ICG: Improving Cover Image Generation via MLLM-based Prompting and Personalized Preference Alignment

Zhipeng Bian[1,2], Jieming Zhu[2*], Qijiong Liu[3], Wang Lin[4], Guohao Cai[2], Zhaocheng Du[2], Jiacheng Sun[2], Zhou Zhao[4], Zhenhua Dong[2]

[1]Huazhong University of Science and Technology [2]Huawei Noah's Ark Lab
[3]Hong Kong Polytechnic University [4]Zhejiang University
bian_zhipeng@hust.edu.cn jiemingzhu@ieee.org liu@qijiong.work
{caiguohao,duzhaocheng,sunjiacheng,dongzhenhua}@huawei.com
{linwanglw,zhaozhou}@zju.edu.cn

Abstract

Recent advances in multimodal large language models (MLLMs) and diffusion models (DMs) have opened new possibilities for AI-generated content. Yet, personalized cover image generation remains underexplored, despite its critical role in boosting user engagement on digital platforms. We propose ICG, a novel framework that integrates MLLM-based prompting with personalized preference alignment to generate high-quality, contextually relevant covers. ICG extracts semantic features from item titles and reference images via meta tokens, refines them with user embeddings, and injects the resulting personalized context into the diffusion model. To address the lack of labeled supervision, we adopt a multi-reward learning strategy that combines public aesthetic and relevance rewards with a personalized preference model trained from user behavior. Unlike prior pipelines relying on handcrafted prompts and disjointed modules, ICG employs an adapter to bridge MLLMs and diffusion models for end-to-end training. Experiments demonstrate that ICG significantly improves image quality, semantic fidelity, and personalization, leading to stronger user appeal and offline recommendation accuracy in downstream tasks. As a plug-and-play adapter bridging MLLMs and diffusion models, ICG is compatible with common checkpoints and requires no ground-truth labels during optimization.

1 Introduction

Large language models (LLMs) and diffusion models (DMs) have driven the rise of AI-generated content (AIGC) in applications such as personal assistants, chatbots, digital art, and cover image generation (Omneky, 2024; Jarsky et al., 2024; Yang et al., 2024). In recommender systems—especially news feeds—blurry, mismatched, or unappealing covers are common, undermining user engagement. Thus, improving cover image generation is critical to enhancing recommendation quality.

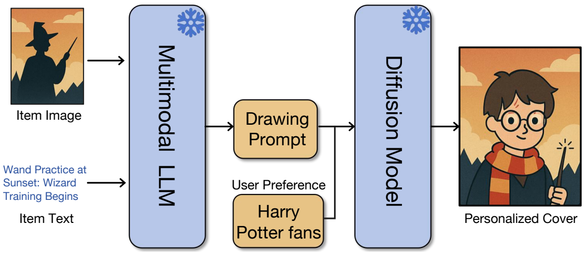

Figure 1: The overall pipeline for cover image generation.

Text-to-image models such as Stable Diffusion (Rombach et al., 2022), Midjourney (mid, 2022), and DALLE-3 (Betker et al., 2023) are widely used by designers and publishers for banner and cover image generation. However, they rely heavily on manually crafted prompts and careful prompt engineering, which limits scalability for platforms handling millions of items, such as news aggregators, streaming services (e.g., Netflix, YouTube), and social media feeds (e.g., TikTok, Instagram). In these scenarios, visually appealing and context-relevant cover images are critical for capturing user attention and improving engagement. As shown in Figure 1, a promising solution is to use multimodal large language models (MLLMs) to automatically extract semantics from raw item content and generate prompts for Stable Diffusion. Despite its simplicity, this pipeline faces several challenges in practical adoption.

Firstly, although MLLM-based prompt generation eliminates the need to manually craft prompts for each item, it still requires careful design of prompt instructions for the MLLMs. Prior works such as BeautifulPrompt (Cao et al., 2023), Promptist (Hao et al., 2023), and UF-FGTG (Hei et al.) aim to automate or refine prompts using large language models, but they focus on improving existing prompt text. In contrast, our task starts from raw

[*] Corresponding Author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12268–12278
November 4-9, 2025 ©2025 Association for Computational Linguistics

item content (e.g., titles), rendering these methods inapplicable. Furthermore, the absence of golden prompt references for cover images limits the possibility of supervised fine-tuning for MLLMs in this setting.

Secondly, the current pipeline is disjointed and lacks end-to-end optimization, leading to issues such as MLLM hallucinations and misalignment with diffusion models, which often result in low-quality or semantically irrelevant covers. This hinders error correction and model refinement. Recent progress in multimodal AI has produced models like MiniGPT-5 (Zheng et al., 2023), SEED-LLaMA (Ge et al., 2024), and Kosmos-G (Pan et al., 2024), which integrate MLLMs with diffusion decoders for unified understanding and generation. However, they still rely heavily on user-crafted prompts. In addition, the scarcity of high-quality cover images limits supervision when generating directly from raw item content.

Thirdly, current text-to-image generation methods lack personalization, often producing covers that fail to reflect user preferences and reduce engagement. For instance, male users may prefer dark, professional styles, while female users may favor pink, cute designs. Aligning covers with individual tastes can boost click-through rates and user experience. Prior work like PMG (Shen et al., 2024) and DiFashion (Xu et al., 2024) explores this direction but has key limitations: (1) Both use the next item's image as the training target, assuming high-quality covers—often untrue in practice; (2) PMG represents preferences as discrete keywords via LLMs, hindering end-to-end optimization and fine-grained preference capture. Consequently, these methods often fail to generate visually appealing, truly personalized outputs.

To address these challenges, we propose ICG, a unified framework for personalized cover generation that integrates MLLMs with reward-based optimization. It leverages item content—comprising a reference image and title—to retain original semantics, while personalization is guided by user interaction histories. Textual and visual inputs are encoded by MLLMs, with meta tokens capturing contextual features that are injected into the diffusion model via an adapter for end-to-end training. User features are fused with context to condition generation. The model is optimized using a differentiable multi-reward framework, combining public aesthetic and relevance scores with a personalized reward model trained on user-item interactions, enabling content-aligned and user-specific generation.

The main contributions of this work are:

(1) We present the first framework that integrates MLLMs with reward learning for personalized cover image generation, demonstrating its effectiveness in recommendation scenarios.

(2) We introduce meta tokens to capture contextual semantics and fuse them with user embeddings via a plug-and-play adapter into a diffusion model. A multi-reward learning framework enables end-to-end training guided by aesthetics, content relevance, and user preference alignment—without requiring explicit supervision.

(3) Extensive experiments show that ICG consistently outperforms prior methods in aesthetics, semantic fidelity, and personalization, leading to improved user engagement.

2 Related Work

2.1 Conditional Image Generation

Conditional image generation enables personalized synthesis from inputs like text, poses, edges, semantic maps, and reference images. Text-guided models such as CLIP encode semantics into latent space. Diffusion models like Stable Diffusion (Rombach et al., 2022) set the current standard. Methods like ControlNet (Zhang et al., 2023) and MoMA (Song et al., 2024) enhance generation with structured control. For personalization, user behavior-based conditioning has been explored. DiFashion (Xu et al., 2024) uses interaction history but assumes high-quality inputs; CG4CTR (Yang et al., 2024) applies reward filtering but lacks end-to-end learning. Both focus on specific domains, whereas our method targets general-purpose cover generation and is thus not directly comparable.

2.2 Automated Assessment of Image Generation

Traditional metrics such as IS (Salimans et al., 2016), FID (Heusel et al., 2017), and CLIP Score (Radford et al., 2021) are widely used to assess image fidelity and text–image consistency, but they fail to capture subjective human preferences. To bridge this gap, several preference-aligned evaluation models have been proposed, including PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), ImageReward (Xu et al., 2023), and the Multi-dimensional Preference Score (MPS) (Zhang et al., 2024). These methods

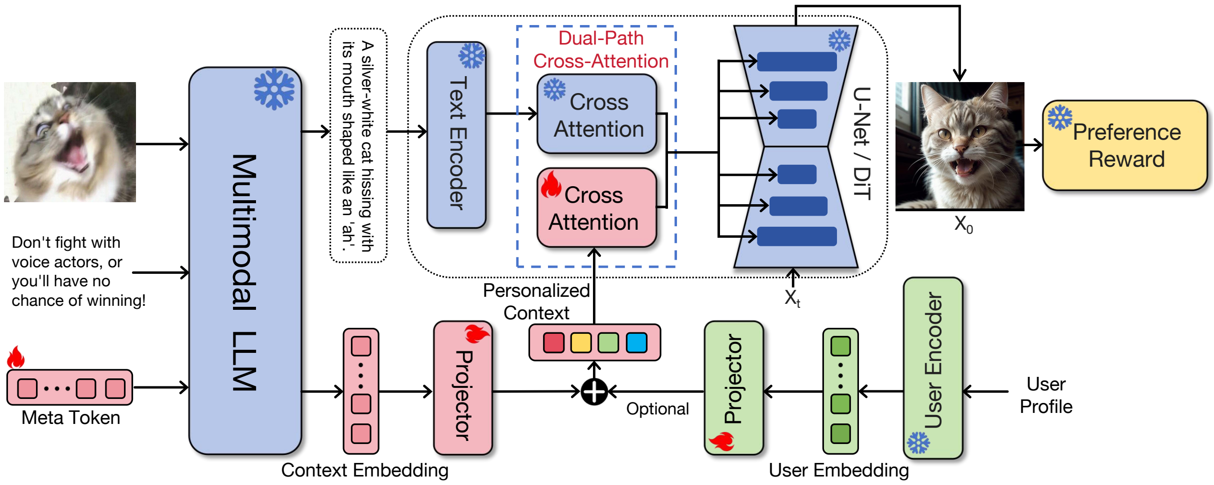

Figure 2: Overview of the proposed method. The model takes a reference image, title, and meta token to generate context embeddings via a Multimodal LLM. Combined with user embeddings, personalized features are injected into a diffusion model through a dual-path adapter. Reward models evaluate the output and guide training via feedback.

fine-tune vision–language models on large-scale human-labeled comparisons, thereby producing scores that better reflect aesthetic appeal or semantic alignment. However, they primarily model *general human preferences*, rather than the *user-specific preferences* that are crucial for personalized generation.

In our framework, we adopt a subset of these preference models—specifically HPSv2 and PickScore—as auxiliary training signals. We select them for their complementary strengths: HPSv2 captures aesthetics aligned with human judgments and PickScore provides robust comparisons for text–image relevance. Together, they form a diverse reward set that improves overall quality and personalization, while our personalized reward module further extends beyond these general signals to incorporate user-specific feedback.

2.3 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) (OpenAI, 2023) extend LLMs to visual inputs via modality-specific encoders and projection layers. Recent studies (Koh et al., 2023; Zheng et al., 2023; Pan et al., 2023) explore three paradigms for image generation: (1) symbolic prompts (Xia et al., 2023), (2) continuous visual features (Li et al., 2024), and (3) discrete tokens (Ge et al., 2024) decoded by VQ-GAN (Esser et al., 2021) or Stable Diffusion. We adopt the continuous approach for its semantic richness and compatibility with diffusion models. More recently, unified understanding–generation models, such as ILLUME (Wang et al., 2024), MetaQuery (Pan et al., 2025), and Janus-flow (Ma et al., 2025), have emerged to jointly perform vision–language reasoning and generation within a single architecture. These models demonstrate strong zero-shot capabilities and deliver impressive quality for generic content creation. However, they typically lack mechanisms for fine-grained user conditioning or preference-aware reward optimization, which are central to personalized cover generation. Our framework is therefore complementary: instead of competing with unified models on general-purpose tasks, we focus on injecting explicit user embeddings and optimizing with multi-reward supervision to achieve personalization-aware outputs, while viewing integration with unified architectures as a promising direction for future research.

3 Methodology

We propose ICG (Figure 2), a framework for generating personalized cover images for short videos and movies based on user preferences. It consists of four key components: (1) MLLM-based context prompting, which extracts features from the reference image and title; (2) personalized prompting, which encodes user profiles and integrates them with context features; (3) context adaptation, which injects the personalized prompt into the diffusion model; and (4) preference alignment learning, which leverages multiple reward models—including a custom personalized reward—for

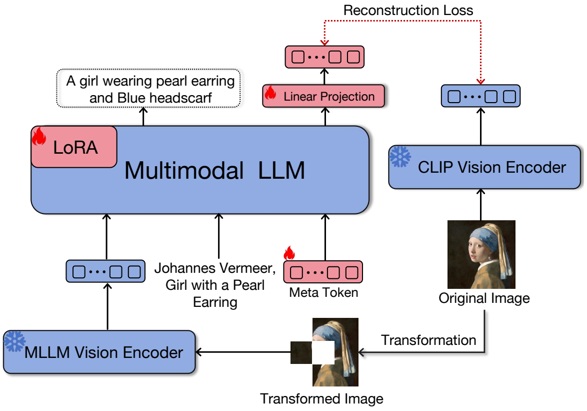

Figure 3: Model designed to train meta token.

supervision.

3.1 MLLM-based Context Prompting

We propose a Multimodal Prompt Generator based on the pre-trained MLLM Qwen2.5VL-7B (Bai et al., 2025) to produce effective prompts for cover generation. The model integrates visual cues from a reference image (I_{ref}) and textual content (T_{title}), guided by a system instruction (T_{sys}) requesting: ***"Please generate a drawing prompt that aligns with the semantics of the specified reference cover and content title."*** This yields an explicit prompt:

$$P_{exp} = MLLM(I_{ref}, T_{title}, T_{sys}).$$

The explicit prompt captures key entities from both modalities, ensuring basic semantic alignment. However, natural language, as a discrete representation, limits expressiveness. To address this, we introduce a meta token block that complements the prompt by capturing fine-grained multimodal context features in continuous space.

To enhance domain-specific understanding, we further design a Multimodal Generative Learning Stage (Figure 3). The MLLM receives I_{ref}, T_{title}, and meta tokens (Koh et al., 2023), which are jointly attended to by text and image tokens. The meta tokens are optimized to approximate the CLIP-encoded embedding of I_{ref} using a reconstruction loss:

$$\mathcal{L} = ||MLLM(V_{enc}(I_{ref}^{trans}), T_{title}, Meta\ Token) \\ -CLIP(I_{ref})||_2^2.$$

where V_{enc} is the MLLM vision encoder, and I_{ref}^{trans} is a transformed version of I_{ref} (e.g., via masking, blurring, or cropping) to enhance robustness. Although CLIP embeddings alone provide strong semantic signals, our reconstruction training enables the MLLM to jointly encode textual context and transformed visual features, enriching semantic alignment and robustness beyond CLIP's single-modal representation. While meta tokens are trained with $\mathcal{L}_{rec}$, other tokens follow a standard next-token prediction objective. Once trained, the MLLM outputs prompt-contextualized embeddings for personalized cover generation.

3.2 User-Profile-based Personalized Prompting

The context representations and explicit text generated by the MLLM are generic and lack personalization, limiting their ability to reflect diverse user preferences. To address this, we introduce a **User-Profile-Based Personalized Prompt Generator**, which encodes user attributes—such as gender, age, occupation, and preferred cover types—as personalized style preferences to guide visual output. For example, a 27-year-old male teacher favoring cartoons and children's movies would receive prompts adapted to cartoon-style aesthetics.

Formally, the multimodal context features (C_{ref}) are projected into N_c hidden embeddings via a linear layer. In parallel, user embeddings (U_{pre}), obtained from a pretrained user encoder (e.g., a two-tower CTR model (Covington et al., 2016)), are projected into N_u embeddings. The two sets are concatenated one-to-one to form the final personalized context prompt C_{ref}^{per}:

$$\begin{aligned} Proj &= LayerNorm\left(Linear(*)\right), \\ C_{ref}^{per} &= Concat\left(Proj(C_{ref}), Proj(U_{pre})\right). \end{aligned}$$

This provides a unified representation for generating covers that are both semantically aligned and user-specific.

3.3 Personalized Context Adaptation

To inject personalized features into the pretrained diffusion model, we adopt a dual-path cross-attention mechanism inspired by Stable Diffusion (Rombach et al., 2022) and DiT (Peebles and Xie, 2023), where text features are integrated into U-Net or transformer blocks via attention layers.

In each cross-attention layer, we introduce an additional branch for the personalized context. The outputs from both text and personalized paths are aggregated to capture general semantics and user-specific preferences. Given query features Z, text features c_t, and personalized features c_p, the updated output is:

$$\mathbf{Z}^{new} = \text{Attention}(\mathbf{Q}, \mathbf{K}^t, \mathbf{V}^t) + \text{Attention}(\mathbf{Q}, \mathbf{K}^p, \mathbf{V}^p).$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$ is the query matrix. $\mathbf{K}^t$, $\mathbf{V}^t$ and $\mathbf{K}^p$, $\mathbf{V}^p$ are key-value pairs derived from c_t and c_p, respectively. While $\mathbf{W}_q$, $\mathbf{W}_k^t$, and $\mathbf{W}_v^t$ are inherited from the original model, $\mathbf{W}_k^p$ and $\mathbf{W}_v^p$ are newly introduced and trained for personalization. To preserve the pretrained model, we freeze all original parameters and train only the newly added projection layers. This lightweight adaptation improves personalization while preserving the generalization ability of pretrained diffusion models. As illustrated in Figure 2, user conditions are optional: context embeddings enhance generation quality, while user embeddings enable personalization when available.

3.4 Personalized Preference Alignment Learning

As real personalized covers are unavailable as ground truth, traditional supervision (e.g., MSE) is not applicable. Inspired by reward learning from human feedback (RLHF), we guide training with multiple reward models. Public reward models (Deng et al., 2024; Wallace et al., 2024) capture general aesthetics but overlook user-specific preferences. To address this, we introduce a personalized preference reward model that provides user-aware feedback, enabling joint optimization through a strategy we term Personalized Preference Alignment Learning.

3.4.1 Training of Personalized Preference Reward Model.

Following prior work, we formulate user preferences as pairwise comparisons. Users with fewer than six interactions are filtered out. For the remaining users, interacted items are ranked by relevance signals (e.g., clicks or ratings). The top k_1 items are labeled as positive and the bottom k_2 as negative, forming up to $k_1 \times k_2$ training pairs.

The reward model is built on CLIP, enhanced with transformer layers and fully connected (FC) heads. Each input includes a title, caption (generated via CLIP-Interrogator[*]), user profile, and image. These inputs are encoded and projected as follows:

$$\begin{aligned}
t = CLIP_{txt}(title)\ ,\ c = CLIP_{txt}(caption),\\
i = CLIP_{vis}(image)\ ,\ u = CLIP_{txt}(user),\\
t_f = FC_t(t)\ ,\ c_f = FC_c(c),\\
i_f = FC_i(i)\ ,\ u_f = FC_u(u),\\
t_t, i_t, u_t = Transformer(concat(t_f, c_f), i_f, u_f),\\
p = FC_{per}(concat(t_t, i_t, u_t)).
\end{aligned}$$

where $CLIP_{txt}$ and $CLIP_{vis}$ are the CLIP text and image encoders, and p is the predicted personalized preference score. The loss is defined as:

$$\mathcal{L} = -\ \mathbb{E}_{U\sim\mathcal{D}}[\ \log\ (\sigma(\ p_m - p_n\ ))\ ]\ .$$

where p_m and p_n are scores for more- and less-preferred items, respectively. To prevent overfitting, only the last few layers of CLIP and the added modules are trained.

3.4.2 Training with Multi-Reward Feedback.

Our goal is to generate covers that are both aesthetically appealing and aligned with user preferences. To achieve this, we employ three reward models: 1) **HPSv2**: Evaluating color vividness and content completeness; 2) **PickScore**: Measuring overall visual aesthetics; 3) **Personalized Reward Model**: Capturing user-specific preferences. Training consists of two stages: 1) **Initialization**: We align personalized features with the diffusion model using a weak CLIP-based reconstruction loss between generated images and their captions; 2) **Reward Feedback Learning**: For each sample $(title_i, ref\_img_i, caption_i)$, we extract personalized features using the multimodal LLM and user encoder. A latent x_t is sampled from Gaussian noise and denoised into image x_0 via the diffusion model. The generated image is evaluated by all reward models. The final training objective is a weighted sum of reward losses:

$$\mathcal{L}_{\text{total}} = \lambda_h \mathcal{L}_h + \lambda_{per} \mathcal{L}_{per} + \lambda_p \mathcal{L}_p + \lambda_r \mathcal{L}_{\text{rec}}\ .$$

where $\mathcal{L}_h$, $\mathcal{L}_{per}$, and $\mathcal{L}_p$ denote losses from HPSv2, the personalized reward model, and PickScore, respectively. $\mathcal{L}_{rec}$ ensures alignment between image and caption. All weights λ are set to 0.25. Only the adapter and projector layers are updated, enabling efficient optimization of both personalization and visual quality.

[*] https://github.com/pharmapsychotic/clip-interrogator

Table 1: Quantitative comparisons. The best results are in **bold** and the second-best results are underlined.

Dataset	PixelRec				MovieLens			
Metric	LPIPS(↓)	SSIM(↑)	FID(↓)	Aesthetics(↑)	LPIPS(↓)	SSIM(↑)	FID(↓)	Aesthetics(↑)
Title+Image Rule-based	0.6446	0.1484	47.74	4.17	0.6512	0.1634	46.24	4.09
Text Inversion (Gal et al., 2022)	0.6282	0.1632	42.23	4.12	0.6345	0.2474	43.27	4.12
PMG (Shen et al., 2024)	0.5411	0.1624	35.18	4.21	0.4140	0.2515	33.93	4.11
ICG	**0.5126**	**0.1724**	**33.06**	**4.87**	**0.4018**	**0.2695**	**31.23**	**4.77**

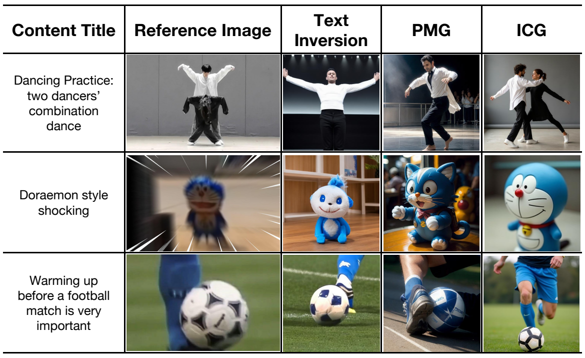
Content Title
Reference Image
Text Inversion
PMG
ICG
Dancing Practice: two dancers' combination dance
Doraemon style shocking
Warming up before a football match is very important

Figure 4: Qualitative comparison. Content titles, reference images and generated covers with different approaches

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate ICG on two public datasets representing short video and movie recommendation scenarios. (1) **PixelRec**[†] is a large-scale video cover dataset; we use its 1M subset containing 0.3M covers across 22 domains, 1M user profiles, and 10M interactions, along with metadata such as clicks, likes, titles, and descriptions. (2) **MovieLens**[‡] includes 86K movies, 0.3M users, and 3.3M ratings, with additional user demographics and movie metadata (titles, genres, and covers). We conduct both automatic and human evaluations. For image quality, we report FID (Heusel et al., 2017) and aesthetic scores using a LAION-trained predictor[§]. For personalization, we compute LPIPS (Zhang et al., 2018) and SSIM (Wang et al., 2004) between generated and reference images. These metrics jointly assess visual appeal, fidelity, and personalization. Human evaluation further validates alignment with real-world user preferences.

4.2 Baselines

We compare ICG with three generative baselines: (1) **Text Inversion** (Gal et al., 2022), which embeds user preferences into word tokens and combines them with textual prompts for diffusion-based generation; (2) **PMG** (Shen et al., 2024), which transforms user-interacted and reference images into text, then extracts preference keywords via a pre-trained LLM to guide generation; and (3) a **rule-based variant of ICG**, which replaces the personalized pipeline with a vanilla Stable Diffusion model. Given only a reference image and title, the MLLM generates a drawing prompt without personalization, highlighting the benefits of end-to-end optimization and MLLM-diffusion integration.

Table 2: The average score of generated covers in human evaluation

	PixelRec	MovieLens
ICG	**2.419**	**2.527**
Title+Image Rule-based	1.978	2.041
Text Inversion	1.952	1.923
PMG	2.152	1.994

4.3 Implementation details

We use Qwen2.5VL-7B (Bai et al., 2025) as the context prompt generator and adopt Stable Diffusion V1.5 or Flux for cover image generation, with adapters initialized from IP-Adapter-SD15 or IP-Adapter-Flux. During multimodal finetuning, only the adapter and projector layers are updated (meta token length = 1, projected dimension = 1024), ensuring compatibility across architectures. The full model is trained for 50,000 iterations using two 32GB GPUs, with a learning rate of 10^{-6} and a guidance scale of 1.0. At inference, we use the DDIM scheduler (Song et al., 2020) with 15 sampling steps and a guidance scale of 7.0. The personalized reward model is trained separately on PixelRec and MovieLens using 0.2M user-item pairs (80%-10%-10% split), optimized with Adam (lr

[†] https://github.com/westlake-repl/PixelRec

[‡] https://grouplens.org/datasets/movielens

[§] https://github.com/christophschuhmann/improved-aesthetic-predictor

Table 3: Quantitative ablation study of multimodal generative learning stage and meta tokens using the LPIPS metric two datasets. N denotes the number of multimodal tokens. The best results are in **bold** and the second-best results are underlined.

N	Finetuning	PixelRec	MovieLens
1	✗	0.4367	0.5491
2	✗	0.4359	0.5482
4	✗	0.4398	0.5526
8	✗	0.4495	0.5689
1	✓	<u>0.4194</u>	**0.5293**
2	✓	**0.4168**	<u>0.5315</u>
4	✓	0.4255	0.5391
8	✓	0.4231	0.5412

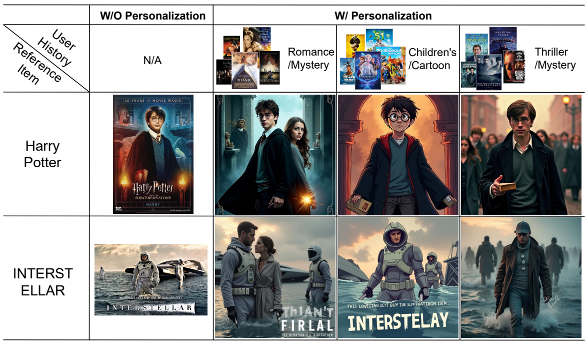

Figure 5: The effectiveness ablation of varying user conditions.

$= 10^{-4}$) and early stopping. It consists of frozen CLIP encoders, two trainable transformer layers (768 hidden size), and fully connected heads, totaling 20M trainable parameters.

4.4 Experimental Results

4.4.1 Qualitative comparison.

Figure 4 presents example outputs from ICG and three baselines, alongside content titles and reference images. ICG consistently achieves superior visual coherence and semantic alignment. In the first example, it accurately conveys the theme and color tone of "Dancing Practice: two dancers' combination dance," while baselines fail to reflect the intended meaning. The second row shows precise reconstruction of a cartoon character, whereas PMG introduces irrelevant details and Text Inversion omits key features. In the third case, it clearly depicts a warm-up scene with a football and full-body figure, effectively grounding the title, which baselines overlook. Additional results on MovieLens are discussed in later ablations.

Table 4: Overall quantitative ablation study of the ICG framework. The best results are in **bold** and the second-best results are underlined.

Dataset	PixelRec		MovieLens	
Metric	LPIPS(↓)	FID(↓)	LPIPS(↓)	FID(↓)
ICG	**0.5126**	<u>33.06</u>	**0.4018**	**31.23**
w/o Meta token	0.5912	39.24	0.5854	37.02
w/o User feature	<u>0.5203</u>	**32.67**	<u>0.4284</u>	<u>31.43</u>
w/o Both	0.5893	38.45	0.5194	36.54

4.4.2 Quantitative comparison.

As shown in Table 1, ICG consistently outperforms all baselines on both PixelRec and MovieLens. It achieves the lowest LPIPS (0.5126, 0.4018) and FID (33.06, 31.23), and the highest SSIM (0.1724, 0.2695), indicating superior personalization, realism, and structural fidelity. While PMG performs reasonably on LPIPS and FID, it lags in aesthetics and personalization. Text Inversion and the rule-based baseline perform worst, with significantly higher LPIPS and FID. ICG also attains the highest aesthetic scores (4.87, 4.77), benefiting from joint supervision by public and personalized reward models.

4.4.3 Human evaluation.

While quantitative and qualitative results confirm the effectiveness of ICG in terms of personalization and image quality, it is necessary to further examine the user-perceived quality of the generated covers. To this end, we conducted a human evaluation comparing ICG with three baselines. A total of 100 volunteers rated 120 anonymized images (30 from each method) on a 1–3 Likert scale, with higher scores indicating stronger visual quality and alignment with user preferences. All images were randomly shuffled to mitigate bias. As shown in Table 2, ICG obtains the highest average scores, indicating stronger user appeal. We emphasize that this evaluation captures *perceived appeal* rather than direct engagement; more detailed experiments on downstream recommendation performance are provided in the following sections.

4.5 Ablation and Analysis

We evaluate the impact of user feature conditions on cover generation by measuring similarity to users' historical items (personalization) and distance to the reference image (fidelity). As shown in Table 4, ICG effectively integrates user preferences, with slightly reduced reference distance

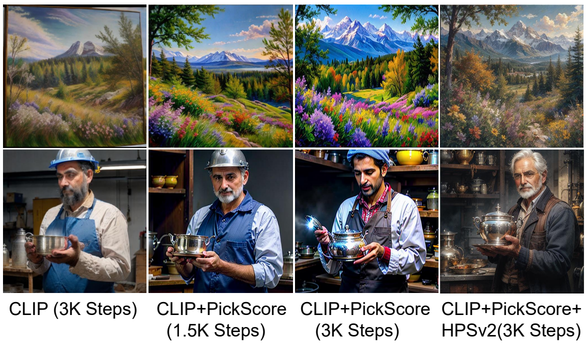
CLIP (3K Steps)
CLIP+PickScore (1.5K Steps)
CLIP+PickScore (3K Steps)
CLIP+PickScore+ HPSv2(3K Steps)

Figure 6: The effectiveness ablation of the proposed CLIP, PickScore and HPSv2 rewards.

Table 5: Quantitative ablation study of the reward models. The best results are in **bold** and the second-best results are underlined.

Metric	LPIPS(↓)	FID(↓)	Aesthetics(↑)
ICG	**0.5126**	**33.06**	**4.87**
w/o CLIP	0.5504	35.76	4.71
w/o HPSv2+PickScore	0.5413	34.87	4.45
w/o Personalized Reward	0.5653	35.81	4.54

in movie scenes—indicating personalization enhances alignment with original content.

We further visualize the impact of user conditions on generation. As shown in Figure 5, for Harry Potter, the model adapts styles such as cartoon, romance, or thriller based on user preferences; for Interstellar, it integrates elements like astronauts, aliens, and oceans. These results demonstrate that ICG tailors cover styles to individual tastes while preserving core semantics.

4.5.1 Meta tokens.

We evaluate the impact of the multimodal generative learning stage and the number of meta tokens (N) on personalization using LPIPS scores on PixelRec and MovieLens. As shown in Table 3, finetuning notably improves performance, especially on MovieLens. For non-finetuned models, larger N improves results, while finetuned models perform best at $N = 2$, with higher values degrading performance—indicating that too many tokens reduce embedding effectiveness. Table 4 further shows that removing meta tokens significantly harms both personalization and image quality, underscoring their importance in capturing multimodal context.

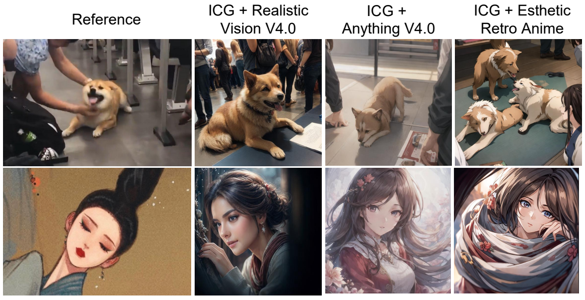

Figure 7: Generated Example Covers. Despite being trained on the base Stable Diffusion v1.5, our model can be seamlessly applied to a range of community checkpoints.

Table 6: Personalized preference prediction accuracy on test sets of PixelRec and MovieLens under different setting

	PixelRec	MovieLens
Personalized Reward Model	**85.2**	**86.2**
Only image	53.8	54.1
Image and title	61.3	67.1
Image and user profile	74.6	78.3
w/o transformers	70.5	72.5

4.5.2 Reward models.

As detailed in the Methodology, the personalized reward model is essential for enabling differentiable training in personalized cover generation. We assess its effectiveness using preference accuracy from pairwise comparisons of user-interacted items ranked by view counts (PixelRec) or ratings (MovieLens). As shown in Table 6, models relying solely on image features perform poorly, while adding titles or user profiles significantly boosts accuracy. Transformer-based fusion yields further gains, underscoring the model's ability to capture multimodal preferences. We further ablate all reward components. Figure 6 shows that using only CLIP similarity introduces visual distortions; adding HPSv2 improves realism but may introduce contrast bias, which PickScore helps mitigate by enhancing smoothness and sharpness. As shown in Table 5, removing CLIP or the personalized reward notably degrades fidelity and alignment, while omitting HPSv2 or PickScore harms aesthetics. These results underscore the complementary roles of all rewards, with the personalized module being critical for modeling user-specific preferences in ICG.

4.5.3 Analysis of compatibility.

A key advantage of our design is that the dual-path cross-attention adapter functions as a lightweight, plug-and-play module. Since the underlying diffusion backbone is kept frozen during training, the

Table 7: Comparison of MMGCN's recommendation performance using different item and user image features. **Best results** are highlighted in bold, and second-best results are underlined.

	Item	User	Recall@10	NDCG@10
w/o image	✗	✗	16.17%	0.0749
Item	✓	✗	17.94%	0.0853
Averaged-user	✓	Average	18.99%	0.0991
Generated-user	✓	Generated	**20.21%**	**0.1016**

adapter learns to modulate attention features without altering the pretrained generative capacity of the model. This enables ICG to generalize beyond the base Stable Diffusion v1.5 checkpoint and seamlessly extend to custom fine-tuned variants derived from the same foundation. In practice, this means that once trained, our adapter can be directly inserted into diverse diffusion checkpoints without additional retraining or parameter adjustment.

As illustrated in Figure 7, ICG works out of the box on widely used community models from HuggingFace and CivitAi (Civitai, 2024), including Realistic Vision V4.0 (Adhik Joshi, 2024), Anything v4 (Xyn AI, 2024), and Esthetic Retro Anime (OneRing, 2024). These results highlight the universality of our module design and its practicality for real-world deployment, where model diversity and user-specific customization are essential.

4.6 Applications in Recommendation Tasks

To complement the human evaluation that primarily measured perceived appeal, we further evaluated the downstream utility of ICG-generated covers in recommendation tasks. Experiments were conducted on the MovieLens dataset using the multimodal recommendation model MMGCN (Wei et al., 2019). We compared four input configurations: (1) *w/o image*, using only item IDs; (2) *Item*, using original item images; (3) *Averaged-user*, incorporating user features from averaged images of previously interacted items; and (4) *Generated-user*, incorporating personalized covers generated by ICG conditioned on user profiles. Evaluation followed standard offline recommendation metrics, including Recall@10 and NDCG@10, which are widely adopted as proxies for user engagement in large-scale recommender systems. Recall@10 measures the proportion of relevant items successfully retrieved within the top-10 recommendations, reflecting coverage of user interests. NDCG@10 (Normalized Discounted Cumulative Gain) further accounts for the ranking positions of relevant items, assigning higher importance to those appearing earlier in the list.

As shown in Table 7, adding visual features improves accuracy across all settings. Notably, the *Generated-user* configuration achieves the strongest results, with improvements of **+2.27% Recall@10** and **+19.1% NDCG@10** over the best baseline. These results demonstrate that the personalized covers produced by ICG not only enhance visual appeal but also translate into measurable gains in recommendation effectiveness, suggesting their potential to improve downstream user engagement. Future work will extend this offline evaluation with online A/B testing to directly assess the impact on click-through rate and other behavioral metrics.

5 Conclusion

We propose ICG, a unified framework for personalized cover generation that integrates multimodal large language models (MLLMs) with diffusion models. By leveraging context and user-profile prompts, it generates outputs aligned with both item semantics and user preferences. A multi-reward learning strategy enables end-to-end optimization without the need for ground-truth labels. Experiments on two datasets demonstrate consistent improvements in image quality, semantic relevance, and personalization. As a plug-and-play module, ICG can be seamlessly integrated into existing diffusion pipelines.

Limitations

While ICG delivers strong improvements, several limitations remain. First, our current framework relies on static user profile embeddings, which limits its sensitivity to short-term preference shifts; integrating session-based or online user modeling is an important next step. Second, although multi-reward supervision enhances personalization, it introduces about 20% additional training cost, and inference latency is roughly 1.5 seconds per image on a V100 GPU. Future work will explore model compression and accelerated diffusion backbones to enable real-time deployment. Finally, our evaluation of user-facing benefits is limited to offline proxies (Recall@10, NDCG@10) and human preference ratings. A direct measurement of engagement through large-scale online A/B testing remains an important avenue for future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No.U24A20326. We also acknowledge partial support from MindSpore (https://www.mindspore.cn), a new deep learning computing framework.

References

2022. Midjourney. https://www.midjourney.com/ Accessed: August 3, 2023.

Adhik Joshi. 2024. realistic vision v40. Accessed: 2024-03-06.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–11.

Civitai. 2024. Civitai. Accessed: 2024-03-06.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, pages 191–198.

Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. 2024. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7423–7433.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024. Making llama SEE and draw with SEED tokenizer. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Nailei Hei, Qianyu Guo, Zihao Wang, Yan Wang, Haofen Wang, and Wenqiang Zhang. A user-friendly framework for generating model-preferred prompts in text-to-image synthesis. In *Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2139–2147.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 6626–6637.

Ivan Jarsky, Valeria Efimova, Ilya Bizyaev, and Andrey Filchenkov. 2024. Conditional vector graphics generation for music cover images. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pages 233–243.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. 2023. Generating images with multimodal language models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Dongxu Li, Junnan Li, and Steven Hoi. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751.

Omneky. 2024. How ai image generation is changing the face of advertising.

OneRing. 2024. era-esthetic retro anime. Accessed: 2024-03-06.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2023. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. 2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2226–2234.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. PMG : Personalized multimodal generation with large language models. In *Proceedings of the ACM on Web Conference 2024 (WWW)*, pages 3833–3843.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. 2024. Moma: Multimodal LLM adapter for fast personalized image generation. *CoRR*, abs/2404.05674.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. 2024. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *CoRR*, abs/2306.09341.

Bin Xia, Shiyin Wang, Yingfan Tao, Yitong Wang, and Jiaya Jia. 2023. Llmga: Multimodal large language model based generation assistant. *arXiv preprint arXiv:2311.16500*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. 2024. Difashion: Towards personalized outfit generation and recommendation. *CoRR*, abs/2402.17279.

Xyn AI. 2024. Anything v4.0. Accessed: 2024-03-06.

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

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. 2024. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *CoRR*, abs/2310.02239.