Pref.↑**	**Attr.↑**
SFT (w.o RL)	0.9772	0.2234	-	0.7553
Vanilla GRPO	0.9791	0.2283	0.5023	0.7593
Nested-GRPO	**0.9906**	**0.2390**	**0.5208**	**0.7784**

Nested-GRPO raises Reward to 0.9906 and DeepConv. to 0.2390, with preference 0.5208 and attribution 0.7784. The conversion gain comes with more user-specific outputs, rather than from a generic high-reward template, consistent with Appendix D. Overall, decomposing the advantage by modality turns the simulator reward into measurable gains in conversion and personalization.

5 Conclusion

We presented RIGC, to our knowledge the first foundation model to unify recommendation and native multimodal generation over a shared user representation. Within a single mixture-of-transformers backbone, RIGC decodes the next-item SID and marketing text, then renders a personalized keyframe via native flow matching. A five-stage curriculum reconciles cross-entropy and flow-matching objectives, and Nested-GRPO further credits each modality by its marginal contribution to a reward grounded in real user feedback. Across industrial and public benchmarks, RIGC surpasses dedicated recommenders in retrieval accuracy and outperforms cascaded and unified baselines in fidelity and personalization, indicating that recommendation and generation reinforce each other over one representation. With compositional restyling, video generation, and online alignment as immediate extensions, we hope RIGC marks a step toward recommender systems that no longer select from a fixed inventory, but create, for each user, content that no inventory could hold.

AI USE STATEMENT

We used generative AI tools only to aid and polish the writing and formatting of this manuscript. The authors reviewed the AI-assisted material and take responsibility for the final content, including its text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The model architecture, training objectives, and optimization procedure are described in Section 3. Training configurations and hyperparameters are provided in Appendix B, while the user-feedback simulator architecture and training details appear in Appendix C. Experimental settings, baselines, and evaluation metrics are described in Section 4.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Yanhua Cheng, Bo Wang, Haotian Zhang, Xinyuan Gao, Zhihui Yin, Ben Xue, Yongzhi Li, Jieting Xue, Ye Ma, Minquan Wang, Jiahui Li, Tianyu Xu, Zhiqiang Liu, Xiao Lin, Shiyang Wen, Changcheng Li, Liu Liu, Quan Chen, Peng Jiang, and Kun Gai. Recommendation as generation: Unifying personalized video generation and recommendation at industrial scale. *arXiv preprint arXiv:2606.25496*, 2026.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

DeepMind. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, Chengguang Lv, Haojia Yu, Haozhe Xie, Hongli Wang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jingcheng Ni, Junxiang Xu, Kaihuan Liang, Lianqiang Shi, Linjun Dai, Linyan Wang, Oscar Qian, Peng Gao, Pengfei Liu, Qingping Sun, Rui Shen, Ruisi Wang, Shengnan Ma, Shuang Yang, Siyi Xie, Siying Li, Tianbo Zhong, Xiangli Kong, Xuanke Shi, Yang Gao, Yongqiang Yao, Yves Wang, Zhengqi Bai, Zhengyu Lin, Zixin Yin, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. SenseNova-U1: Unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

Wang-Cheng Kang and Julian J. McAuley. Self-attentive sequential recommendation. In *ICDM*, pp. 197–206, 2018.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Fei Liu, Xinyu Lin, Hanchao Yu, Mingyuan Wu, Jianyu Wang, Qiang Zhang, Zhuokai Zhao, Yinglong Xia, Yao Zhang, Weiwei Li, Mingze Gao, Qifan Wang, Lizhu Zhang, Benyu Zhang, and Xiangjun Fan. RecoWorld: Building simulated environments for agentic recommender systems. In *WWW*, pp. 650–659, 2026.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Mahesh Sathiamoorthy. Recommender systems with generative retrieval. In *NeurIPS*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. PMG: Personalized multimodal generation with large language models. In *WWW*, pp. 3833–3843, 2024.

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 *CIKM*, pp. 2400–2409, 2024.

Yexing Xu, Wei Feng, Shen Zhang, Haohan Wang, Yuxin Qin, Yaoyu Li, Ao Ma, Yuhao Luo, Lu Wang, Xudong Ren, Haoran Wang, Run Ling, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, Longguang Wang, and Yulan Guo. Design your ad: Personalized advertising image and text generation with unified autoregressive models. In *CVPR*, pp. 472–483, 2026a.

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*, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zijian Zhang, Shuchang Liu, Ziru Liu, Rui Zhong, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Qidong Liu, and Peng Jiang. LLM-powered user simulator for recommender system. In *AAAI*, volume 39, pp. 13339–13347, 2025.

Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. KuaiSim: A comprehensive simulator for recommender systems. In *NeurIPS*, volume 36, pp. 44880–44897, 2023.

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 *ICDE*, pp. 1435–1448, 2024.

Guorui Zhou, Honghui Bao, Jiaming Huang, Jiaxin Deng, Jinghao Zhang, Junda She, Kuo Cai, Lejian Ren, Lu Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rongzhou Zhang, Ruiming Tang, Shiyao Wang, Wuchao Li, Xiangyu Wu, Xinchen Luo, Xingmei Wang, Yifei Hu, Yunfan Wu, Zhanyu Liu, Zhiyang Zhang, Zixing Zhang, Bo Chen, Bin Wen, Chaoyi Ma, Chengru Song, Chenglong Chu, Defu Lian, Fan Yang, Feng Jiang, Hongtao Cheng, Huanjie Wang, Kun Gai, Pengfei Zheng, Qiang Wang, Rui Huang, Siyang Mao, Tingting Gao, Wei Yuan, Yan Wang, Yang Zhou, Yi Su, Zexuan Cheng, Zhixin Ling, and Ziming Li. OpenOneRec technical report. *arXiv preprint arXiv:2512.24762*, 2025a.

Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, Weifeng Ding, Wuchao Li, Xinchen Luo, Xingmei Wang, Zexuan Cheng, Zixing Zhang, Bin Zhang, Boxuan Wang, Chaoyi Ma, Chengru Song, Chenhui Wang, Di Wang, Dongxue Meng, Fan Yang, Fangyu Zhang, Feng Jiang, Fuxing Zhang, Gang Wang, Guowang Zhang, Han Li, Hengrui Hu, Hezheng Lin, Hongtao Cheng, Hongyang Cao, Huanjie Wang, Jiaming Huang, Jiapeng Chen, Jiaqiang Liu, Jinghui Jia, Kun Gai, Lantao Hu, Liang Zeng, Liao Yu, Qiang Wang, Qidong Zhou, Shengzhe Wang, Shihui He, Shuang Yang, Shujie Yang, Sui Huang, Tao Wu, Tiantian He, Tingting Gao, Wei Yuan, Xiao Liang, Xiaoxiao Xu, Xugang Liu, Yan Wang, Yi Wang, Yiwu Liu, Yue Song, Yufei Zhang, Yunfan Wu, Yunfeng Zhao, and Zhanyu Liu. OneRec technical report. *arXiv preprint arXiv:2506.13695*, 2025b.

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. *arXiv preprint arXiv:2508.20900*, 2025c.

Hengji Zhou, Yufeng Liu, Ye Liu, Yong Xu, Lianghao Xia, and Liqiang Nie. Navigating user behavior toward personalized multimodal generation. *arXiv preprint arXiv:2606.24196*, 2026.

APPENDIX

A DATA CONSTRUCTION DETAILS

A.1 SOURCE DATA AND MULTIMODAL ANNOTATIONS

The source logs for IndusAds-MM contain 200,000 users, 72,552,922 logged item occurrences, and 115,637,662 actions within those occurrences. Metadata and annotation records cover 4,233,391 advertisement identifiers (`photo_id`). Offline Qwen-based services derive marketing understanding c from video frames, speech, and recognized text, and first-frame description d from the keyframe. These source media and transcripts create supervision but are not historical model inputs. We join s, c, d, I by `photo_id` and verify SID consistency. A SID may identify multiple advertisements, so it cannot by itself identify the matching text or image.

A.2 TEXT AND MEDIA FILTERING

Text fields must be nonempty, contain 50–1,000 characters, and pass checks for protocol fragments, control characters, repeated spans, and anomalous compression. Only pass-status annotations enter alignment training. We also exclude normalized text that maps to conflicting SIDs from text-to-SID tasks.

Keyframes must be readable, at least 256×256 pixels, and have aspect ratios in $[1/4, 4]$; near-black images are rejected. Of 2,602,115 candidate images, 2,592,082 pass these checks: 2,583,890 for training and 4,096 each for validation and testing. Validated media identities are reused across task views.

A.3 TEMPORAL CONSTRUCTION AND SPLIT PROTOCOL

Prediction points and visible history. We order each user's occurrences by their earliest action time and use the last active calendar day in UTC+8 for prediction targets. Each prediction cutoff is the target occurrence's earliest action time. Only actions strictly before that cutoff may enter the history; occurrences at the same millisecond cannot appear in one another's context. We retain up to 256 recent mapped occurrences and remove the oldest first if the rendered context exceeds its token budget. The most recent 50 occurrences in joint examples may include available industry categories and product names; recommendation examples use industry categories. Recommendation contexts have a 16,384-token limit. Profile inputs use allowlisted demographic, geographic, interest, and category attributes, with raw user identifiers excluded.

History-only task protocol. The primary task is autonomous next-item recommendation and generation. At inference, the model receives only the history before τ, allowlisted profile attributes, and metadata for historical items. The target SID, text, image, item and media identifiers, complete assistant response, and future interactions are withheld. In joint supervised training, the target response appears only on the assistant side as output supervision, including teacher-forced response prefixes, never in the input context. During evaluation, targets remain hidden until generation finishes and are then provided only to scorers. The response is serialized as $s \to c \to d \to I$, with the image marker after all textual fields.

User and item splits. Recommendation training, validation, and testing contain 155,000/5,000/40,000 users, respectively, assigned by deterministic salted hashing. Joint training combines the first two groups into 160,000 training users; the former 40,000 test users are divided into 5,000 validation and 35,000 test users. Final validation and test users are therefore absent from both training phases. Alignment reserves approximately 2% of SIDs for validation. Keyframe validation and test sets each contain 2,048 images with SIDs absent from keyframe training and 2,048 new creatives with keyframe-training SIDs.

A.4 TASK VIEWS AND TRAINING MIXTURES

Semantic alignment and recommendation. Table 5 gives the five alignment views and their Stage 2 sampling shares. SID-to-text views sample a SID group and one associated creative; other

views sample advertisement records. Stage 2B mixes alignment with auxiliary replay at 80:20. Stage 3 samples next-SID, alignment, and replay examples at 60:20:20; its five alignment views use 35/5/25/25/10% in table order.

Table 5: Alignment pools and Stage 2 sampling shares.

Input → output	Pool unit	Entries	Share
$s \to c$	SID group	2,434,056	35%
$s \to (c, d)$	SID group	2,354,534	5%
$c \to s$	Advertisement	2,520,887	25%
$(c, d) \to s$	Advertisement	2,437,643	15%
$(s, I) \to d$	Advertisement	2,560,494	20%

Replay draws on visual question answering from The Cauldron, multi-image tasks from Mantis-Instruct, instructions from Infinity-Instruct, and Chinese advertisement OCR.

Conditional keyframe generation. The 2,583,890 training keyframes are assigned once to five views: $(s, d) \to I$, $d \to I$, $(s, d, c) \to I$, $(s, c) \to I$, and $s \to I$, in proportions 45/10/30/10/5%. We add 645,973 text–image pairs from Fine-T2I, giving 3,229,863 records in total.

Joint recommendation and generation. Joint business examples require a valid SID, both text annotations, and the matching first frame. The 1,141,831 complete examples are mixed 1:1 with auxiliary understanding, instruction, generation, and editing records selected deterministically without replacement. The latter include SEED-Data-Edit OpenImages and Unsplash image edits. The resulting 2,283,662 records are consumed once.

A.5 INDUSADS-MM EVALUATION-SET CONSTRUCTION

Evaluation cases come from held-out users with complete target tuples. We exclude users and source records from an earlier 50-case diagnostic set, then select at most one case per user by a deterministic ranking. The release contains 200 development and 3,000 test cases, with a 50-case subset nested in development. Serialized inference contexts and target records are stored in separate views; the release checks user disjointness and input–target separation for all 3,200 cases.

B TRAINING CONFIGURATION

Each stage starts from the preceding stage's weights. The trainable parameter groups are summarized in Table 6; Table 7 gives the optimization budgets.

Table 6: **Progressive training and supervision.** Scales are rounded: Stage 3 counts recommendation examples, while Stages 4 and 5 include auxiliary data. AR denotes next-token loss and FM pixel-space flow matching.

Stage	Supervision (scale)	Updated parameters	Loss
1. SID initialization	—	New SID rows initialized	—
2A. SID alignment	SID–text/image alignment	SID rows	AR
2B. Semantic adaptation	Alignment + auxiliary replay	SID rows, top-4 FFNs/norms, projector	AR
3. Recommendation	History → SID (1.94M)	SID rows, top-12 understanding blocks, projector	AR
4. Keyframe generation	SID/text → I (3.23M)	Generation blocks and image modules	FM
5. Joint optimization	History → (s, c, d, I) (2.28M; 1:1 business/auxiliary)	Full model	0.1 AR + FM

Table 7: **Optimization settings.** Stages 2A/2B count supervised tokens; later stages cover the stated data once. Learning rates are peak values.

Stage	Learning rate	Training budget / coverage
2A	10^{-4}	1.53×10^9 supervised tokens
2B	5×10^{-6}	1.90×10^9 supervised tokens
3	Blocks: 2×10^{-6}; projector: 10^{-6}; SID rows: 5×10^{-6}	1,944,998 recommendation examples, with alignment and replay
4	10^{-5}	3,229,863 image-generation records
5	2×10^{-5}	2,283,662 joint and auxiliary records

Optimization. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, zero weight decay, and gradient-norm clipping at 1.0. Stages 2A–4 use constant learning rates after warmup: none in Stage 2A, 57 million supervised tokens in Stage 2B, 7.36 million in Stage 3, and 96,896 generated images in Stage 4. Stage 5 packs up to 32,768 tokens and uses a cosine schedule after a 100-step initialization interval. We report the final raw weights.

Image objective and conditioning. Generation targets are normalized to $[-1, 1]$. For N image patches, we set $\sigma_R = \sqrt{N/64}$ and sample $t = \mathrm{sigmoid}(\eta)$ with $\eta \sim \mathcal{N}(-0.8, 0.8^2)$. The squared velocity error in Eq. 3 is averaged over patch channels, weighted by $1/\sqrt{N}$ per patch, and normalized across generated images. Stage 4 text-to-image training drops the text condition with probability 10%. Stage 5 applies the same text-condition drop to business examples; auxiliary tasks retain their own conditioning settings.

C Details of Simulator

C.1 Simulator Training

Network architecture. The simulator is a user–item two-tower network with a shared scoring layer, specified stage-wise in Table 8 (the Item Tower is the content tower of Sec. 3.3). The *input* stage feeds the user profile and a length-50 history to the user tower, and the candidate SID, instruction, and reference keyframe to the item tower. At *encoding*, all text is encoded by a shared pre-trained BGE-M3 encoder, SIDs by a shared learnable embedding, and the keyframe by a pre-trained SigLIP ViT-SO400M; every branch is mapped to a 1024-dimensional working size. The user *tower* is a 4-layer bidirectional Transformer over the profile and an 8-layer causal Transformer over the history; the item tower fuses SID with instruction by 16-head cross-attention and then gated-fuses the keyframe through a residual MLP. *Merge* concatenates the last history state with the mean-pooled profile into a 2048-dimensional user vector, while the item tower keeps a 1024-dimensional embedding. *Scoring* takes the element-wise product of the two projected vectors and predicts the three interest levels defined in Section 3.3: low, medium, and high. Pre-trained encoders account for $\approx$1.45B of the $\approx$1.6B total parameters; the remaining $\approx$150M (towers, SID embeddings, scoring) are trained from scratch.

Table 8: **Architecture of the user-feedback simulator.** Each row is one stage, from input to scoring. Building blocks are shown in brackets, with the numbers of blocks stacked. $a|b$ denotes the User-Tower and Item-Tower output dimensions when they differ. †BGE-M3 weights are shared across profile, history, and instruction; ‡SID embeddings are shared across towers.

layer name	output dim	User Tower	Item Tower
input	—	profile; history $\times 50$	SID; instruction; keyframe
encoding	1024	BGE-M3†; SID embed.‡	BGE-M3†; SigLIP-SO400M; SID embed.‡
towers	1024	$\begin{bmatrix}\text{16-head Attn, bi}\\ \text{FFN}\end{bmatrix}\times 4$ $\begin{bmatrix}\text{16-head Attn, causal}\\ \text{FFN}\end{bmatrix}\times 8$	$\begin{bmatrix}\text{CrossAttn, 16H}\\ \text{(SID, inst)}\end{bmatrix}$ $\begin{bmatrix}\text{GatedFusion}\\ \text{residual MLP}\end{bmatrix}$
merge	2048 | 1024	$\text{concat}(h^{\text{last}}_{\text{hist}}, h^{\text{mean}}_{\text{prof}})$	h_{item}
scoring	3	$(2048\to1024\to1024)\ \odot\ (1024\to1024)\ \to$ residual MLP $\to$ 3-class logits	

Training Setup. We train with a single AdamW optimizer (weight decay 10^{-4}) whose state is kept across five chronological daily epochs under DDP on 8 GPUs. The global batch size is 2,048 (256 per GPU; no gradient accumulation). A warmup–cosine schedule is applied over a 9,165-step horizon: a 500-step linear warmup from 1% of the peak learning rate to 10^{-4}, then cosine decay to 10^{-6}. Gradients are clipped to ℓ_2 norm 1.0, dropout is 0.1, and updates use mixed-precision AMP; BGE-M3 and SigLIP are loaded in bf16 and cast to fp32 before the towers. On the first epoch the two encoders are frozen and only the remaining ≈150M parameters are updated; from the second epoch they are unfrozen and share the same learning-rate curve.

C.2 Results

The simulator is trained as a three-class predictor of the interest levels in Section 3.3—low, medium, and high—on logged user feedback, and later serves as both the offline evaluation environment and the Nested-GRPO reward model. Figure 4 reports one-vs-rest AUC for each class together with the macro-average over 3,000 training steps. All four curves rise sharply in the first few hundred steps and then plateau. Low, the majority class, is easiest and saturates near 0.91; medium and the rarer high class converge more slowly, with higher variance, to about 0.89 and 0.88. Macro-average AUC stabilizes near 0.89, indicating that the simulator is sufficiently calibrated to provide dense, class-aware rewards for subsequent policy optimization and a test-set score as the offline counterpart of live online evaluation.

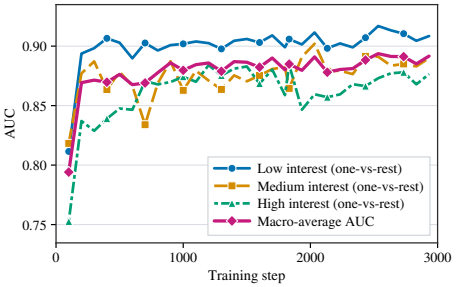


Figure 4: **Simulator training dynamics.** One-vs-rest AUC of the low, medium, and high classes and their macro-average over 3,000 training steps.

D Credit Assignment Problem

Vanilla GRPO scores each rollout with one scalar r and copies the same group-relative advantage A onto every token of that rollout—the SID, the instruction, and the keyframe. Table 9 shows why this is the wrong update when r is jointly determined by the three. Rollouts y_1 and y_2 have low r, so GRPO gives them a negative A; y_4 has high r, so it gets a positive A. The low reward of y_1, however, comes from a weak instruction: its SID and keyframe are already good. Because A is shared, the policy is nonetheless trained to emit that SID and that keyframe *less* often. The same happens in reverse for y_2: the instruction is good, yet it is down-weighted together with the weak SID and keyframe. The optimizer therefore cannot tell which component to keep and which to change. Correct modalities are unlearned, incorrect ones are reinforced whenever they ride along with a high-r triplet, and the same SID (or instruction, or keyframe) can receive opposite gradients

across rollouts—the SID of y_1 is punished, the same-quality SID of y_4 is rewarded. The resulting policy gradient is biased and high-variance, which slows RL training and can degrade modalities that were already reliable. We therefore propose Nested-GRPO: a nested group of $K \times M \times L$ rollouts isolates the marginal contribution of SID, instruction, and keyframe, and each modality is updated with its own group-relative advantage (Section 3.3).

Table 9: Four rollouts whose SID, instruction, and keyframe differ in quality. Vanilla GRPO sees only the scalar r of the full triplet, so it cannot tell *which* of the three components made the reward high or low. A checkmark (resp. cross) marks a high-quality (resp. low-quality) component.

	SID	Instruction	Keyframe	r
y_1	✓	✗	✓	low
y_2	✗	✓	✗	low
y_3	✓	✓	✗	medium
y_4	✓	✓	✓	high

E RIGC Input–Output Examples

Figures 5–7 illustrate RIGC's end-to-end personalized recommendation and generation through three input–output examples. Each figure pairs the user profile and interaction history with the SID, *Marketing Understanding*, *KeyFrame Description*, and generated image, making the connection between user context, item representation, and multimodal content visible.

For readability, we show selected profile fields and the last two of 256 interactions, abbreviate district names, and translate action labels into English. The text outputs retain their complete English translations; image captions are also displayed in English.

Case 1: Household repair. Agriculture, hardware, and home-related interests provide a plausible context for the leak-repair creative (Figure 5). The marketing text emphasizes ease of use and immediate results; the description and image express this practical appeal through a leaking pipe, spray cans, and a repair-oriented headline.

Case 2: Sneakers. Sports and apparel interests, reinforced by two recent clothing-store interactions, are consistent with the sneaker content (Figure 6). The text emphasizes materials and workmanship, while the description and image foreground the shoe's mesh texture, contrasting sole, and side logo in a product close-up.

Case 3: Cartoon hair accessory. Pet and anime interests offer a plausible stylistic connection to the cartoon hair clip (Figure 7), although recent drama-app interactions provide limited product-category evidence. The text emphasizes cuteness and companionship; the description and image convey this emotional appeal through the character-shaped clip, stacked close-ups, and a playful pink banner.

Together, these examples expose RIGC's complete output structure and show how practical, quality-oriented, and emotional appeals appear across text and images. They complement the quantitative evaluation in Section 4 and the visual gallery in Appendix F; individual examples alone do not establish recommendation accuracy or personalization gains.

01 Input

User Profile

Demographics: Male; age segment: 7; province / city: Jilin; district: Y.C.

Video interests · L1: Agriculture & Rural Life = 0.9363; Sports = 0.6647; Science & Law = 0.6586; Anime & Comics = 0.5833; Travel = 0.5133.

Shopping interests · L1: Auto & Farm Machinery Aftermarket = 0.471; Hardware & Electrical Equipment = 0.4693; Home & Living = 0.2298; Flowers / Plants & Agricultural Supplies = 0.2276; Outdoor / Fishing Gear = 0.2133. …

Interaction History … / last 2 of 256 interactions

255 [d=1, h=10] <sid_begin><sid_07923><sid_10870><sid_18241><sid_30696><sid_end>
Business Services / Industry & Agriculture · Ruixi Crop Supplies Store — Actions: Watched for 5 seconds

256 [d=1, h=10] <sid_begin><sid_06840><sid_15726><sid_22646><sid_29673><sid_end>
Shop / Auto Accessories · Chepu Official Flagship Store — Actions: Watched for 5 seconds

02 Text Output

SID: <sid_begin><sid_01091><sid_15740><sid_21019><sid_24694><sid_end>

Marketing Understanding

This video precisely targets ordinary households troubled by leaking homes, seepage through roofs, and rain leaking around window frames. Their main concerns are hassle, being ripped off, and paying for a repair professional to visit. The central marketing narrative therefore uses foolproof operation and immediate results to overcome their reluctance to tackle the problem.

The opening does not show the product directly. Instead, suspenseful English lyrics accompany the large headline "How to fix a house leak," creating anticipation that a mysterious solution is about to be revealed. Within three seconds, the contrast between an English song and a leaking-home problem captures attention. The middle section moves into a rigorous demonstration: first, an exaggerated high-pressure pipe rupture uses a stress test to demonstrate the product's pressure resistance; then the video moves to common leak sites such as roofs, window frames, and sheet-metal roofing. The visible "one spray stops the leak" effect builds trust, showing viewers that a simple action can solve a major problem and removing the perceived barrier of professional repair.

The ending reinforces credibility with a "not a fake experiment" disclaimer and uses brand terms such as "Leak-Repair King" and "Golden Armor" as memorable cues to encourage orders. The inferred background music is an English pop song with a strong beat. The commentary is calm and objective, with a hint of amazement, creating a dependable atmosphere of scientific verification and everyday usefulness that makes viewers feel, "This really works, and I can handle it myself."

03 Image Output



KeyFrame Description

A 9:16 portrait opening frame of a smartphone-shot advertisement, without black borders, uses an outdoor scene with overlaid headline text at the top. A medium shot is taken with the camera level at the subjects' waists. A large, thick black cylindrical pipe stretches horizontally across the center, supported by a metal stand, with water spraying downward from both ends.

Three main figures stand behind the pipe. On the left, a short-haired man in a white short-sleeved T-shirt and dark shorts holds a yellow can-shaped object in both hands and looks down at it. In the middle, a man wearing a black baseball cap, sunglasses, and a black-and-white vertically striped sleeveless top raises his right hand while his left arm hangs naturally. On the right, a woman with a high ponytail wears a gray athletic tank top, holds a red can-shaped object in her right hand, and lets her left arm hang naturally. All three wear sneakers and stand on wet ground. Blurred trees and other people standing farther away are visible behind them.

A large, bold yellow headline at the top reads "HOW TO FIX A HOUSE LEAK". Immediately below, bold white text reads "If only I had found this sooner". Both lines have black outlines and extend horizontally across the upper frame. Natural daylight gives the scene an overall warm tone. Water on the ground creates distinct reflections; faces and clothing are evenly lit, and the background trees are in natural light.

Figure 5: **Case 1: Household repair.** User context and RIGC outputs, with a practical marketing appeal expressed through a leak-repair demonstration.

01 Input

User Profile

Demographics: Male; age segment: 1; province / city: Chongqing; district: Y.T.M.A.C.

Video interests · L1: Sports = 1.4663; Gaming = 0.8665; Cars = 0.8644; Beauty = 0.5108; Appearance = 0.4298.

Shopping interests · L1: Personal Care & Cleaning = 0.3559; Makeup = 0.332; Apparel / Footwear = 0.321; Skincare = 0.3129; Phones / Digital / Computers & Office = 0.3058. …

Interaction History … / last 2 of 256 interactions

255 [d=0, h=15] <sid_begin><sid_04155><sid_12406><sid_22600><sid_28142><sid_end>
Shop / Clothing & Headwear · Hanlichen Flagship Store — Actions: Watched for 5 seconds

256 [d=0, h=16] <sid_begin><sid_02691><sid_09413><sid_22823><sid_31202><sid_end>
Shop / Clothing & Headwear · Fetse Fashion Store — Actions: Watched for 5 seconds

02 Text Output

SID: <sid_begin><sid_03001><sid_10388><sid_17297><sid_31685><sid_end>

Marketing Understanding

This video targets young consumers who follow fashion trends but worry about counterfeit or poor-quality products. Their main concerns are making a bad purchase and being ripped off, so the marketing narrative builds strong trust through forceful unboxing and close-ups of product details.

The opening gets straight to the point: hands forcefully tear open the shoebox, accompanied by an exclamation of "Damn!" This creates an immediate impression of exceptional shoe quality, using an unexpected stress test to capture attention within three seconds. The middle section moves through product details: the mesh upper and tongue branding, close-ups of the outsole tread and side Swoosh, then fingers pressing the shoe to demonstrate rebound. Step by step, doubts about authenticity give way to reassurance about solid materials and fine workmanship. The combination of stress testing and detail shots reinforces perceived quality.

At the end, the presenter faces the camera and delivers a rapid sales pitch emphasizing "affordable" and "grab it now," using price and urgency to move hesitant viewers toward ordering. The inferred background music is energetic, with a strong beat and electronic or hip-hop elements. The excited, exaggerated, down-to-earth delivery creates both the delight of discovering a hidden gem and an impulsive sense of missing out by not buying.

03 Image Output



KeyFrame Description

A 9:16 portrait opening frame of a smartphone-shot advertisement, without black borders, set indoors. An eye-level close-up shows the side of a sneaker. A hand that appears to belong to someone around twenty enters from the lower left, gripping the heel and holding the shoe slightly left of center. The skin tone is natural, the nails are neatly trimmed, and part of the forearm is visible at the wrist.

The sneaker is black and white, with a predominantly black mesh upper, a white Nike Swoosh on the side, black laces, a white midsole, and a black outsole. Black reinforcements appear at the toe and heel, and "NIKE" lettering is visible on the side. An open shoebox lined with white tissue paper sits behind it, against a light gray wall with black line decorations above.

A large white sans-serif caption across the top reads "Damn! The quality of these shoes is amazing!!", extending horizontally above the shoebox. Light from above creates distinct highlights on the shoe and hand, with soft shadows. The overall color tone is cool, and the image is sharp.

Figure 6: **Case 2: Sneakers.** User context and RIGC outputs, connecting sports and apparel interests with a quality-oriented product presentation.

01 Input

User Profile

Demographics: Female; age segment: 5; province: Beijing; city: Beijing; district: F.D.

Video interests · L1: Celebrity & Entertainment = 1.3905; Pets = 0.9684; Anime & Comics = 0.7327; Education = 0.6085; Sports = 0.3864.

Shopping interests · L1: Home & Living = 0.4253; Apparel / Footwear = 0.383; Fresh Food = 0.2834; Children's / Baby / Family Matching Clothing = 0.2828; Snacks / Nuts / Regional Specialties = 0.2817. …

Interaction History … / last 2 of 256 interactions

255 [d=1, h=19] <sid_begin><sid_05798><sid_08827><sid_17135><sid_25739><sid_end>
Media & Information / Audio & Video Apps · Midnight Drama
Actions: Watched for 5 seconds

256 [d=1, h=21] <sid_begin><sid_07413><sid_12248><sid_23109><sid_28830><sid_end>
Media & Information / Audio & Video Apps · Midnight Drama
Actions: Watched for 5 seconds · Clicked product

02 Text Output

SID: <sid_begin><sid_05180><sid_14131><sid_17057><sid_28027><sid_end>

Marketing Understanding

This video precisely targets young women and students who seek a sense of occasion in everyday life and love cute things. They no longer choose hair accessories solely for their traditional functions; they also seek the emotional value of cuteness. The entire video uses a soft-sell approach combining product recommendations in everyday settings with emotional resonance. It avoids overt sales pitches and adopts the tone of sharing a good find to lower viewers' defenses.

The opening immediately addresses a pain point. A pink caption bar accompanies the playfully pleading voice-over, "What's wrong with buying me one?" This instantly creates the sweet feeling of being doted on by a friend. The cute, slightly bossy tone captures attention within three seconds and prompts viewers to imagine, "I want to be treated like that too." The middle section presents the product through a montage of different settings. First, close-ups show the hair clip's soft, plush texture. Next, a demonstration shows how adorably doll-like it looks when clipped into hair. Multiple colors and desktop arrangements follow. The exaggerated claim "You won't find anything this cute anywhere in heaven or on earth" reinforces a sense of scarcity. It redefines the clip from a practical tool into a little corner-dwelling creature that can play with you, giving the product the role of a companion.

At the end, the social cue "My good friend, I'm tagging you" encourages sharing and conformity. The rhetorical question "You'll get one for me too, won't you?" provides the final nudge, triggering a purchase impulse through interaction. Overall, the inferred background music is light, soothing Japanese-style vlog music. The voice-over sounds cute, intimate, and a little pleading, like a close female friend. Together with bright, soft filters and close-ups, it creates a warm, comforting, girlish atmosphere that unconsciously makes viewers want to own a piece of this cuteness.

03 Image Output



KeyFrame Description

A 9:16 vertical, smartphone-filmed opening advertisement frame with no black borders. The image consists of two stacked panels showing successive states of the same scene. A close-up focuses on the top of a person's head in an indoor setting. In the upper panel, a white, round-headed cartoon hair clip stands upright at the center of the crown. Its face has small black dots for eyes and a mouth, and its arms are folded. A long, black, strip-like object resembling a hair tip extends from the right side of its body, curving toward the back-right. The person's hair is dark brown and smooth, with the scalp visible at the roots.

A pink, rounded rectangular banner is overlaid across the middle of the image. Its centered white text reads, "What's wrong? Buy me one. What's wrong with that?" The lower panel has essentially the same composition as the upper one. The clip sits slightly lower, the black hair tip has a more pronounced curve, and the clip's facial expression is clearly visible. Light-colored curtains with vertical stripes are behind the person's head, blurred out of focus. A line of extremely small white text reading "Advertising content is subject to actual conditions" appears along the bottom edge of the lower panel. The lighting is soft and even, with a warm color palette. The hair has a natural sheen, while the clip has a matte surface with no obvious reflections.

Figure 7: **Case 3: Cartoon hair accessory.** User context and RIGC outputs, with text and imagery emphasizing cuteness and companionship.

F ADDITIONAL VISUAL RESULTS OF RIGC

Figures 8–9 present 90 RIGC-generated images in three galleries of 30 examples each. The collection spans everyday scenes, product promotions, and gaming content, with varied compositions and photographic, illustrated, and stylized renderings. These examples provide a qualitative view of RIGC's visual range, complementing controlled comparisons in Figure 3. User-specific alignment and comparative performance are evaluated through paired examples and quantitative results in Section 4.

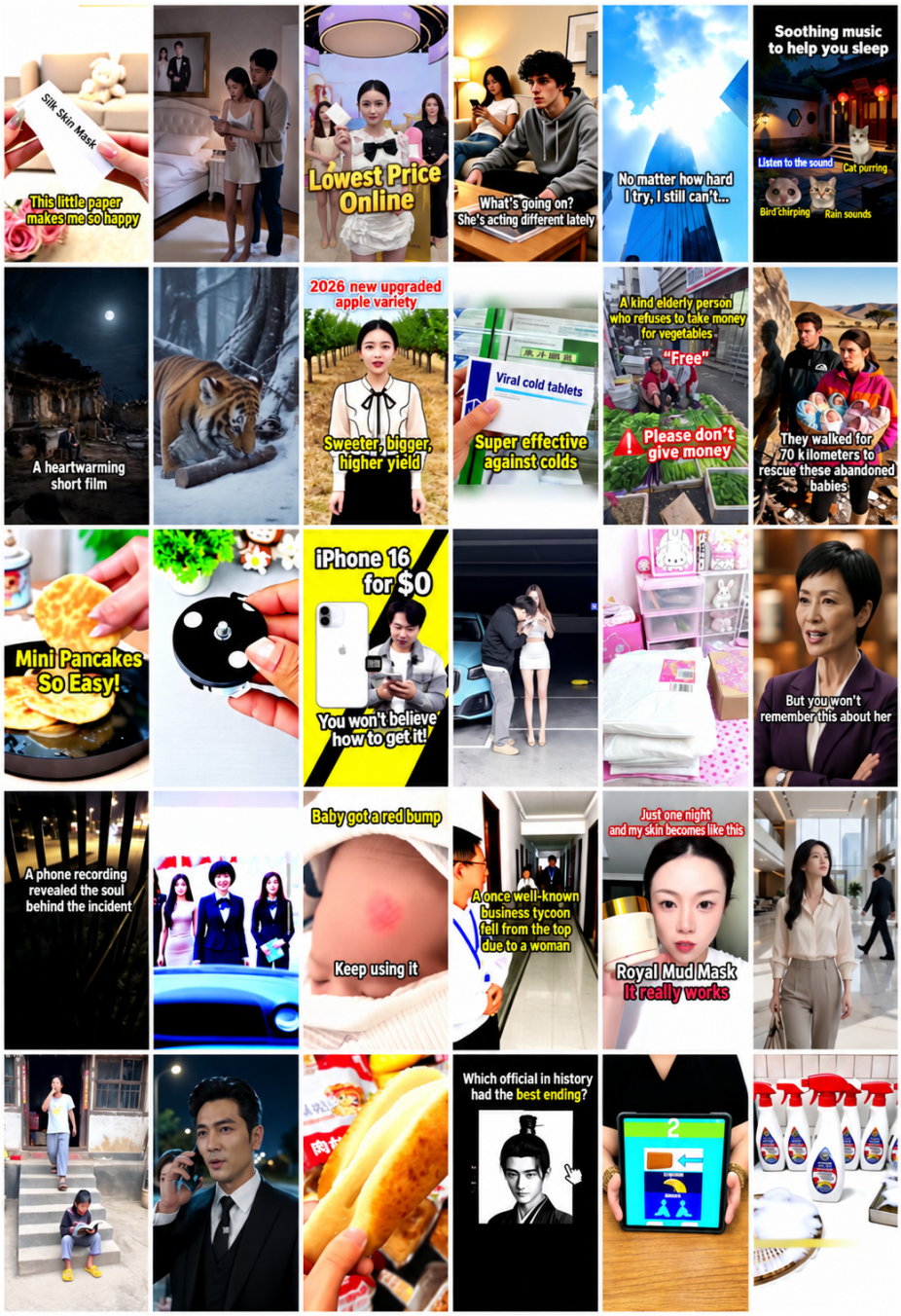


Figure 8: **RIGC-generated visual examples in portrait format.** Thirty examples arranged in five rows and six columns depict everyday scenes, narrative content, animals, food, and promotional creatives, illustrating varied subjects and compositions.

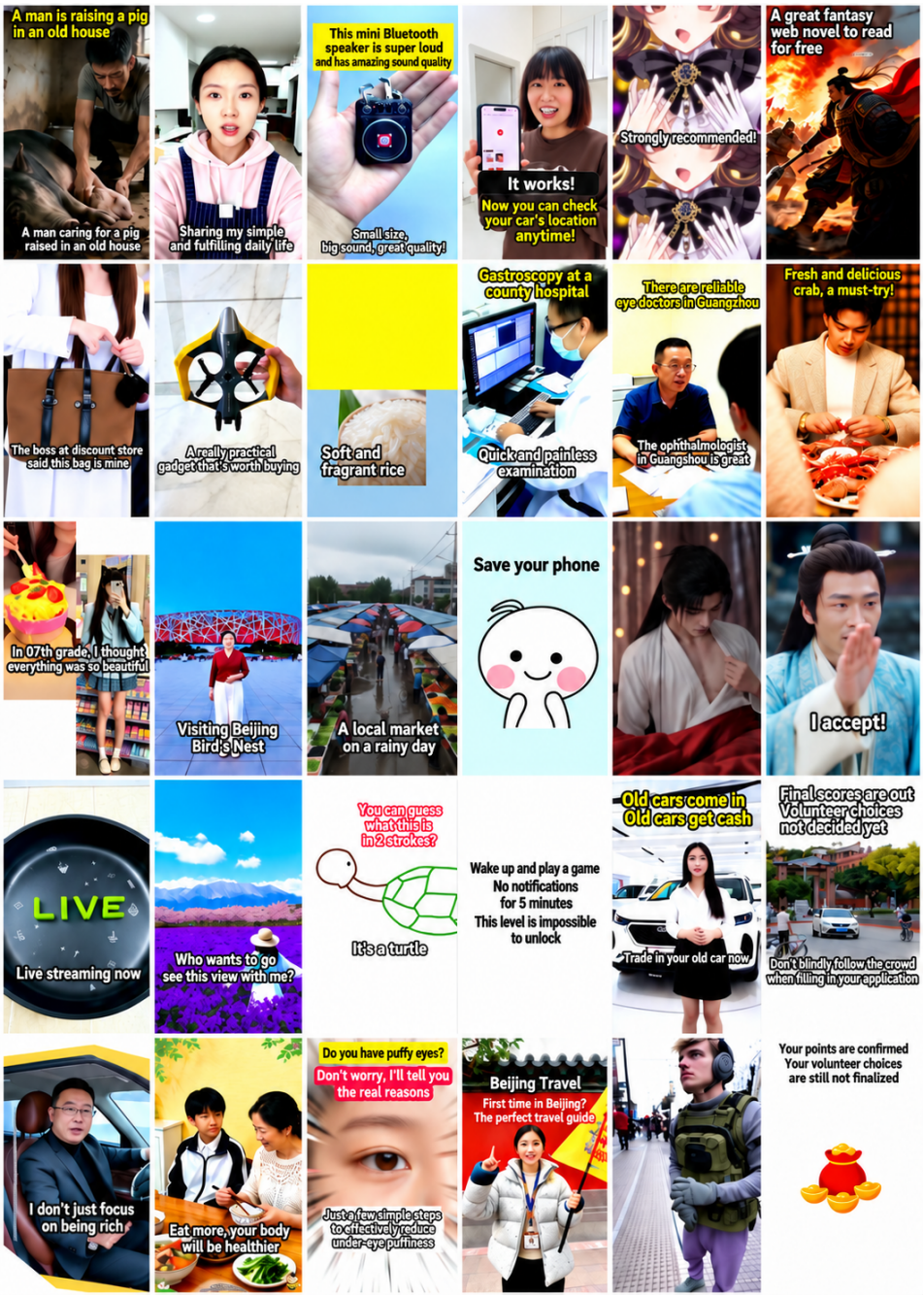


Figure 9: **Additional RIGC-generated visual examples in portrait format.** Thirty examples arranged in five rows and six columns cover product presentations, daily life, travel, entertainment, and service promotions, complementing the preceding galleries.

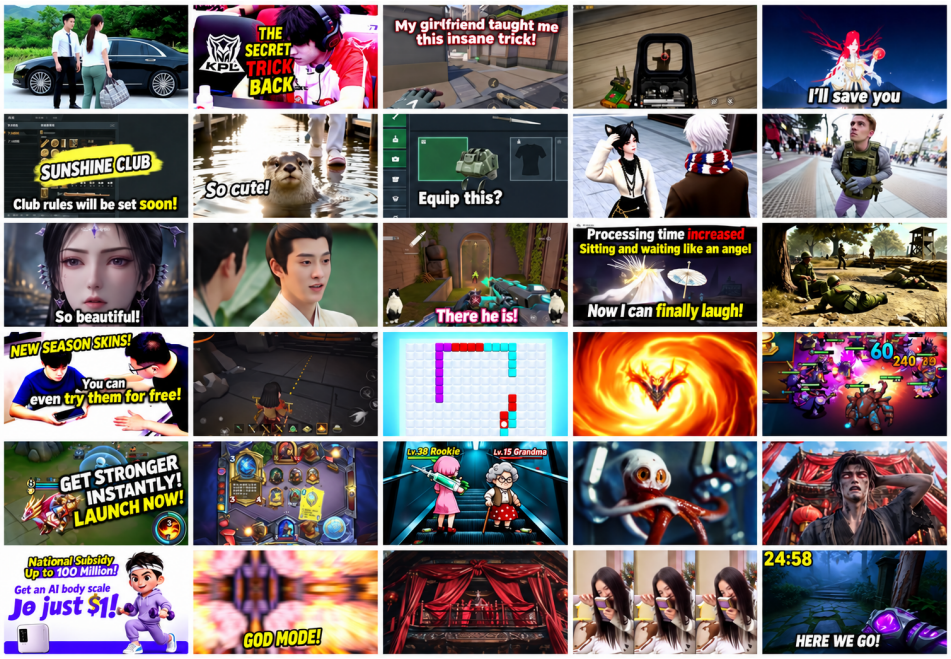


Figure 10: **RIGC-generated visual examples in landscape format.** Thirty examples arranged in six rows and five columns include gaming scenes, fantasy characters, stylized animation, and live-action-style imagery, illustrating varied rendering styles and scene layouts.