Personalized Visual Content Generation in Conversational Systems

Xianquan Wang[1*], **Zhaocheng Du**[2*], **Huibo Xu**[1*], **Shukang Yin**[1], **Yupeng Han**[1], **Jieming Zhu**[2], **Kai Zhang**[1†], **Qi Liu**[1]

[1]University of Science and Technology of China
[2]Huawei Noah's Ark Lab

{xqwang, xhbxhb, xjtupanda, yupenghan}@mail.ustc.edu.cn
zhaochengdu@huawei.com jiemingzhu@ieee.org
{kkzhang08, qiliuql}@ustc.edu.cn

Abstract

With the rapid progress of large language models (LLMs) and diffusion models, there has been growing interest in personalized content generation. However, current conversational systems often present the same recommended content to all users, falling into the dilemma of "one-size-fits-all." To break this limitation and boost user engagement, in this paper, we introduce PCG (**P**ersonalized Visual **C**ontent **G**eneration), a unified framework for personalizing item images within conversational systems. We tackle two key bottlenecks: the depth of personalization and the fidelity of generated images. Specifically, an LLM-powered *Inclinations Analyzer* is adopted to capture user likes and dislikes from context to construct personalized prompts. Moreover, we design a dual-stage LoRA mechanism—*Global LoRA* for understanding task-specific visual style, and *Local LoRA* for capturing preferred visual elements from conversation history. During training, we introduce the visual content condition method to ensure LoRA learns both historical visual context and maintains fidelity to the original item images. Extensive experiments on benchmark conversational datasets—including objective metrics and GPT-based evaluations—demonstrate that our framework outperforms strong baselines, which highlight its potential to redefine personalization in visual content generation for conversational scenarios like e-commerce and real-world recommendation.

1 Introduction

The rise of large language models (LLMs) [33, 34] has sparked conversational systems across industries. For instance, e-commerce platforms can recommend preferred items through multi-turn conversations [29, 5], while music apps interactively suggest album cover [3]. They operate by dynamically analyzing users' inclinations (likes or dislikes) during interactions [13]: when users show strong interest in specific item categories, or strongly advocate for a specific item, the dialog prioritizes related content; conversely, explicit rejections reduce recommendations about those items [11, 20]. Usually, the systems also include the **item image** to attract users, as shown in Figure 1.

A picture is worth a thousand words; images are the most intuitive and vivid representation of items. Unfortunately, current conversational systems **presented identical item representations** to all users, even though different users prefer different points of the same item. For example, the

[*]Equal contribution
[†]Correspondence to: Kai Zhang (kkzhang08@ustc.edu.cn)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

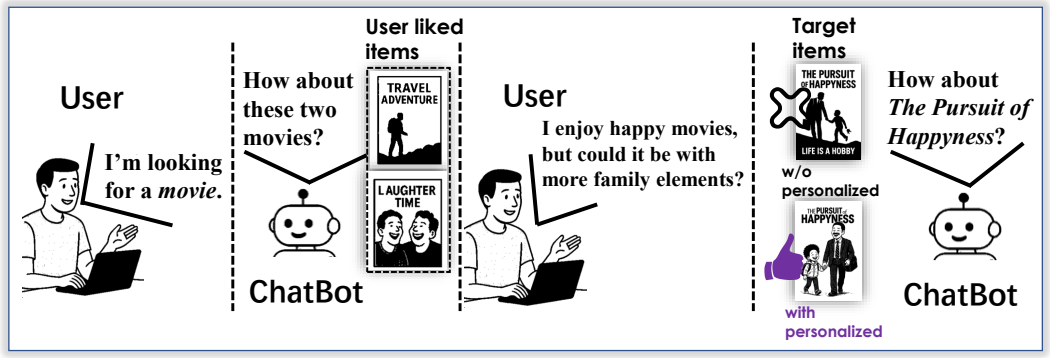

Figure 1: A toy example of content generation personalization in **movie conversational systems**. In this example, the personalization focus on visual content.

parenting, adventure, and inspirational themes of movie *The Pursuit of Happyness* attract different users. Incorporating personalized image content into the system is expected to significantly enhance the system's level of personalization, which is expected to lead to a more satisfying user experience, increase user engagement, and improve alignment between user intent and system output—particularly in application domains such as entertainment [30], e-commerce [17], and digital assistants [36].

However, little to no work has systematically explored personalized visual content generation within conversational systems, and achieving this goal is far from easy. In the conversation process, generating the visual content that reflects the user's preferences on the target item faces three main challenges. First, **how to extract the user's inclinations that can be understood by content generator** based on the conversation content. These inclinations consist of two parts: one is the positive preference, or what is known as "likes"; the other is the aversion to certain styles, or "dislikes." These inclinations may reflect on certain style types or may be deeply hidden in historical images. This requires a thorough understanding and analysis of the conversation history to accurately and effectively summarize the user's inclinations. Another challenge is **how to incorporate the user's preferred visual elements** into the image generation process, as even if a user enjoys "comedy" films, the preferred elements may vary greatly. The image references in conversational systems can provide beneficial guidance. Finally, **how to preserve the identity** in the generation process, ensuring that while personalization is reflected, the identity characteristics (such as the actor or semantic information) remain as consistent as possible during the inclinations' conditioning process.

To address the three challenges mentioned above, we constitutes one of the earliest comprehensive attempts to form a unified framework **PCG** (**P**ersonalized Visual **C**ontent **G**eneration) in conversational systems. **First**, we utilize a language model to summarize and analyze past interaction items, extracting both positive and negative user preferences. In this process, the language model acts as a analyzer, with word weight modeling to ensure that the output remains within expectations. These summaries form user preference descriptions, which guide the generation process to include or exclude content. **Second**, integrating previously favored content elements into the generation is also challenging. Based on the contextual understanding ability of base models, we propose a strategy that directly incorporates historical item images by concatenating each other on a unified canvas. This allows the generation process to be conditioned on visual preferences without modifying the model architecture. **Finally**, ensuring identity consistency requires the model to understand the relation between the target item and historical images. To facilitate this, we propose the sequence invariance strategy for historical items condition, and global-local LoRA mechanism for guidance, which help the model better capture this relationship.

For evaluation, we adopt two commonly used benchmark datasets in conversational systems to validate the effectiveness of our method. In the objective experiment, we use various low-level and high-level metrics to assess the degree of personalization and fidelity of the generated outputs. In the GPT evaluation study, we further evaluate image aspects such as style and coherence. Experimental results demonstrate the effectiveness of our approach, and this work may open up a new research direction at the intersection of conversational AI and visual generation. The code is publicly available at `https://github.com/xqwustc/PCG`.

2 Related Work

2.1 Conversational Systems

Conversational systems [19] have been widely explored in fields such as robotics, smart shopping assistants, and recommender systems [25, 32, 27]. With the advancement of language models [37, 35] for text [16] and language processing, these systems have evolved from handling simple dialogues to understanding complex semantics [33]. The rise of large language models has enabled deeper comprehension, allowing systems to better understand user needs and empathize with user emotions [12]. This enhances the sense of engagement and satisfaction during interactions [38]. However, most current systems focus only on interaction at each turn, with little attention paid to personalizing rich visual content based on the conversations. Our work addresses this gap by introducing a novel approach to content personalization within conversational systems.

2.2 Content Generation

With the rapid development of large language models and visual generation models, personalized content generation has drawn increasing attention [21, 31]. For instance, personalization in large models enables the generation of responses tailored to users' specific needs and preferences. The primary goal of such personalization is to enhance user satisfaction by producing outputs that better align with individual requirements and expectations. A key function of this type of personalization is to generate text content that matches the user's writing style [14, 1]. These approaches have been applied in various contexts, such as user-level personalization based on interaction history, role-level customization for specific personas, and global preference alignment (*e.g.*, ethics, factuality). However, these efforts have been predominantly limited to text-based personalization.

Personalization of visual content has also been widely explored. For example, several studies have used Diffusion-based models [6, 23, 4] or the FLUX architecture [26] to customize image generation. Their success is built on the foundation of **Diffusion Transformers (DiTs)**. These methods support prompt-driven generation and style transfer, among other features. Recently, large multimodal models [2, 10] have integrated both understanding [15] and generation capabilities. However, they still lack the ability to actively infer user preferences and rely heavily on explicit user input. As highlighted in pioneering works [24, 28, 34], some studies have attempted to explore personalization in the multimodal domain (*e.g.*, images). Nevertheless, these works fail to model user interests from **conversational history** [7], leaving a blank in personalizing visual content within multi-turn conversations.

3 Preliminaries

Let $\mathcal{P}$ be the set of participants and $\mathcal{I}$ the set of items. A *conversation* is represented as a sequence of interaction turns: $C = (p_t, s_t, i_t)_{t=1}^{T}$, where at each turn t, a participant $p_t \in \mathcal{P}$ (either a *user* or a *chatbot*) produces an utterance s_t, which may mention a subset of items $i_t \subseteq i$. The set i_t can be empty if no items are referenced. The conversation captures the alternating turns between the user and the agent, along with the evolving context of item mentions.

The ultimate goal of this work is to enable *personalized content generation* for the **target item** recommended by a chatbot in a multi-turn dialogue. Specifically, after interacting with a user across multiple conversational turns, the chatbot recommends a target item i^*. Based on the entire conversation history, the system aims to generate personalized content associated with i^* that aligns with the user's preferences expressed during the conversation.

While the definition of "content" can be broad, in this paper, we focus on generating *visual content*, such as posters, thumbnails, or cover images. The generated visual should not only represent the semantics of the item i^*, but also incorporate stylistic or thematic elements that resonate with the user's individual taste. Formally, the task can be defined as: $v_{i^*}^* = \arg\max_{v \in \mathcal{V}_{i^*}} \text{Score}(v \mid C)$.

Ideally, the target $v_{i^*}^*$ should receive the highest scores across all metrics, such as similarity to the conversation history or the user's preferences. For visual content, the score can refer to the **semantic similarity** between the generated image [28] and the original image, or between the image and items the user has previously liked [34].

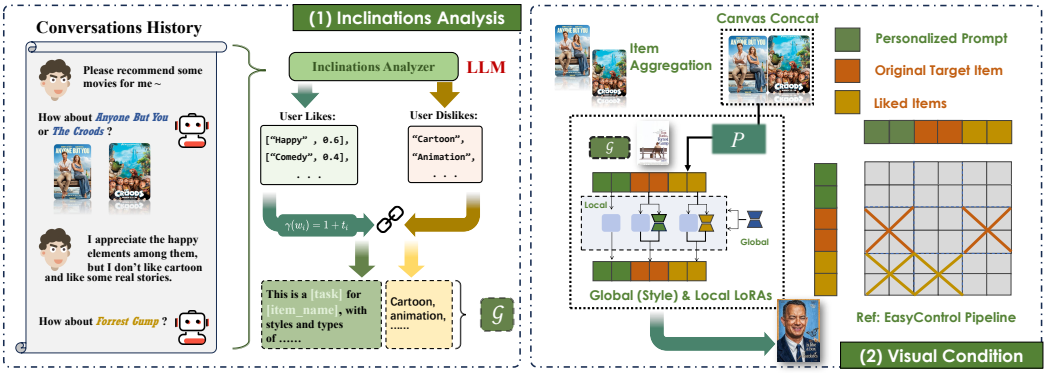

Figure 2: The PCG pipeline consists of two main components. On the left, the inclination analyzer summarizes the user's likes and dislikes. On the right, both global and local LoRAs control the style condition, the original target item condition, and the liked items condition. The personalized prompt has access to the other two tokens, while each of the two tokens is only aware of its own state.

4 Methods

In this section, we introduce the role of each component in generating the final output. First, the analyzer summarizes the user's conversation inclinations (*e.g.*, likes and dislikes), providing a deep understanding of past inclinations. Then, to better integrate previously preferred item content, we leverage DiTs' contextual understanding capabilities. Moreover, by using the sequence invariance of historical interactions under shuffled order, the model improves its ability to retrieve relevant interest-related content.

4.1 Inclinations Analyzer

During conversations, the user's inclinations primarily manifest in two ways: **(1) explicit statements**, such as "I want a xxx type of item" or "I don't like items with xxx style," and **(2) reactions** to the item candidates provided by the system, for instance, liking *Titanic* or *Flipped*, which indirectly suggests a preference for romantic items or items about love. These two types of preferences complement each other, and analyzing them in a unified manner is essential for accurately understanding the user's inclinations (both likes and dislikes).

Let the user's positive inclination (preference) word set be:

$$\mathcal{W} = \{w_1, w_2, \ldots, w_n\}. \tag{1}$$

Each word w_i is associated with a non-negative preference score $t_i = s(w_i) \in \mathbb{R}_{\geq 0}$, which reflects the user's level of interest in that item. To normalize the scores into a valid probability distribution (as shown in Figure 2), we define the user's tendency distribution $\mathcal{T}$ as:

$$\mathcal{T} = \{(w_i, t_i) \mid w_i \in \mathcal{W}\}. \tag{2}$$

For example, in the movie poster scenario, the user's original preference scores could be:

$$\mathcal{T} = \{(\text{“Sci-fi”}, 0.375),\ (\text{“Warm”}, 0.25),\ (\text{“Mystery”}, 0.25),\ (\text{“Comedy”}, 0.125)\}.$$

The inclination analyzer converts users' explicit statements and implicit feedback into quantifiable preference weights through a *decay-aware weighting mechanism*. Given a normalized tendency distribution $\mathcal{T} = \{(w_i, t_i)\}$, we compute the final recommendation weights as:

$$\gamma(w_i) = 1 + t_i. \tag{3}$$

This design provides two distinctive advantages over conventional normalization approaches: **Non-Zero Preservation**: By establishing a baseline weight of 1.0 for all attributes, our method prevents complete elimination of low-probability preferences. Formally, we have $\gamma(w_i) \geq 1.0, \quad \forall w_i \in \mathcal{W}$.

Therefore, even marginally preferred attributes (*e.g.*, $t_i < 0.1$) retain influence in the generation process. And for **Relative Emphasis Modulation**: The additive weighting compresses extreme differences while maintaining ordinal relationships. For any two attributes (w_j, w_k):

$$\frac{\gamma(w_j)}{\gamma(w_k)} = \frac{1+t_j}{1+t_k} < \frac{t_j}{t_k}, \quad \text{when } t_j > t_k. \tag{4}$$

This property mitigates over-specialization in recommendations, especially in the process of generation process. After obtaining each w_i and its corresponding $\gamma(w_i)$, we aim to ensure the generation model effectively captures this relationship. Studies show that DiT-based models can leverage (word, weight) pairs for prompt weight adjustment [18]. Hence, we design a **positive prompt** template, denoted as $\mathcal{G}_{\text{pos}}$, which can be formulated as:

$$\mathcal{G}_{\text{pos}} = \text{This is a [task] for [item\_name], with styles and types of } (w_1 : \gamma(w_1)), \ldots,$$

where $w_i \in \mathcal{W}$. In practice, we choose the top-K highest weight words to construct the personalized prompt. For the **negative prompt** $\mathcal{G}_{\text{neg}}$, it mainly includes the types disliked by the user. This is used to reduce the appearance of elements that may cause discomfort. Since these cases are relatively rare, we directly use the disliked types to form the prompt without weighting the negative words.

This prompt acts as a textual supervision signal, guiding the generation process during the distribution mapping phase. $\mathcal{G} = \{\mathcal{G}_{\text{pos}}, \mathcal{G}_{\text{neg}}\}$ forms the overall **personalized prompt**.

However, relying solely on textual supervisory signals introduces several issues, such as the model's inability to capture the inherent semantic information of the original target item and integrate elements from previously liked item content. We additionally incorporate the content and textual information of both the target item and historically liked items for generation condition.

4.2 Visual Content Condition

To enable the model to condition on both the original item and historical liked items, it is essential to incorporate historical information into the generation process. Specifically, the personalized prompt should guide the generation jointly with the visual content, while maintaining as much independence as possible between the original item and historical items. This is because the generation process introduces personalized elements based on the original item. A straightforward approach is to use Group Diffusion Transformers [8]. However, such methods have two major drawbacks: they are heavily influenced by the VAE and fail to make historical visual elements visible to each other during the condition generation process.

Inspired by [9], we propose a method to allow textual and visual modeling to jointly guide the generation process without modifying any model architecture. Specifically, historical visual content is concatenated before encoding and then fed into the encoder as a single image. This approach is motivated by [9], which demonstrated that DiT can effectively handle concatenated images and merged prompts, capturing contextual relationships within them.

4.2.1 Multi-Item Aggregation

Given N item images $\{I_i\}_{i=1}^{N}$ a user liked in the conversation, we spatially concatenate them into a unified canvas I_{cat}:

$$I_{\text{cat}} = \mathcal{C}(I_1, I_2, ..., I_N) \in \mathbb{R}^{H \times (N \cdot W) \times C}, \tag{5}$$

where $\mathcal{C}(\cdot)$ denotes row-wise concatenation operator. Similarly, we could also define a column-wise operator. In general, for vertically oriented visual items, we can stack them row by row for concatenation, while for horizontally oriented items, we concatenate them along the column dimension. This approach ensures that the aspect ratio of the resulting image does not vary significantly, enabling the text to better guide the image generation process. It is important to note that during the generation process, the fused information should not depend on the order of concatenation. Therefore, the training process should be designed to minimize sensitivity to the order of concatenation.

4.2.2 Sequence Invariance

To make the model more robust to the order of input features while keeping their meaning unchanged, we propose a mechanism called **Sequence Invariance**. This method ensures that the model's output does not depend on the order in which input items are concatenated. We implement this by randomly changing the order of concatenated input images during training.

We define the concatenated image I_{cat} as a block matrix as in Equation 5 with I_1 to I_N. Each block $I_i \in \mathbb{R}^{H\times W\times C}$ is an individual input image, and $I_{\text{cat}} \in \mathbb{R}^{H\times(N\cdot W)\times C}$ represents all N images placed side by side. To apply Sequence Invariance, we use a permutation matrix $P \in \{0,1\}^{N\times N}$ to reorder the blocks in I_{cat}:

$$I_{\text{cat, shuffled}} = P \cdot I_{\text{cat}}, \tag{6}$$

where P satisfies:

$$P_{ij} = \begin{cases} 1 & \text{if block } j \text{ moves to position } i, \\ 0 & \text{otherwise,} \end{cases} \quad P^\top P = PP^\top = I_N. \tag{7}$$

Here, P is an orthogonal matrix, and I_N is the $N \times N$ identity matrix. This operation permutes the positions of $I_1, I_2, \ldots, I_N$ without changing their content.

During training, a new permutation matrix P is randomly sampled for each batch. P is chosen uniformly from the set: $\mathcal{P}_N = \{P \in \{0,1\}^{N\times N} \mid P^\top P = I_N\}$. This ensures all possible input orders are equally used in training and helps the model learn order-independent features.

We construct P from a random permutation $\pi \in S_N$, the set of all orderings of N elements:

$$P_{ij} = \begin{cases} 1 & \text{if } \pi(j) = i, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

At inference time, we use the identity matrix $P = I_N$ to keep the original input order:

$$I_{\text{cat, shuffled}} = P \cdot I_{\text{cat}} = I_N \cdot I_{\text{cat}} = I_{\text{cat}}. \tag{9}$$

In Appendix C.2, we provide an in-depth analysis of the Sequence Invariance.

4.2.3 Global and Local LoRA

Based on concatenated images, how these images guide the generation process is a key challenge. In typical text-to-image generation tasks, textual prompts alone often serve as the condition. While this enables diverse outputs, it may also result in large deviations from the original visual content. A common strategy in the community is to first train a task-specific LoRA, which, through fine-tuning, embeds useful prior knowledge into the model for better alignment with the task.

In our case, we aim to condition the generation not only on the original image but also on historical visual information. This motivates the design of two types of LoRA modules: a **global LoRA** to control the overall style and semantics, and a **local LoRA** that selectively influences the content branch to better fuse visual conditions. The global LoRA is obtained based on the target task, though integrating visual features into the local LoRA is non-trivial.

Thanks to the powerful plugin system of `EasyControl`, we can fully leverage this dual-conditioning strategy during generation. As shown in Figure 2, built on FLUX, `EasyControl` provides additional LoRA signals for branches such as subject and background, enabling the model to incorporate user-specific visual history. Meanwhile, the global LoRA is applied across noise, personalized prompts, and visual condition branches to ensure consistency with the task. By jointly using a task-specific global LoRA and a visually-oriented local LoRA, our framework effectively balances personalized preferences with the faithful rendering of target item content during generation. Specifically, we adopt the same architecture as `EasyControl`, where the tokens from the personalized prompt attend to both the subject and background through local LoRA. In contrast, the subject and background tokens only interact with their own content. This design is important for our task because it ensures that control is consistently guided by the prompt. It also keeps the blending of subject and background under textual supervision, preventing mutual interference.

Global LoRA In this stage, the base model parameters θ are frozen, and only the style LoRA parameters ϕ_{style} are trained. No visual or dialogue conditioning is used. Our model is based on FLUX.1-dev [2], which uses a flow-matching loss. Therefore, the same loss function is adopted as:

$$\mathcal{L}_{\text{style}} = \mathbb{E}_{t,\, x_0 \sim \mathcal{N}(0,I),\, x_1} \left\| u_{\theta+\phi_{\text{style}}}(x_t, t) - (x_1 - x_0) \right\|_2^2, \tag{10}$$

where x_0 denotes Gaussian noise, x_1 denotes the original data sample, and x_t is an intermediate state obtained by interpolating between them, typically defined as $x_t = (1-t)x_0 + tx_1$. The style LoRA parameters ϕ_{style} are optimized using the gradient $\nabla_{\phi_{\text{style}}} \mathcal{L}_{\text{style}}$.

Training Subject/Background LoRA (Local) with Frozen Global LoRA After the style LoRA ϕ_{style} is obtained and fixed, we introduce two local LoRA modules, ϕ_{subj} and ϕ_{bg}, to enable personalized visual conditioning. The context $F_\Theta(C)$ encodes cues about the subject and background. The goal is to guide the generation from the original distribution to the target distribution, conditioned on the dialogue history C and the referenced images, using the following loss:

$$\mathcal{L}_{\text{local}} = \mathbb{E}_{t,\, x_0 \sim \mathcal{N}(0,I),\, x_1,\, F_\Theta(C)} \left\| u_{\theta+\phi_{\text{style}}+\phi_{\text{subj}}+\phi_{\text{bg}}}(x_t, t, F_\Theta(C)) - (x_1 - x_0) \right\|_2^2, \tag{11}$$

where gradients are applied to ϕ_{subj} and ϕ_{bg} (*i.e.*, $\nabla_{\phi_{\text{subj}}} \mathcal{L}_{\text{local}} \neq 0$, $\nabla_{\phi_{\text{bg}}} \mathcal{L}_{\text{local}} \neq 0$), while ϕ_{style} remains frozen.

Inference During the inference process, it is necessary to balance the influence strength of the subject and background elements on the generated content. Let the weights of these two local LoRAs be represented by λ_1 and λ_2, respectively. Generally, λ_1 is set to be slightly larger than λ_2 since the background serves as a supplementary role. We preset $\lambda_1 = 1$ and $\lambda_2 = 0.85$.

5 Experiment

In this section, we introduce the base model, comparison models, and details of the datasets used in our experiments. Although many conversational systems exist in the field, high-quality conversation data with available visual content **remains scarce**. Following previous works of conversational recommender systems [22, 39], we conduct experiments on two conversational recommendation datasets set in movie scenarios. The hyperparameters we adopted are shown in Appendix G.

5.1 Dataset

Dataset	Dialogs	Utterances	Avg. Turns
E-ReDial	756	12,003	15.9
Inspired	1,001	35,811	10.73

Table 1: Data overview of conversational datasets in the movie scenario.

The two datasets are classic benchmarks for movie conversational recommender systems, containing many high-quality interactions with the systems.

For both datasets, the original data was randomly split into training, validation, and test sets with a ratio of 8:1:1. The generated results were evaluated on the test set.

5.2 Evaluation Metrics

Moreover, we use CLIP to measure similarity between images and texts (CLIP-T), as well as between images (CLIP-I), in order to evaluate the semantic and visual similarity between the generated and original images. The Deep Inception Score (DIS) Score is also defined to assess the cosine similarity of the last layer of two images after passing through the Inception model.

In our evaluation, different image similarity metrics are computed using various image resolutions. Specifically, Inception-based metrics such as FID and DIS are processed at the resolution, while CLIP-based metrics (CLIP score and Image CLIP score) and LPIPS are evaluated at 224×224. For MS-SSIM, we also maintain the resolution for image pairs to capture fine-grained structural similarities. Moreover, for LPIPS calculation, we use AlexNet as the base model.

[2] `https://huggingface.co/black-forest-labs/FLUX.1-dev`

Table 2: The objective evaluation results of E-ReDial.

Types	Models	Historical Liked		Target					FID↓
		MS-SSIM ↑	CLIP-T↑	DIS↑	CLIP-T↑	CLIP-I↑	MS-SSIM↑	LPIPS↓	
Text-to-Image	SD3	0.1285	26.52	84.97	28.82	74.14	0.1286	0.6831	21.32
	FLUX.1-dev	0.1741	25.02	82.56	27.49	68.14	0.1875	0.6692	26.51
Image-to-Image	EasyControl-SUBJ	0.1437	25.97	87.94	30.72	77.92	0.2166	0.5561	20.20
	EasyControl-BG	0.1514	24.62	80.20	26.96	64.72	0.1765	0.6742	25.01
	PCG (Ours)	0.1589	26.52	85.76	30.52	73.94	0.1938	0.6238	21.88

Table 3: The objective evaluation results of Inspired.

Types	Models	Historical Liked		Target					FID↓
		MS-SSIM ↑	CLIP-T↑	DIS↑	CLIP-T↑	CLIP-I↑	MS-SSIM↑	LPIPS↓	
Text-to-Image	SD3	0.1145	26.12	84.24	27.81	69.31	0.1236	0.6810	20.72
	FLUX.1-dev	0.1819	25.36	83.25	27.17	68.32	0.1767	0.6673	25.66
Image-to-Image	EasyControl-SUBJ	0.1521	25.87	88.52	30.70	79.08	0.2201	0.5371	19.74
	EasyControl-BG	0.1906	24.89	81.09	26.38	61.62	0.1746	0.6947	25.41
	PCG (Ours)	0.1764	26.23	85.55	29.56	72.20	0.1872	0.6291	22.41

5.3 Main Experiment (Objective Evaluation)

The main experiment primarily involves the analysis of objective metrics related to the content. We focus on two key aspects: the integration of historical liked items into the generated content and the fidelity to the target item intended for interaction. Specifically, for EasyControl, we evaluate the generative performance in two separate settings: **-SUBJ**, where only information from the target item is used for conditioning, and **-BG**, where only background information is used as the conditioning input. Both settings are based on the EasyControl framework. The experiment on E-ReDial and Inspired can be found in Table 2 and 3.

As shown in Table 2, the conventional text-to-image methods struggle to maintain high fidelity to the target item. Specifically, the similarity between the generated movie posters and the captions of the target item's posters, measured by CLIP-T, is around 27–29, generally lower than that achieved by image-to-image models. This highlights the importance of effectively conditioning on the target item to ensure model fidelity. On the other hand, text-to-image models perform relatively well in integrating historical interests. For example, the FLUX.1-dev model achieves high SSIM similarity scores, and SD3 outperforms in the CLIP-T metric when compared to historical liked items. For EasyControl-BG, it can be observed that solely using concatenated historical items does not improve historical relevance compared to FLUX.1-dev. This indicates that a single LoRA for background signals cannot generate a coherent guiding signal with the prompt, meaning it fails to maintain fidelity or integrate historical interests.

On E-ReDial, when using the target item directly as the image condition along with user historical items as text prompts, the historical information can be partially integrated. This approach achieves an MS-SSIM of 0.1437 against historical images, showing noticeable improvement compared to SD3. However, the integration is still not thorough enough, as evidenced by the two historical interest metrics where EasyControl-SUBJ underperforms compared to our proposed PCG framework. This further validates the effectiveness of our method. On Inspired in Table 3, PCG excels at preserving historical user preferences. While the baseline FLUX.1-dev can partially reflect user history, its semantic fidelity (measured by CLIP-T) is significantly lower than PCG's. EasyControl-SUBJ shows strong performance in capturing details of the target item, but this comes at the cost of weaker integration of information from previously liked items. These results highlight PCG's ability to strike a strong balance between fidelity and personalization.

Overall, text-to-image methods exhibit promising results in capturing historical interests, likely because explicitly incorporating user history into prompts enhances generative capability. In contrast, image-to-image methods demonstrate superior fidelity, possibly due to stronger perceptual alignment with image signals. Achieving optimal performance in both aspects simultaneously is challenging, as there is an inherent trade-off between them. Nevertheless, PCG effectively integrates historical information while maintaining high fidelity, proving its overall superiority.

Table 4: GPT evaluation score (subset).

Model	V	I	T	D
FLUX.1-dev	3.00	2.00	2.25	2.75
SD-3	4.50	4.50	4.75	4.50
EC-SUBJ	2.00	1.25	1.50	2.25
EC-BG	3.25	3.25	3.00	3.75
PCG	5.00	5.00	4.75	5.00

Table 5: GPT evaluation score (full test set).

Model	V	I	T	D
FLUX.1-dev	3.94	3.23	3.26	3.97
SD-3	3.48	3.13	3.11	3.40
EC-SUBJ	4.14	3.82	3.86	4.03
EC-BG	3.81	3.61	3.31	3.95
PCG	4.34	4.01	4.17	4.38

5.4 GPT Evaluation

Due to the subjective nature of personalized generation, we use GPT-4o to evaluate the generated results, aiming to make the assessment more objective and targeted. When GPT-based models are unavailable, the Gemini series models serve as good alternatives. In this part of the experiment, we first randomly select a subset of generated dialogues from the test set and compare them with other generated results. The relevant prompts are provided in Appendix I. According to the scores from GPT on E-ReDial subset (shown in Table 4), our proposed PCG performs well across all evaluation aspects, indicating a good overall balance between fidelity and personalization. This gap is substantial, demonstrating the exceptional performance of PCG in specific contexts and scenarios.

On the full E-ReDial test set, PCG shows a smaller score gap compared to other base models. However, PCG still achieves the best performance across all four dimensions. Apart from PCG, EC-SUBJ closely follows in each dimension, showing strong consistency with the objective metrics evaluation. For example, its performance in objective metrics is outstanding, and it also scores highly in subjective evaluation, particularly in coherence and details. EC-BG performs well in the Integration dimension, indicating that background as a condition is effective. However, due to a lack of good synergy with other techniques, it slightly underperforms compared to the base FLUX.1-dev model in visual presentation metrics.

5.5 Hyperparameter Analysis

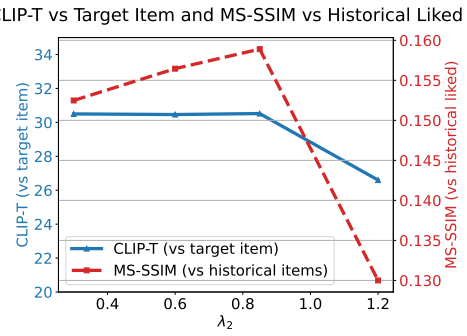

Figure 3: The impact of λ_2 on E-ReDial dataset.

We analyze the impact of λ_1 and λ_2 on the generated results. In our experiments, we fix $\lambda_1 = 1.0$ and vary λ_2, which controls the background strength. Overall, increasing λ_2 improves both fidelity and personalization. However, when $\lambda_2 > 1$, the performance drops significantly. Notably, the CLIP-T vs. target item score remains stable for $\lambda_2 < 1$, but shows a clear decline at $\lambda_2 = 1.2$. This is likely because the background becomes too dominant, leading to visual clutter, or what we refer to as chaotic "patterned images." Thus, selecting an appropriate λ_2 is crucial to balance fidelity and personalization for different subjects. Further hyperparameter analysis of it, and the impact of training steps on LoRA are provided in Appendix F.1.

5.6 Visual Comparison

In Figure 4, we present the generation results of five models based on four conversations. Two of the original conversations are shown in Appendix J. It can be observed that PCG achieves a good balance between fidelity to the original image and personalization.

For example, in the second row, with the case of *Superbad*, the user's favorite movies include typical teen comedies such as *Meet the Parents* and *The Hangover*. PCG preserves the high-school vibe and playful tone of *Superbad*, while enhancing the character positions and background colors in the

Figure 4: Visual Comparison of PCG with other models. Here EC represents EasyControl.

poster, making it more vibrant and youthful. In contrast, the outputs of other models show significant style changes or simply replace the background without maintaining the original atmosphere. In the fourth row, the target item is the classic *Taken* poster, which features a strong cold tone, tense atmosphere, and the iconic composition of the protagonist standing alone. PCG retains the dark tone and tension of the original *Taken* poster while integrating the frontal portrait style of historical items, effectively capturing elements of the user's inclinations.

Overall, PCG generates results that naturally combine historical interests while maintaining high fidelity to the target item. In comparison, EC-SUBJ performs similarly, but it is slightly less effective in capturing the overall style and transferring the atmosphere.

6 Conclusion

To address the lack of visual diversity in conversational systems, we propose **PCG**, a novel personalized generation framework that integrates user preferences while maintaining fidelity to the original item. Specifically, PCG leverages the strong language understanding capabilities of large models to efficiently infer user likes and dislikes, which are then used to construct personalized prompts guiding the generation process. In addition, to better incorporate visual elements that users may prefer, we introduce a visual content conditioning method. This method uses both global and local LoRAs to align the generation with the target item and historically liked items. We evaluate PCG on standard benchmarks, focusing on both fidelity to the target item and personalization based on user history. Our results demonstrate that PCG show great generation ability, which may offer a promising new direction for improving user experience in conversational systems.

Acknowledgment

This research was partially supported by the National Natural Science Foundation of China (Grants No. 62406303, 62525606), the Anhui Province Science and Technology Innovation Project (202423k09020010), and the USTC Bihe Youth Program for Interdisciplinary Innovation (BH-202518).

References

[1] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*, 2024.

[2] 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.

[3] Zhikang Dong, Xiulong Liu, Bin Chen, Pawel Polak, and Peng Zhang. Musechat: A conversational music recommendation system for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12775–12785, 2024.

[4] Zhongyi Fan, Zixin Yin, Gang Li, Yibing Zhan, and Heliang Zheng. Dreambooth++: Boosting subject-driven generation via region-level references packing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11013–11021, 2024.

[5] Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135*, 2024.

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

[7] Hamed Hematian Hemati and Hamid Beigy. Consistency training by synthetic question generation for conversational question answering. *arXiv preprint arXiv:2404.11109*, 2024.

[8] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. Group diffusion transformers are unsupervised multitask learners. 2024.

[9] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

[10] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chengru Song, Dai Meng, Di Zhang, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In *ICLR*, 2024.

[11] Sunghwan Kim, Tongyoung Kim, Kwangwook Seo, Jinyoung Yeo, and Dongha Lee. Stop playing the guessing game! target-free user simulation for evaluating conversational recommender systems. *arXiv preprint arXiv:2411.16160*, 2024.

[12] Nils Klüwer, Irina Nalis, and Julia Neidhardt. Context over categories: Implementing the theory of constructed emotion with llm-guided user analysis. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[13] Ivica Kostric, Krisztian Balog, and Filip Radlinski. Generating usage-related questions for preference elicitation in conversational recommender systems. *ACM Transactions on Recommender Systems*, 2(2):1–24, 2024.

[14] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024*, pages 3367–3378, 2024.

[15] Kunxi Li, Zhonghua Jiang, Zhouzhou Shen, ZhaodeWang ZhaodeWang, Chengfei Lv, Shengyu Zhang, Fan Wu, and Fei Wu. MadaKV: Adaptive modality-perception KV cache eviction for efficient multimodal long-context inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13306–13318, Vienna, Austria, July 2025. Association for Computational Linguistics.

[16] Li Liu and Paul Fieguth. Texture classification from random features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):574–586, 2012.

[17] Yuanxing Liu, Weinan Zhang, Baohua Dong, Yan Fan, Hang Wang, Fan Feng, Yifan Chen, Ziyu Zhuang, Hengbin Cui, Yongbin Li, et al. U-need: A fine-grained dataset for user needs-centric e-commerce conversational recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2723–2732, 2023.

[18] Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6808–6817, 2024.

[19] Julia Masche and Nguyen-Thinh Le. A review of technologies for conversational systems. In *Advanced Computational Methods for Knowledge Engineering: Proceedings of the 5th International Conference on Computer Science, Applied Mathematics and Applications, ICCSAMA 2017 5*, pages 212–225. Springer, 2018.

[20] Tendai Mukande, Esraa Ali, Annalina Caputo, Ruihai Dong, and Noel E O'Connor. Mmcrec: Towards multi-modal generative ai in conversational recommendation. In *European Conference on Information Retrieval*, pages 316–325. Springer, 2024.

[21] Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. *arXiv preprint arXiv:2503.02450*, 2025.

[22] Mathieu Ravaut, Hao Zhang, Lu Xu, Aixin Sun, and Yong Liu. Parameter-efficient conversational recommender system as a language processing task. *arXiv preprint arXiv:2401.14194*, 2024.

[23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

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

[25] Yueming Sun and Yi Zhang. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 235–244, 2018.

[26] Cong Wang, Jiaxi Gu, Panwen Hu, Haoyu Zhao, Yuanfan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. Easycontrol: Transfer controlnet to video diffusion for controllable generation and interpolation. *arXiv preprint arXiv:2408.13005*, 2024.

[27] Lingzhi Wang, Shafiq Joty, Wei Gao, Xingshan Zeng, and Kam-Fai Wong. Improving conversational recommender system via contextual and time-aware modeling with less domain-specific knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[28] Xianquan Wang, Likang Wu, Shukang Yin, Zhi Li, Yanjiang Chen, Hufeng Hufeng, Yu Su, and Qi Liu. I-am-g: Interest augmented multimodal generator for item personalization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21303–21317, 2024.

[29] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1929–1937, 2022.

[30] Pontus Wärnestål. User evaluation of a conversational recommender system. In *Proceedings of the 4th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 32–39, 2005.

[31] Yuxiang Wei, Yiheng Zheng, Yabo Zhang, Ming Liu, Zhilong Ji, Lei Zhang, and Wangmeng Zuo. Personalized image generation with deep generative models: A decade survey. *arXiv preprint arXiv:2502.13081*, 2025.

[32] Feng Wu, Guoshuai Zhao, Tengjiao Li, Jialie Shen, and Xueming Qian. Improving conversation recommendation system through personalized preference modeling and knowledge graph. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[33] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.

[34] Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, pages 264–274, 2025.

[35] Heng Yu, Junfeng Kang, Rui Li, Qi Liu, Liyang He, Zhenya Huang, Shuanghong Shen, and Junyu Lu. CA-GAR: Context-aware alignment of LLM generation for document retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5836–5849, Vienna, Austria, July 2025. Association for Computational Linguistics.

[36] Subiya Zaidi, Sawan Rai, and Kapil Juneja. A review of existing conversational recommendation systems. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pages 22–26. IEEE, 2024.

[37] Yi Zhan, Qi Liu, Weibo Gao, Zheng Zhang, Tianfu Wang, Shuanghong Shen, Junyu Lu, and Zhenya Huang. Coderagent: Simulating student behavior for personalized programming learning with large language models. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 293–301. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Main Track.

[38] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 84–93, 2024.

[39] Lixi Zhu, Xiaowen Huang, and Jitao Sang. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. In *Proceedings of the ACM on Web Conference 2025*, pages 4653–4661, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and the introduction, we clearly state our contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We reported our limitations in Section A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We reported our theory assumptions and proofs in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Method and Experimental details are discussed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code link in the Introduction section.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Dataset information is displayed in Section 5.1, while implementation details such as hyperparameters, are discussed in Section 5.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the experiments compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper conforms to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We've declared it in Appendix B.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The models and data used in this paper are open source.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used are cited in the References.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the code link in the Introduction section.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: As described in the main text, our proposed PCG framework utilizes a large language model as a core component, the *Inclinations Analyzer*.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Limitations

Although our method demonstrates its effectiveness through subjective and objective evaluations, hyperparameter analysis, ablation experiments, and result visualizations, it has some limitations. The approach is highly dependent on the quality of the original images (fortunately, the images used in this study are of high resolution), the ability of the base LLM for interest analysis, and its relatively slow generation speed, which can be a challenge for real-time conversations.

We hope that the ideas presented in this paper will offer new insights for personalized content generation in conversational systems, pave the way for further developments, and inspire more excellent work. In our future work, we will also focus on refining the generation of human-related details and addressing the current issue where imperfect content fidelity may cause generated outputs to stray somewhat from the original intent.

B Broader Impact

The proposed visual content generation framework, PCG, demonstrates a significant advancement in personalized content synthesis by leveraging user preferences to create visually compelling and contextually aligned content. By integrating liked items with the target movie, PCG enhances visual consistency and thematic relevance, bridging the gap between user-specific taste and cinematic representation. This approach not only amplifies user engagement but also introduces novel ways for users to interact with media content, offering a more immersive and personalized experience.

The main advantage of PCG lies in its ability to accurately capture thematic elements from both the target movie and the user's preferred visual styles. Unlike traditional generation methods that merely focus on stylistic transfer, PCG effectively balances semantic alignment with visual aesthetics, leading to outputs that are not only artistically pleasing but also contextually meaningful. This capability opens up new opportunities in content marketing, fan engagement, and interactive media, allowing for dynamically generated visuals that resonate with individual tastes.

However, the method also introduces potential ethical considerations. Personalized content generation, if not regulated properly, may inadvertently promote biased representations based on user history, reinforcing stereotypes or unbalanced views. Furthermore, the use of copyrighted materials for personalized synthesis raises questions about intellectual property rights and fair use, particularly when generated content resembles original artworks. Addressing these challenges requires clear guidelines on data privacy, responsible use of AI in media creation, and proper attribution for derivative works to ensure ethical and fair application.

C Theoratical Support for PCG

C.1 Flow Matching Loss Formulation

C.1.1 Loss Analysis on Global LoRA

We provide the flow matching loss on Global LoRA as:

$$\mathcal{L}_{\text{style}} = \mathbb{E}_{t,\, x_0 \sim \mathcal{N}(0,I),\, x_1} \left\| u_{\theta+\phi_{\text{style}}}(x_t, t) - (x_1 - x_0) \right\|_2^2 . \tag{12}$$

This loss function trains the LoRA parameters ϕ_{style} to predict the velocity field that transforms noise samples $x_0 \sim \mathcal{N}(0, I)$ into data samples $x_1 \sim p_{\text{data}}$. The intermediate state x_t is defined through linear interpolation:

$$x_t = (1 - t)x_0 + tx_1. \tag{13}$$

The gradient with respect to the style LoRA parameters ϕ_{style} is computed as:

$$\nabla_{\phi_{\text{style}}} \mathcal{L}_{\text{style}} = \mathbb{E}_{t, x_0 \sim \mathcal{N}(0,I), x_1} \left[2(u_{\theta+\phi_{\text{style}}}(x_t, t) - (x_1 - x_0)) \cdot \nabla_{\phi_{\text{style}}} u_{\theta+\phi_{\text{style}}}(x_t, t) \right] . \tag{14}$$

This gradient drives the parameter updates through gradient descent:

$$\frac{d\phi_{\text{style}}}{d\tau} = -\nabla_{\phi_{\text{style}}} \mathcal{L}_{\text{style}}. \tag{15}$$

Under the LoRA parameterization $\phi_{\text{style}} = BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$, the gradients decompose as:

$$\nabla_A \mathcal{L}_{\text{style}} = B^T \nabla_{\phi_{\text{style}}} \mathcal{L}_{\text{style}}, \tag{16}$$

$$\nabla_B \mathcal{L}_{\text{style}} = \nabla_{\phi_{\text{style}}} \mathcal{L}_{\text{style}} A^T. \tag{17}$$

This low-rank parameterization constrains updates to a compact subspace, enabling efficient and modular style adaptation while preserving the base model's core capabilities.

C.1.2 Application to Local LoRAs

The gradient flow for Local LoRA (Subject and Background LoRAs) is identical to that of Style LoRA and will not be repeated here.

C.2 Sequence Invariance as Flow Matching Regularization

In our work, *how sequence invariance affects the training or inference process* remains an open question. In fact, the sequence invariance strategy can be interpreted as a regularization method in the flow matching framework.

Proposition 1 (Sequence Invariance Regularization)**.** *Let $\pi \in S_N$ be any permutation of the input source sequence. Define the permuted source distribution as q_0^π. The ideal goal is:*

$$\mathcal{L}_{FM}(q_0^\pi, q_1) = \mathcal{L}_{FM}(q_0, q_1), \quad \forall \pi \in S_N, \tag{18}$$

where $\mathcal{L}_{FM}$ is the flow matching loss.

Introducing random permutation matrices P during training is equivalent to minimizing the expected flow matching cost:

$$\mathbb{E}_{\pi\sim Uniform(S_N)}[\mathcal{L}_{FM}(q_0^\pi, q_1)]. \tag{19}$$

This encourages the model to learn representations that are invariant to input orderings.

C.2.1 Group Theory Perspective on Sequence Invariance

From a group theory standpoint, sequence invariance corresponds to invariance under the symmetric group S_N.

A function f is sequence-invariant if and only if:

$$f(P \cdot x) = f(x), \quad \forall P \in \mathcal{P}_N. \tag{20}$$

This can be enforced using the Reynolds operator:

$$\hat{f}(x) = \frac{1}{N!} \sum_{\pi\in S_N} f(P_\pi \cdot x). \tag{21}$$

This suggests that random permutations during training project the model toward the invariant function space.

C.3 The Choice of λ_1 and λ_2.

Let $\lambda_1 = 1$ and $\lambda_2 = 0.85$ be the fusion weights used during inference. Here we prove why we could set the λ_1 for Global LoRA and change λ_2. These settings can be justified using Bayesian perspective and Information Bottleneck view.

C.3.1 Bayesian Optimal Weight Allocation for Multimodal Information Fusion

To explain why we set $\lambda_1 > \lambda_2$, we analyze the theoretical rationale behind this choice from a Bayesian perspective. Given subject data D_{subj} and background data D_{bg}, the target is to estimate the posterior $p(x|D_{\text{subj}}, D_{\text{bg}})$. By Bayes' rule:

$$p(x|D_{\text{subj}}, D_{\text{bg}}) \propto p(D_{\text{subj}}|x)p(D_{\text{bg}}|x)p(x). \tag{22}$$

Taking logarithms and it could be obtained:

$$\log p(x|D_{\text{subj}}, D_{\text{bg}}) \propto \log p(x) + \log p(D_{\text{subj}}|x) + \log p(D_{\text{bg}}|x). \tag{23}$$

Assuming these are parameterized by neural networks, the optimal fusion weights are inversely proportional to the conditional entropy:

$$\lambda_1 \propto \frac{1}{\mathcal{H}(D_{\text{subj}}|x)}, \quad \lambda_2 \propto \frac{1}{\mathcal{H}(D_{\text{bg}}|x)}. \tag{24}$$

Since subject data usually provides more certain information, this supports the setting $\lambda_1 = 1 > \lambda_2 = 0.85$.

C.3.2 Information Bottleneck View of Fusion

To explain why we fix $\lambda_1 = 1$ and vary λ_2, we provide a proof from the Information Bottleneck perspective, showing that the ratio between the two (rather than their specific values) has the greatest impact on the fusion data. From the information bottleneck principle, optimal fusion maximizes relevant information while minimizing redundancy:

$$\max_{\phi} I(Z;Y) - \beta I(Z;X), \tag{25}$$

where X represent the input data, Y represent the target output, and Z represent the learned representation. We define β as an entropy-weighted trade-off parameter. Then the ratio of fusion weights satisfies:

$$\frac{\lambda_1}{\lambda_2} \approx \frac{I(Z_{\text{subj}};Y)}{I(Z_{\text{bg}};Y)} \cdot \frac{1-\beta_{\text{bg}}}{1-\beta_{\text{subj}}}. \tag{26}$$

D Pseudo Code

D.1 Training Pseudo Code

Algorithm 1 Training Process with Sequence Invariance

Input: N item images $\{I_i\}_{i=1}^{N}$, historical content
Step 1: Spatially concatenate images to form $I_{\text{cat}} \in \mathbb{R}^{H\times(N\cdot W)\times C}$
Step 2: Generate permutation matrix P via:
    Sample $\pi \sim S_N$ & construct P where $P_{ij} = \mathbb{I}[\pi(j) = i]$
    Apply shuffling: $I_{\text{cat, shuffled}} = P \cdot I_{\text{cat}}$
Step 3: Train Global LoRA ϕ_{style}:
    Freeze base model parameters θ
    Optimize ϕ_{style} using flow-matching loss $\mathcal{L}_{\text{style}}$
Step 4: Train Local LoRAs ϕ_{subj} and ϕ_{bg}:
    Freeze ϕ_{style}
    Optimize ϕ_{bg} using local loss $\mathcal{L}_{\text{local}}$
Step 5: Update model parameters θ and LoRAs using gradient descent

D.2 Inference Pseudo Code

The inference phase involves several modules, including user inclinations analysis, integration of historical items, as well as flow matching transformation and decoding processes. The following pseudo code illustrates the main parts of these processes in detail.

Algorithm 2 Inference Process with Global and Local LoRA

Input: Dialogue history C, reference images $\{I_{\text{ref}}\}$ including $\{I_{\text{cat}}\}$
Step 1: Load pretrained components:
  Base model parameters θ (FLUX.1-dev)
  Global LoRA ϕ_{style} (frozen, pretrained)
  Local LoRAs: ϕ_{subj}, ϕ_{bg} (trained)
Step 2: Set LoRA weights:
  $\lambda_1 = 1.0$ // Subject LoRA weight
  $\lambda_2 = 0.85$ // Background LoRA weight (supplementary)
Step 3: Analyze inclinations:
  $F_\Theta(C) = \text{InclinationsAnalyzer}(C)$
Step 4: Initialize noise and timestep:
  Sample $x_0 \sim \mathcal{N}(0, I)$ // Start from Gaussian noise at $t = 0$
  Set timestep schedule $\{t_i\}_{i=0}^{T}$ with $t_0 = 0$, $t_T = 1$, $t_i \in \mathbb{R}$
Step 5: Flow Matching inference loop:
for $i = 0$ **to** $T - 1$ **do**
  Combine LoRA parameters:
    $\phi_{\text{combined}} = \phi_{\text{style}} \cup \lambda_1 \cdot \phi_{\text{subj}} \cup \lambda_2 \cdot \phi_{\text{bg}}$
  Apply EasyControl framework:
    Inject personalized prompt tokens to both subject & background branches with $\{I_{\text{ref}}\}$
    Ensure subject/background tokens interact within own content
  Predict velocity field:
    $u_{t_i} = u_{\theta+\phi_{\text{combined}}}(x_{t_i}, t_i, F_\Theta(C))$ // $t_i \in \mathbb{R}$
  Update latent via Euler method:
    $\Delta t = t_{i+1} - t_i$
    $x_{t_{i+1}} = x_{t_i} + \Delta t \cdot u_{t_i}$
end for
Step 6: Decode final image:
  $I_{\text{generated}} = \text{Decoder}(x_1)$ // x_1 is the sample at $t = 1$
Output: Generated image $I_{\text{generated}}$

E Contextual Understanding Through Concatenation

We suggest that concatenation allows the DiTs model to capture the semantic information of previously liked items. Although [9] provides empirical evidence of its effectiveness, it lacks a deep analysis of the underlying mechanism. Here, we explain how DiTs integrates semantic context through concatenation by analyzing how the transformer's attention mechanism interprets background information.

E.1 Context Modeling Ability of Transformer Attention Mechanism

DiT, which is based on the Transformer architecture, leverages the Self-Attention mechanism that naturally captures long-range dependencies. When input sequences, such as image features or text tokens, are concatenated, each token can attend to every other token in the sequence through attention weights. This process enables the establishment of global semantic relationships. Specifically, the model computes a Query vector for each token, which is compared to the Key vectors of all other tokens to calculate similarities. The resulting similarity scores are used to weight and aggregate the corresponding Value vectors. This allows the model to focus on key information within the sequence, such as object features or scene styles, and incorporate this information into the current task. Additionally, the multi-head attention mechanism decomposes the attention computation into multiple subspaces, each focusing on different dimensions of context (*e.g.*, color, spatial arrangement). These multi-dimensional semantic insights are then integrated to generate coherent outputs.

E.2 Input Concatenation Strategy Activating the Model's Contextual Representation Ability

DiT activates the model's contextual representation ability by concatenating image or text tokens, providing explicit contextual cues. Unlike traditional methods that concatenate only attention

tokens, DiT directly concatenates multiple images into a single input. This approach preserves both spatial and semantic relationships between images in the feature space. It enables the model to understand the overall theme of the image set and maintain consistency in style and logic during generation. In the case of PCG's Personalized prompt design, the model extracts a global semantic framework through a text encoder and transforms it into contextual constraints for image generation. Furthermore, pretraining allows the model to link abstract concepts in text with image features. When multiple image prompts are provided, the model can use its pretraining knowledge to infer the semantic relationships between the images. As a result, pretraining DiT enables efficient multi-image understanding through *In-Context Learning*.

F More Experiments

F.1 Hyperparameter Analysis

In this section, we conduct an additional hyperparameter analysis.

F.1.1 Impact of Local LoRA Training Steps on Generation Quality

Table 6: The influence of Local LoRA training steps on E-ReDial.

Steps	Historical Liked		Target					FID↓
	MS-SSIM↑	CLIP-T↑	DIS↑	CLIP-T↑	CLIP-I↑	MS-SSIM↑	LPIPS↓	
100	0.1470	26.38	86.15	30.39	76.01	0.1768	0.6354	21.91
300	0.1568	26.64	85.94	30.37	74.76	0.1770	0.6448	22.89
500	0.1542	26.54	86.23	30.30	75.47	0.1773	0.6393	22.67
700	0.1511	26.68	86.03	30.48	74.71	0.1689	0.6445	22.86

The number of training steps for LoRA significantly affects the generation results. If too large, it may lead to overfitting, while if too small, it may fail to capture the fusion of background information, lacking relevant capabilities. Therefore, during training, we select the most suitable checkpoint based on the generation performance on the validation set, which is then used for generation on the test set.

From Table 6, it can be observed that as the number of training steps increases, the model's ability to fuse historical liked items initially increases and then decreases, but overall it shows an upward trend. Comparing the MS-SSIM and CLIP-T scores relative to historical items, we see an overall improvement in history fusion capability compared to training for 100 steps. However, there is a noticeable decline in subject fidelity, as shown by the decrease in DIS and CLIP-related metrics for the target item. This indicates that achieving a balance between personalization and fidelity requires selecting the most appropriate training steps based on the validation set for different tasks.

F.1.2 Deeper Impact of Variations in λ_1 and λ_2

In Section 5.5 of the main text, we have analyzed the impact of different values of λ_2 on quantitative metrics, and here we present its effect on generation quality. As shown in Figure 6, the generated results vary significantly with changes in λ_2. When λ_2 is around 0.3, the generated image incorporates some information from the liked items, but it still primarily reflects elements related to the target item. As λ_2 increases to around 0.6–0.8, a balance is achieved between fidelity to the target and personalization. However, when λ_2 exceeds 1, the generated result is mainly a combination of liked items, as the strong guidance from background information limits the reflection of the original target item.

F.2 Ablation Studies

In this chapter, we conduct various ablation studies on components and strategies, such as Sequence Invariance, to demonstrate the effectiveness of each part.

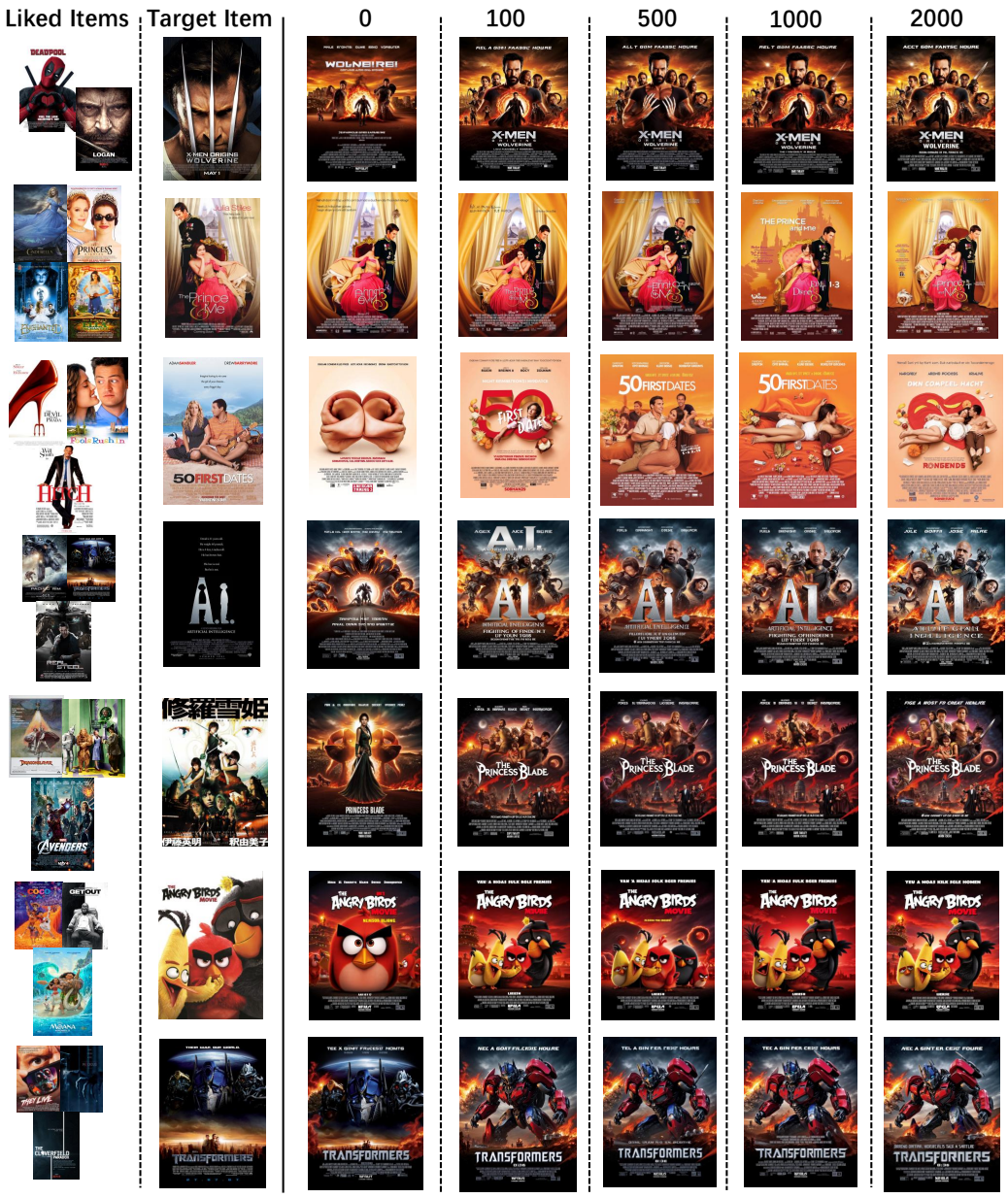

Figure 5: Generation results with different training steps for Local LoRA.

F.2.1 The Impact of Sequence Invariance

In this experiment, we examine the impact of using (w/) and not using (w/o) Sequence Invariance (SI) on quantitative metrics, under the setting of training for 1000 steps with Local LoRA. As shown in Table 7, after applying SI, the generated images better integrate historical information, achieving higher scores on MS-SSIM and CLIP-T. Additionally, SI improves fidelity, as reflected in better performance on metrics like SSIM. While these differences may not be immediately noticeable in the generated results, the metrics clearly demonstrate the effectiveness of SI.

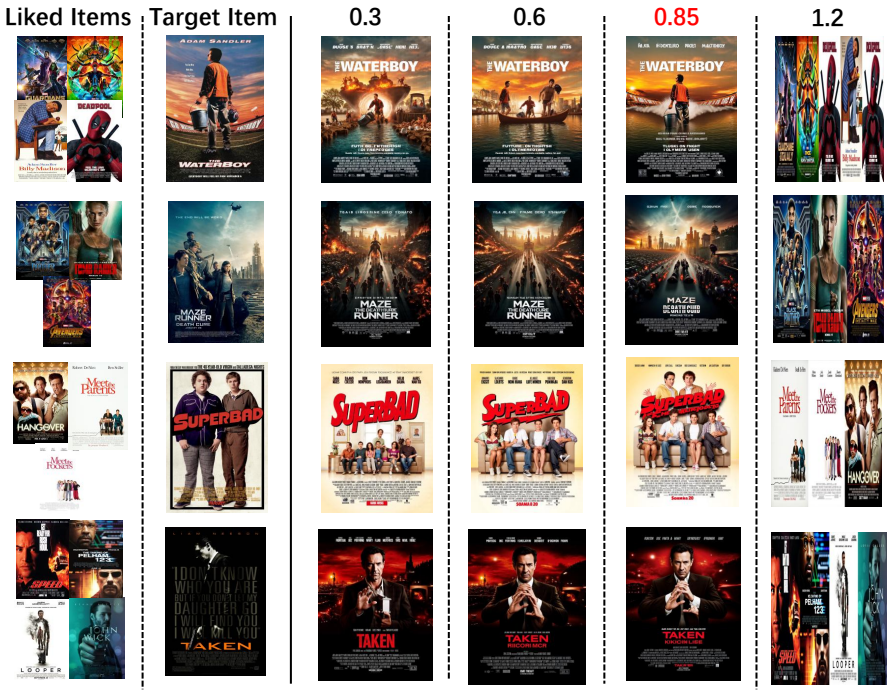

Figure 6: The generation results with different λ_2.

Table 7: The influence of Sequence Invariance on E-ReDial.

Method	Historical Liked		Target					FID↓
	MS-SSIM↑	CLIP-T↑	DIS↑	CLIP-T↑	CLIP-I↑	MS-SSIM↑	LPIPS↓	
1000 steps w/o SI	0.1464	26.53	85.92	30.33	75.67	0.1670	0.6496	22.68
1000 steps w/ SI	0.1488	26.69	85.92	30.34	75.52	0.1710	0.6487	22.49

F.2.2 The Impact of Main Components

In this section, we focus on the impact of the following three main components on generation quality metrics: Global LoRA, Negative Inclinations and Score Reweighting. These components play a crucial role in both the fidelity and personalization of the generated results.

Table 8: The ablation study of E-ReDial.

Method	Historical Liked		Target					FID↓
	MS-SSIM↑	CLIP-T↑	DIS↑	CLIP-T↑	CLIP-I↑	MS-SSIM↑	LPIPS↓	
w/o Global LoRA	0.1521	26.45	87.29	29.11	71.31	0.1890	0.5993	21.13
w/o Negative Prompt	0.1601	26.47	85.63	30.55	73.89	0.1945	0.6257	21.98
w/o Score Reweight	0.1594	26.46	85.72	30.39	73.39	0.1970	0.6250	21.92

From the ablation study in Table 8, we find that Global LoRA is the most important component, playing an essential role in both personalization and fidelity. This is because generating movie posters requires a strong community/pre-trained LoRA to provide relevant guidance; without it, the model struggles to meaningfully connect content with the poster. Next, the Score Reweighting mechanism significantly impacts the fidelity of the generated results. Without Score Reweighting, the generated outputs, compared to target items, show a decrease in quality, and the CLIP-T score relative to historical liked items also drops. This indicates that reweighting helps generate higher-quality content. Lastly, the impact of Negative Prompt is minimal, likely because other components compensate for the negative influence on generation.

In terms of visualizing the generated results, we can clearly observe that Global LoRA has the most significant impact on the generation quality, while the effects of Negative Prompt and Score Reweighting are less apparent visually, despite the differences shown in the evaluation metrics.

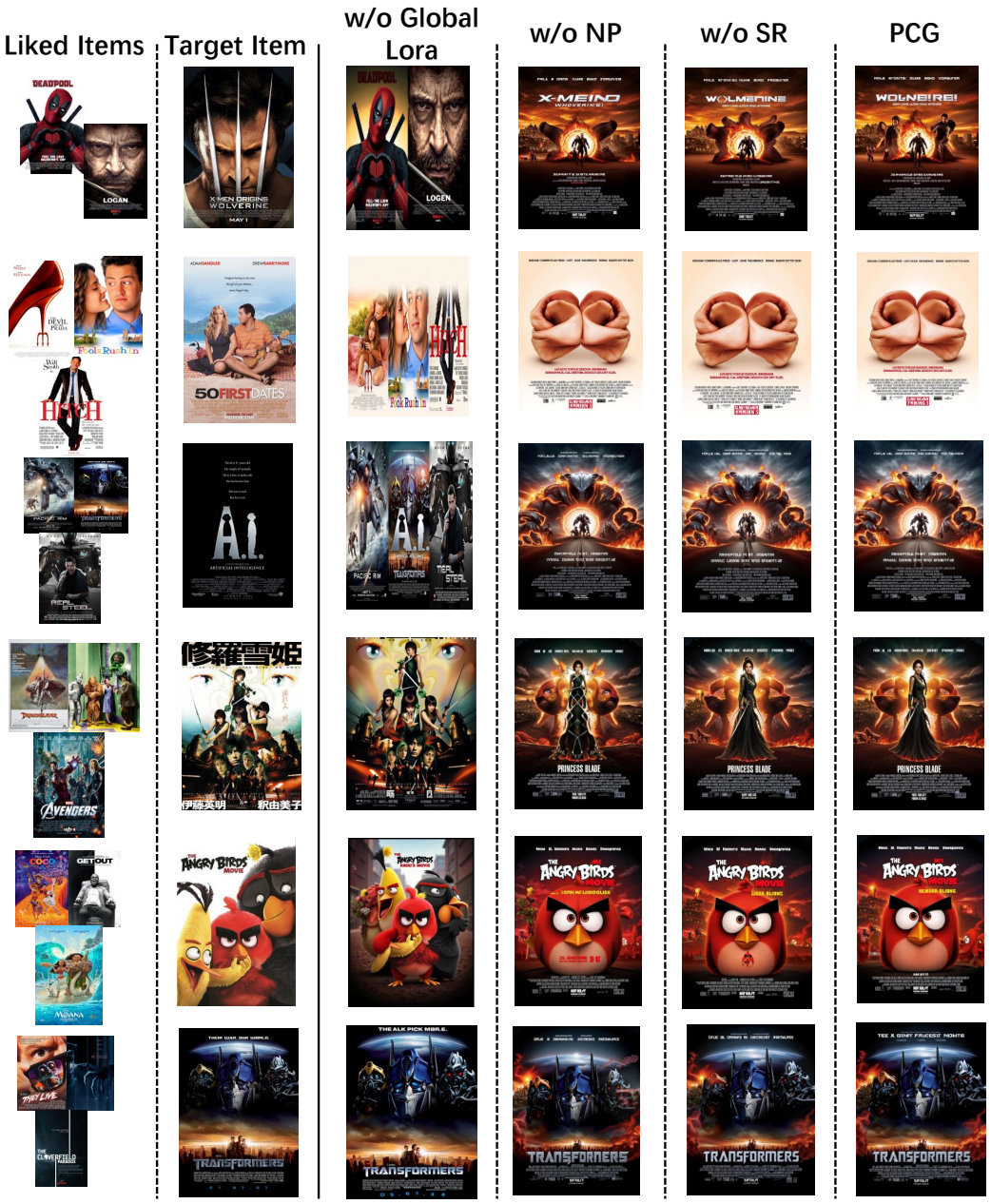

Figure 7: Ablation studies. Here NP represents Negative Prompt, and SR represents Score Reweighting.

G Implementation Details

As mentioned in Section H, we use Qwen3-8B [3] as the LLM to generate user inclinations and GPT-4o for evaluation. Here, we provide a detailed description of the configuration used for Qwen model. We

[3] https://huggingface.co/Qwen/Qwen3-8B

generate outputs with the following parameters: a maximum of 128 new tokens, sampling enabled with a temperature of 0.7, top-p sampling with a probability of 0.8, top-k sampling with a limit of 20, and a minimum probability of 0.0.

When training PCG LoRA based on EasyControl, we strictly follow its recommended settings. The overall learning rate is set to 1×10^{-4} (based on the FLUX.1-dev pre-trained model, using a single A100-80G GPU with a batch size of 1). The two types of LoRAs in the two-stage training share the same learning rate.

The optimizer used is AdamW with the parameters:

$$\beta_1 = 0.9, \quad \beta_2 = 0.999, \quad \text{weight decay} = 1 \times 10^{-4}.$$

The dimension of the low-rank matrices is set to 128.

H Inclinations Analyzer Prompt

```
Role: You are an expert. Please extract the user's movie-related
    common styles, colors, types words and their probabilities from
    the following dialogue. Output format: a list, each element is [
    descriptive word, probability], probability is a float between 0
    and 1. Note that the descriptive word should NOT be the movie
    names. For example, a user liked movie Forrest Gump, the output
    should be like this: [['encourage', 0.8], ['comedy', 0.3]].

Dialogue content: {dialogue_text}

Role: You are an expert. Please extract the user's disliked movie-
    related common styles, colors, types words and their probabilities
     from the following dialogue. Output format: Some words that the
    user dislikes, each word is a string. If the user doesn't dislike
    any movie, please return an empty list. For example, a user
    disliked movie Saving Private Ryan, the output may be like this:
    war, horror.

Dialogue content: {dialogue_text}
```

Listing 1: Few-shot prompt template for inclinations analysis.

I GPT-4o Evaluation Prompt

```
Role: I first provide you some historical liked images and the target
    item image. Then I will provide you with a set of movie posters.
    Please evaluate them based on the following four dimensions, with
    each dimension being rated on a scale of 1 to 5. Be objective and
    fair in your analysis and provide clear reasoning for each rating.
     Finally, calculate the total score and rank the posters
    accordingly. The evaluation dimensions are as follows:

1. Visual Aesthetics
Are the color combinations harmonious and visually impactful?

Is the composition balanced and layered?

Does it have the quality and feel of a cinematic movie poster?

2   . Integration Level
Are the elements from the Target Item and History Liked Items
    naturally integrated? Like the colors, elements and roles of
    History Liked Items.
```

```
Does it appear seamless and unified, as if they belong to the same
    universe?

3   . Thematic Coherence
Does it successfully convey the main storyline of the Target Item
    while highlighting the characteristics of the Liked Items?

Is the background story immediately understandable to the viewer?

4   . Detail Representation
Are the background, characters, lighting, and text details well-
    refined?

Are there any obvious signs of poor blending or noticeable seams?

Scoring Criteria:
1: Very Poor - Completely uncoordinated

2: Poor - Noticeable flaws

3: Fair - Basic integration but lacks refinement

4: Good - Visually consistent and well-detailed

5: Excellent - Perfectly blended with cinematic quality
```

Listing 2: Prompt for GPT Evaluation.

J Original Conversations


```
[
  {
    "assistant": "Hi",
    "user": "Hello! How are you? Could you recommend some comedy's? I'
    ve had a tough day."
  },
  {
    "assistant": "I recommend [Meet the Parents (2000)], since  i t s
    a funny comedy and it can make you laugh. It tells the story of
    Greg, a future son-in-law and male nurse, who meets his strict
    father-in-law."
  },
  {
    "user": "No, I haven't."
  },
  {
    "assistant": "You have to see it. If you're interested in more
    family comedies, also try [Meet the Fockers (2004)]  the
    hilarious sequel. Another great option is [The Hangover (2009)],
    where three buddies wake up from a bachelor party in Las Vegas
    with no memory and the groom missing."
  },
  {
    "user": "I will check them out."
  },
  {
    "user": "Oh, [The Hangover (2009)] was funny"
  },
  {
    "assistant": "I recommend [Superbad (2007)], since it's hilarious
    and funny. It's about three high schoolers trying to bring alcohol
     to a party to impress their peers. I'm sure after watching, you
    can't stop laughing."
```

```
  },
  {
    "user": "Have not seen that either. Will have to check it out."
  },
  {
    "user": "Thank you for the recommendations"
  },
  {
    "assistant": "You're welcome, take care"
  }
]
```

Listing 3: Conversation for Row-3

```
[
  {
    "user": "Hey!",
    "assistant": "Hi!"
  },
  {
    "user": "I am in need of some good action movies, not superhero
    movies though or [The Fast and the Furious] movies. I love movies
    like [Speed (1994)] with Keanu Reeves, or [The Taking of Pelham
    123 (2009)] with Joh Travolta"
  },
  {
    "user": "Any suggestions?"
  },
  {
    "assistant": "Since you like action movies, then I think this
    movie will suit your taste: [Looper (2012)]. The sci-fi is not
    foolproof as the director would like it to be, but the film is
    appealing and entertaining enough with a nice blend of action and
    thrills. I believe that you will enjoy yourself while watching it
    ."
  },
  {
    "user": "I have with Bruce Willis! That one is such a great movie
    !"
  },
  {
    "assistant": "If you're looking for great action movies, I
    recommend [John Wick (2014)]. It's about a first-class killer who
    is forced to go back to work to get revenge after his retirement.
    It's one of Keanu's best action movies, so you should really try
    it!"
  },
  {
    "user": "I've heard of it and I know the sequel came out, but have
     never seen it"
  },
  {
    "user": "What is that about?"
  },
  {
    "assistant": "As you have watched [Looper (2012)], then you can
    try [John Wick (2014)]. It's one of Keanu's best! It was directed
    by Chad Stahelski, David Leitch. Other actors include Michael
    Nyqvist, Alfie Allen and Willem Dafoe. I believe this movie will
    give you an unprecedented experience. It's a similar storyline to
    [Taken (2008)]. A man seeking vengeance for his loved one. It's
    really good."
  },
  {
```

```
    "user": "That one sounds really good!"
  },
  {
    "user": "I think the story line is really good thank you"
  },
  {
    "assistant": "You're welcome!"
  },
  {
    "assistant": "I hope you check it out! Enjoy your day!"
  }
]
```

Listing 4: Conversation for Row-4