IterMeme: Expert-Guided Multimodal LLM for Interactive Meme Creation with Layout-Aware Generation

Yaqi Cai[1] , **Shancheng Fang**[2*] , **Yadong Qu**[1] , **Xiaorui Wang**[2] , **Meng Shao**[1] , **Hongtao Xie**[1]
[1]University of Science and Technology of China
[2]YuanShi Technology
{cyaqi, qqqyd}@mail.ustc.edu.cn, {fangsc, mshao1, htxie}@ustc.edu.cn, harrywxr@outlook.com

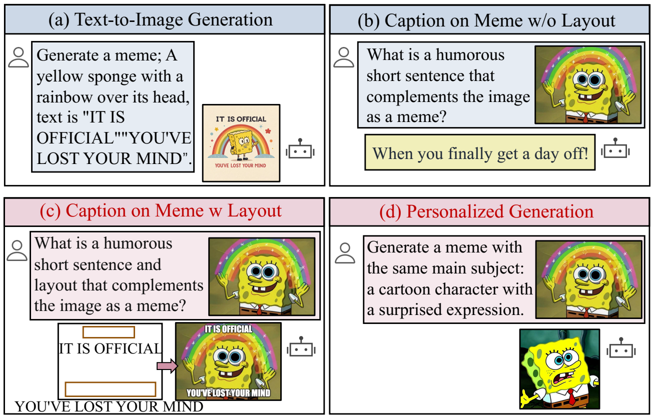
Abstract

Meme creation is a creative process that blends images and text. However, existing methods lack critical components, failing to support intent-driven caption-layout generation and personalized generation, making it difficult to generate high-quality memes. To address this limitation, we propose IterMeme, an end-to-end interactive meme creation framework that utilizes a unified Multimodal Large Language Model (MLLM) to facilitate seamless collaboration among multiple components. To overcome the absence of a caption-layout generation component, we develop a robust layout representation method and construct a large-scale image-caption-layout dataset, MemeCap, which enhances the model's ability to comprehend emotions and coordinate caption-layout generation effectively. To address the lack of a personalization component, we introduce a parameter-shared dual-LLM architecture that decouples the intricate representations of reference images and text. Furthermore, we incorporate the expert-guided M^{3}OE for fine-grained identity properties (IP) feature extraction and cross-modal fusion. By dynamically injecting features into every layer of the model, we enable adaptive refinement of both visual and semantic information. Experimental results demonstrate that IterMeme significantly advances the field of meme creation by delivering consistently high-quality outcomes. *The code, model, and dataset will be open-sourced to the community.*

1 Introduction

Memes, as visual symbols that convey emotions through images, have become an increasingly important communication medium on social media. Consequently, automating meme creation to align with user intentions has emerged as a key research focus.

High-quality memes typically not only present engaging visual content, but also incorporate humorous text components to express users' emotions. However, existing meme generation models lack mechanisms for producing fully integrated components, posing challenges in generating high-quality memes. The incomplete components include ***auto-text-less***, denoting the inability to autonomously generate humorous visual captions for memes as shown in fig. 1 (c), and ***personalization-less***, highlighting the absence of support for personalized meme generation as shown in fig. 1 (d).


(a) Text-to-Image Generation
IT IS OFFICIAL
YOU'VE LOST YOUR MIND
(b) Caption on Meme w/o Layout
(c) Caption on Meme w Layout
IT IS OFFICIAL
YOU'VE LOST YOUR MIND
IT IS OFFICIAL
YOU'VE LOST YOUR MIND
(d) Personalized Generation

Generate a meme; A yellow sponge with a rainbow over its head, text is "IT IS OFFICIAL""YOU'VE LOST YOUR MIND".

What is a humorous short sentence that complements the image as a meme?

When you finally get a day off!

What is a humorous short sentence and layout that complements the image as a meme?

Generate a meme with the same main subject: a cartoon character with a surprised expression.

Figure 1: Examples include (a) text-to-image generation, (b) meme captioning without layout, along with incomplete components: (c) meme captioning with layout and (d) personalized generation. We integrate all these components within a unified framework.

Specifically, some advanced text-to-image (T2I) generation models [Labs, 2024] and tools [ByteDance, 2024; Ideogram, 2024; Recraft, 2024] are capable of incorporating simple text into images based on user input as shown in fig. 1 (a). However, these models require users to explicitly define the text to be generated and often struggle with precise text rendering. Meanwhile, studies [Zhong *et al.*, 2024; Chen *et al.*, 2024] leveraging the powerful understanding capabilities of MLLMs have demonstrated the ability to creatively generate humorous captions with emotional enhancements for memes as shown in fig. 1 (b). Nevertheless, these approaches lack essential layout control for captions, a critical component for effective meme creation. Furthermore, Most Existing approaches [Ye *et al.*, 2023; Wang *et al.*, 2024b] for meme personalization primarily focus on injecting additional image features into diffusion models, overlooking the interactive collaboration in meme creation. While

[*] Corresponding author.

Kosmos-G [Pan *et al.*, 2023] and CAFE [Zhou *et al.*, 2024b] have introduced interactive personalized generation methods within MLLMs, the challenge of integrating multi-task capabilities into a unified architecture remains to be explored.

To address these limitations, we innovatively propose ***IterMeme***, an end-to-end interactive meme creation framework, which leverages a unified MLLM to enable collaborative task execution and resource sharing across multiple meme creation components, creating high-quality memes through user interaction. On the one hand, to address the issue of *auto-text-less*, we empower MLLM to directly perform joint generation of captions and layouts by effectively capturing the emotional semantics embedded in memes. Additionally, We develop an automated parsing pipeline that extracts memes into structured <image, caption, layout>triples, addressing the scarcity of datasets with content-aware layouts. Using this pipeline, we construct ***MemeCap***, the first large-scale dataset that considers the correlation between captions and meme layouts. On the other hand, to tackle the issue of *personalization-less*, we innovatively propose enhancing MLLM with integrated personalization capabilities. Specifically, we implement a parameter-shared dual-LLM architecture that separately processes reference images and textual descriptions. This architecture not only efficiently decouples the complex representations within the LLM but also further captures detailed information from reference images through reconstruction. To incorporate target IP features, we introduce ***M^{3}OE***, an expert-guided module specialized in personalized generation. This module comprises a meticulously designed feature extraction module and a feature fusion module. The feature extraction module captures fine-grained IP visual representations from reference images, while the feature fusion module aligns and integrates these features with textual semantics to produce unified multimodal representations. Additionally, we introduce the feature extraction and fusion modules into every layer of the LLM, enabling dynamic adjustments to the visual and semantic information across different layers. Comprehensive evaluations demonstrate IterMeme's effectiveness in automated caption generation, precise layout control, and high-quality personalized meme creation, marking a major breakthrough in the field.

In summary, our contributions are as follows:

- We propose IterMeme, the first end-to-end interactive meme creation framework leveraging unified MLLM. The expert-guided module M^{3}OE effectively addresses the challenge of interactive meme personalization.
- We introduce MemeCap, the first large-scale dataset for meme creation, constructed via an automated parsing pipeline to address the scarcity of training data in this domain.
- Experiments demonstrate that our approach achieves competitive performance, paving the way for new possibilities in meme creation.

2 Related Work

Unified multimodal model. Multimodal models designed to unify visual tasks have recently gained significant attention within the research community. These models exploit the synergy between understanding and generation tasks, enabling mutual enhancement and overall performance improvements. For instance, studies such as Chameleon [Team, 2024] and Unified-io 2 [Lu *et al.*, 2024] tokenize multimodal data into a shared semantic space, facilitating seamless integration of visual and textual modalities. DreamLLM [Dong *et al.*, 2023] incorporates an external diffusion decoder, allowing direct sampling from the multimodal space to enable interleaved content generation. Transfusion [Zhou *et al.*, 2024a] and Show-o [Xie *et al.*, 2024] unify autoregressive text modeling with discrete diffusion modeling, utilizing a single Transformer architecture.

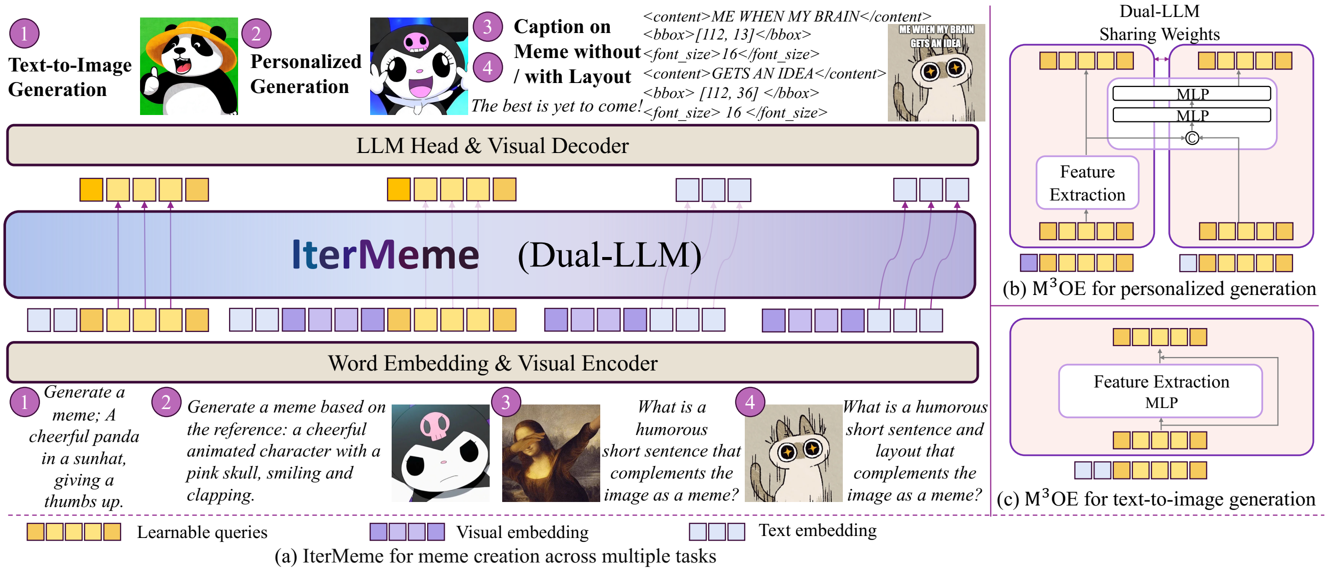
Meme caption generation. Meme captioning represents a practical application of humor generation [Amin and Burghardt, 2020]. Recent advancements in this field include CLoT [Zhong *et al.*, 2024], which explores creative captioning capabilities, and XMeCap [Chen *et al.*, 2024], which leverages multi-granularity similarities between images and captions to enhance multi-image meme captioning performance in LLMs through reinforcement learning. State-of-the-art MLLMs [Wang *et al.*, 2024a; Liu *et al.*, 2024] have also demonstrated effectiveness in addressing meme captioning tasks. However, they lack the ability to control layout, a critical aspect of meme creation.

Personalized image generation. Diffusion models [Ho *et al.*, 2020; Rombach *et al.*, 2022; Zhang *et al.*, 2024a; Zhang *et al.*, 2024b] have made significant advancements in personalized generation. For instance, DreamBooth [Ruiz *et al.*, 2023] binds special tokens to specific concepts, while Textual Inversion [Gal *et al.*, 2022] captures new concepts by learning new embeddings in the input space. However, both methods require retraining whenever a new concept is introduced, which limits their scalability. Approaches such as IPAdapter [Ye *et al.*, 2023], InstantID [Wang *et al.*, 2024b], and ConsistentID [Huang *et al.*, 2024] address this limitation by training additional image encoders, enabling inference with just a single image. Despite these improvements, these methods fall short in enabling personalized interactions. CAFE [Zhou *et al.*, 2024b] takes a step further by leveraging LLMs to process diverse or even ambiguous user inputs; however, integrating multitasking, including interactive personalization, into a unified MLLM remains challenging.

3 Methods

3.1 Problem Setup

The goal of meme creation is to generate a meme that combines visual elements, such as images and text, based on a user-provided textual description of attributes, expressions, actions, and other relevant details, along with an optional reference image. The input consists of a textual description of the meme, denoted as T, and an optional reference image. The generated meme, should include N captions $\{C_i\}_{i=0}^{N-1}$ and corresponding layouts $\{B_i\}_{i=0}^{N-1}$. If a reference image I_{ref} is provided, the generated meme must preserve the IP of the reference to ensure content consistency.


1 Text-to-Image Generation
2 Personalized Generation
3 Caption on Meme without
4 / with Layout
The best is yet to come!
<content>ME WHEN MY BRAIN</content>
<bbox>[112, 13]</bbox>
<font_size>16</font_size>
<content>GETS AN IDEA</content>
<bbox> [112, 36] </bbox>
<font_size> 16 </font_size>
ME WHEN MY BRAIN GETS AN IDEA
LLM Head & Visual Decoder
IterMeme (Dual-LLM)
Word Embedding & Visual Encoder
1 Generate a meme; A cheerful panda in a sunhat, giving a thumbs up.
2 Generate a meme based on the reference: a cheerful animated character with a pink skull, smiling and clapping.
3 What is a humorous short sentence that complements the image as a meme?
4 What is a humorous short sentence and layout that complements the image as a meme?
Learnable queries
Visual embedding
Text embedding
Dual-LLM Sharing Weights
MLP
MLP
C
Feature Extraction
Feature Extraction MLP

(a) IterMeme for meme creation across multiple tasks

(b) M^{3}OE for personalized generation

(c) M^{3}OE for text-to-image generation

Figure 2: **An overview of IterMeme.** We employ a unified MLLM for meme creation, enabling seamless collaboration across components. The expert-based M^{3}OE module (b) integrates personalized generation capabilities into the MLLM, utilizing a dual-LLM architecture to extract fine-grained IP information and achieve cross-modal alignment and fusion. Additionally, (c) M^{3}OE enhances T2I generation quality.

3.2 The Framework of IterMeme

In this section, we present a detailed overview of the proposed interactive meme creation framework, IterMeme, as shown in fig. 2. Through user interaction, the framework captures users' intentions and supports the entire workflow, spanning from meme content understanding to creative generation, while also addressing the *auto-text-less* and *personalization-less* challenges mentioned in section 1 by exploring innovative solutions.

To achieve this, we propose leveraging a unified MLLM, which introduces a special token to automatically determine when to generate an image. When the t -th token is predicted as the special token, a series of learnable queries q are used to query the LLM $\mathcal{F}_\theta$ based on the preceding input sequence $x_{<t+1}$. This process is expressed as $p := \mathcal{F}_\theta(q, x_{<t+1}, V)$, where V represents the set of visual embeddings from the images in the previous sequence. The result p serves as the conditional input for Stable Diffusion (SD) to generate images[1].

Layout representation. To enable the MLLM to directly generate layouts, an appropriate layout representation is required. Many studies [Hsu *et al.*, 2023; Yang *et al.*, 2024] represent layouts as bounding boxes defined by the top-left and bottom-right coordinates $[x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}]$. However, even slight numerical perturbations can affect character spacing and font appearance. Considering that text on memes is typically arranged horizontally with uniform letter spacing and a single font color, and to enhance the robustness of the model, we represent layouts as tuples based on the midpoint coordinates and font size: $B = ([x_{\text{center}}, y_{\text{center}}], \text{fontsize})$, where $x_{\text{center}} = \frac{x_{\text{min}}+x_{\text{max}}}{2}$, $y_{\text{center}} = \frac{y_{\text{min}}+y_{\text{max}}}{2}$, and $\text{fontsize} = y_{\text{max}} - y_{\text{min}}$. It is worth noting that, given a fixed letter spacing, these two representations are reversible.

Dual-LLM architecture. For personalized generation, the user provides inputs in the form of a reference image and a textual description $< I_{\text{ref}}, T >$. The LLM produces an embedding $p := \mathcal{F}_\theta(q, x_T, V_{\text{ref}})$, which serves as the input to SD ϵ_θ. This embedding effectively integrates semantic information from both the reference image and the text description but cannot fully capture the fine-grained visual features of the reference image. Therefore, to preserve the critical details in the reference image, it becomes essential to reconstruct it accurately. To achieve this with minimal architectural disruption, we employ a dual-LLM architecture with shared parameters, where the reference image and textual description are explicitly decoupled and fed separately into the two branches of the architecture. This results in distinct output embeddings: $p_{\text{ref}} := \mathcal{F}_\theta(q_{\text{ref}}, V_{\text{ref}})$ for the reference image and $p_T := \mathcal{F}_\theta(q_T, x_T)$ for the text description. When the target of generation is I , the training loss is defined as the sum of the mean squared errors (MSE) for the reference image reconstruction loss and the target generation loss: $L_{\text{dual}} = L_{\text{rec}} + L_{\text{gen}}$, where $L_{\text{rec}} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}\|\epsilon - \epsilon_\theta(p_{\text{ref}}, \mathcal{E}(I_{\text{ref}}), t)\|$ and $L_{\text{gen}} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}\|\epsilon - \epsilon_\theta(p_T, \mathcal{E}(I), t)\|$, with $\mathcal{E}$ representing the encoder.

M^{3}OE. To prevent the parameter updates of the reconstruction process and the target generation process from interfering with each other, and to ensure the target generation branch incorporates essential reference image information for consistent generation, we design M^{3}OE , a module dedicated to consistent personalized generation that seamlessly integrates these capabilities into the MLLM. M^{3}OE consists of a feature extraction module Φ and a feature fusion module Ψ . The feature extraction module is a multilayer perceptron (MLP)

[1] For simplicity, we omit the projection layer between the LLM and SD.

layer with residual connections, specifically designed for the reference image branch of the dual-LLM to effectively extract visual features and capture fine-grained details. The feature fusion module includes two MLP layers. The first layer is designed without residual connections, while the second incorporates residual connections to ensure stability and expressive capability during feature fusion. For its input, the concatenation of intermediate embeddings from the reference image and textual description branches in the dual-LLM architecture is used.

Multi-layer mechanism. To enable the framework to adapt dynamically to varying levels of visual and semantic information, we introduce feature extraction and fusion modules at each layer of the LLM. Specifically, for the reference image branch, the i -th layer output of the LLM for the learnable queries, which is represented as $\hat{q}_{\text{ref}}^{i+1} := \mathcal{F}_\theta^i(q_{\text{ref}}^i, V_{\text{ref}})$, is passed through the feature extraction module to obtain $q_{\text{ref}}^{i+1} = \Phi^i(\hat{q}_{\text{ref}}^{i+1})$. The resulting intermediate embedding q_{ref}^{i+1} is then concatenated with the i-th layer output of the LLM for the textual description branch $\hat{q}_T^{i+1} := \mathcal{F}_\theta^i(q_T^i, x_T)$, and fed into the feature fusion module, producing the updated embedding $q_T^{i+1} = \Psi^i(\hat{q}_T^{i+1}, q_{\text{ref}}^{i+1})$.

Training strategy. We adopt a two-stage training approach to optimize the proposed framework.

In the first stage, we utilize the MemeCap dataset introduced in section 3.3, along with a subset of the I2T category from the Oogiri-GO dataset introduced in [Zhong *et al.*, 2024]. To facilitate structured output, we introduce three pairs of special tokens to encapsulate caption content, central coordinates, and font size, effectively extending the original LLM vocabulary to constrain the output format. Detailed instruction fine-tuning templates and data transformation processes are provided in the appendix. During this stage, only the LLM is fine-tuned to ensure a comprehensive understanding of input memes, generate humorous captions, and achieve accurate layout control. The optimization process is guided solely by the language model loss.

In the second stage, we use the mixed data of T2I and personalized generation that we have constructed, with the construction process detailed in the appendix. As observed in section 4.3, for the T2I data, the feature extraction module in M^{3}OE can focus on enhancing image generation, leading to improved visual fidelity and precise text alignment. Therefore, we modify the output of M^{3}OE as follows:

$$\text{M}^3\text{OE}(q_{\text{ref}}, q_T, \tau) = \begin{cases} \Psi(\Phi(q_{\text{ref}}), q_T), & \text{if } \tau \text{ is pers. gen.,} \\ \Phi(q_T), & \text{if } \tau \text{ is T2I,} \\ \text{null}, & \text{otherwise,} \end{cases}$$

where τ represents the type of data. To enhance the model's visual representation capability while preserving the knowledge acquired by the LLM during the first stage, we train the proposed M^{3}OE, SD, and the conditional projection layer.

3.3 MemeCap Data Construction

As discussed in section 1, existing models lack the capability to independently determine the content or layout of captions. To address this limitation, we propose the MemeCap dataset in this section. Specifically, we collect a substantial dataset of Chinese and English memes from the internet. The primary challenge lies in decomposing these memes into <image, caption, layout>triples. To tackle this, we develop an innovative data construction pipeline capable of accurately separating text from images and establishing a precise one-to-one correspondence between captions and their layouts.

Image scraping and cleaning. Our pipeline selects data sources based on language. Chinese memes come from platforms like Weibo[2], while English memes are gathered from imgflip[3]. On imgflip, each meme template includes multiple user-generated memes with their corresponding upvote counts. We select the top 10 upvoted memes for each template to ensure quality. To eliminate inappropriate content such as pornography, violence, or gore, an initial filtering step uses the Qwen2-VL [Wang *et al.*, 2024a] model, followed by manual review to maintain data cleanliness and safety.

Text removal. Text removal is a critical step in data construction. To ensure high-quality text removal, we design a two-step removal process. First, a text erasure model produces an initial removal result. Then, SDXL [Podell *et al.*, 2023] inpainting refines the non-text regions and incomplete text-erased areas. The binarized mask for inpainting is generated based on the results of *Text Detection and Recognition*. Experiments in the appendix show that relying solely on text erasure causes blurred text regions, negatively impacting subsequent training, while solely using inpainting risks preserving distorted characters. The combination of these two steps effectively resolves these issues.

Text detection and recognition. Precise text detection and recognition are crucial for the two-step removal process. To extract erased text regions, we calculate the pixel differences between the original meme I_{ori} and the preliminary erasure result, followed by applying a threshold to remove edge noise, producing I_{diff}. Text recognition is then performed on both I_{ori} and I_{diff} using PaddleOCR[4], generating the corresponding text layouts B_{ori} and B_{diff}. Finally, the Intersection over Union (IoU) [Yu *et al.*, 2016] between B_{ori} and B_{diff} is calculated to filter out layouts with low overlap, while safety filtering is applied to the captions.

The MemeCap dataset constructed through this pipeline includes 13,977 English samples and 22,457 Chinese samples.

4 Experiments

4.1 Experimental Settings

Implementation details. The unified MLLM for understanding and generation is built upon DreamLLM [Dong *et al.*, 2023] as the base model, with SD2.1 [Rombach *et al.*, 2022] serving as the visual decoder. In the first training stage, the AdamW [Loshchilov, 2017] optimizer is employed with a learning rate of 1×10^{-5} and a weight decay parameter of 0. The batch size is set to 16, and gradient accumulation is performed over 2 steps to optimize computational efficiency. In

[2] https://weibo.com/

[3] https://imgflip.com/

[4] https://github.com/PaddlePaddle/PaddleOCR

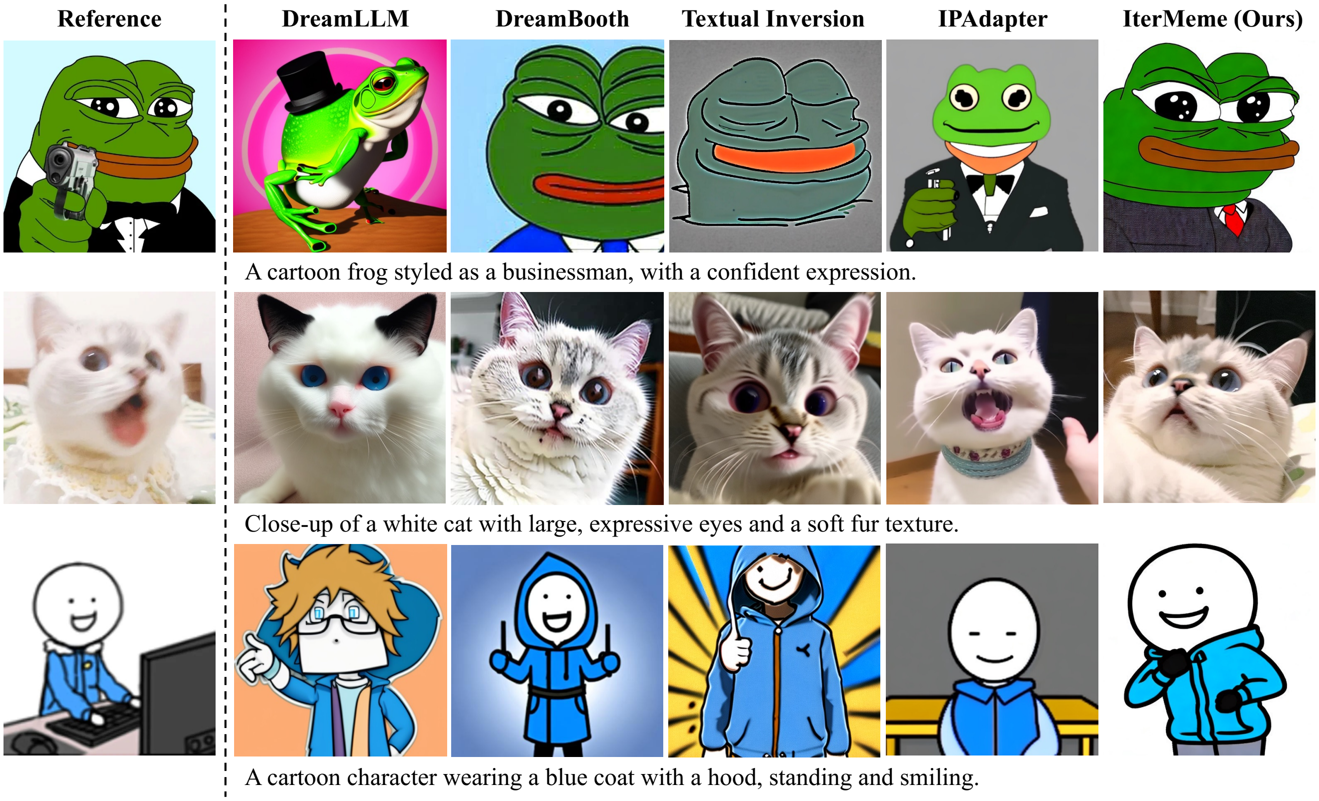

Reference
DreamLLM
DreamBooth
Textual Inversion
IPAdapter
IterMeme (Ours)
A cartoon frog styled as a businessman, with a confident expression.
Close-up of a white cat with large, expressive eyes and a soft fur texture.
A cartoon character wearing a blue coat with a hood, standing and smiling.

Figure 3: **Qualitative comparison of personalized generation.** We compare four distinct IPs and their corresponding prompts, demonstrating that our method achieves the most consistent IP generation.

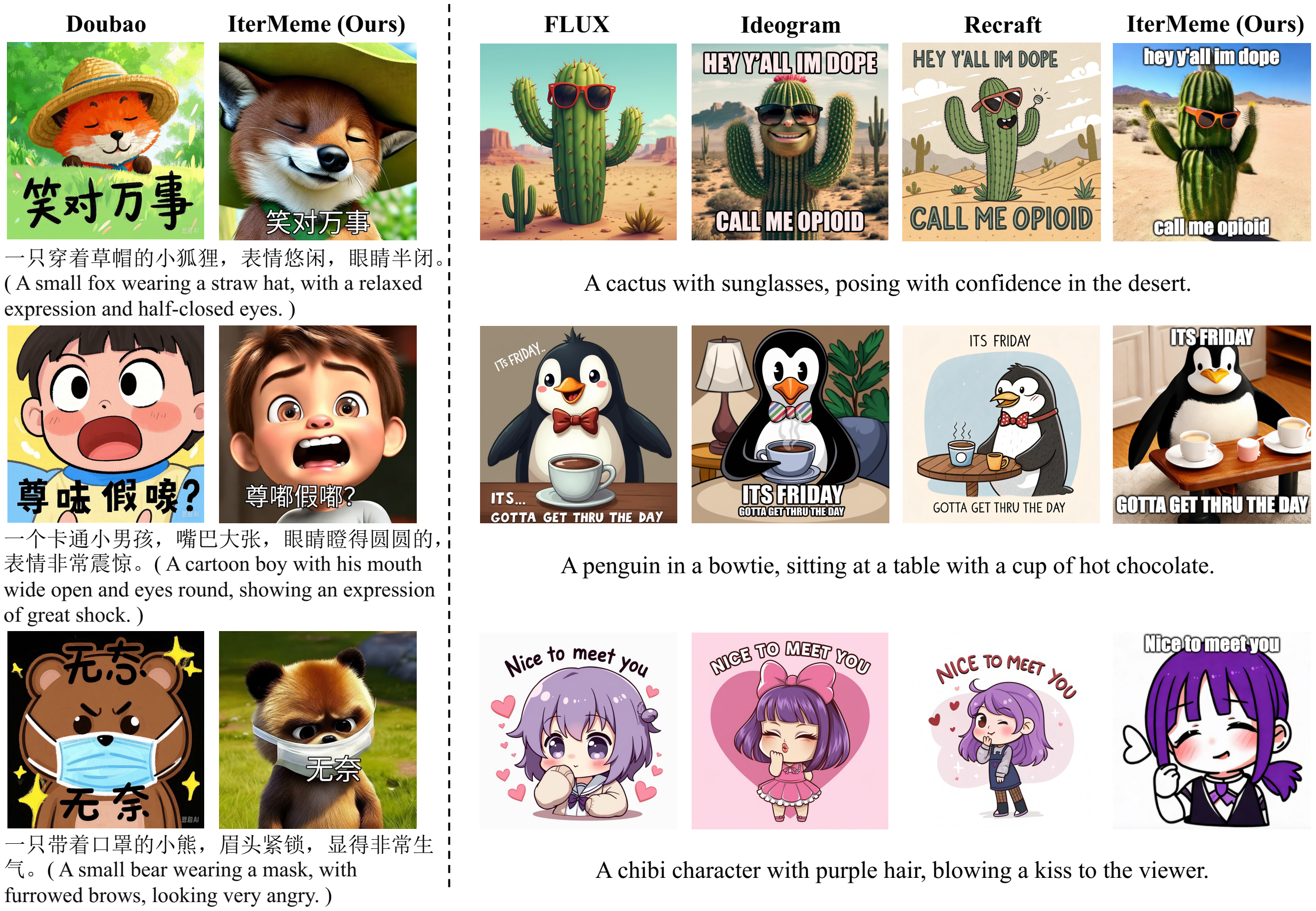
the second training stage, the newly introduced M^{3}OE module is initialized to 0 and trained using the same hyperparameter settings as the first stage.

Evaluation datasets and metrics. Our evaluation dataset comprises 60 independently collected IPs, spanning categories such as real people, cartoons, and anime. These IPs are excluded from the training set and are used as reference images. For each IP, we provide a corresponding prompt that encompasses various attributes and expressions to ensure a diverse range of test scenarios. To thoroughly assess the performance of our framework in meme creation, we follow the setup of XMeCap [Chen *et al.*, 2024], evaluating the performance of the generated captions from four key perspectives: Informativeness, Relevance, Creativity, and Humorous. These assessments are conducted using GPT-4o [Achiam *et al.*, 2023], with detailed evaluation instructions provided in the appendix. For layout generation capabilities, we introduce metrics of Effectiveness and Readability, which are further described in the appendix. Additionally, we use the CLIP-T [Radford *et al.*, 2021] metric to quantify text-image alignment in personalized generation and the DINO [Oquab *et al.*, 2023] and CLIP-I metrics to evaluate IP preservation. A user study is included to assess the overall quality of the generated memes. We invite users from diverse backgrounds to select their preferred memes based on text alignment, image quality, and overall appeal.

Baselines. We select different baseline models tailored to each specific task. To evaluate caption generation capabilities, we select DreamLLM[Dong *et al.*, 2023], CLoT[Zhong *et al.*, 2024], LLaVA-v1.6 [Liu *et al.*, 2024], and Qwen2-VL[Wang *et al.*, 2024a] as baseline models. Due to the lack of existing methods for layout generation, we adopt a random layout generation approach, where three constrained random numbers are generated for each caption to represent the center coordinates and font size. Additionally, we utilize the zero-shot [Kojima *et al.*, 2022] and few-shot [Brown *et al.*, 2020] capabilities of MLLM for layout generation. For personalized generation, we select DreamLLM [Dong *et al.*, 2023], DreamBooth [Ruiz *et al.*, 2023], Textual Inversion [Gal *et al.*, 2022], and IP-Adapter [Ye *et al.*, 2023] as baselines. Furthermore, we include state-of-the-art methods such as Doubao [ByteDance, 2024], FLUX [Labs, 2024], Ideogram [Ideogram, 2024], and Recraft [Recraft, 2024] for overall quality comparisons.

4.2 Comparison to Baselines

Quantitative comparison. We present our quantitative comparison results in table 2. The results for caption generation in table 1b show IterMeme surpasses DreamLLM by 22.8% in Creativity and 32.4% in Humorous, demonstrating the effectiveness of the proposed MemeCap dataset. While CLoT [Zhong *et al.*, 2024] excels in Creativity by encouraging remote association thinking in LLMs, its lower scores in



一只穿着草帽的小狐狸，表情悠闲，眼睛半闭。(A small fox wearing a straw hat, with a relaxed expression and half-closed eyes.)

一个卡通小男孩，嘴巴大张，眼睛瞪得圆圆的，表情非常震惊。(A cartoon boy with his mouth wide open and eyes round, showing an expression of great shock.)

一只带着口罩的小熊，眉头紧锁，显得非常生气。(A small bear wearing a mask, with furrowed brows, looking very angry.)

A cactus with sunglasses, posing with confidence in the desert.

A penguin in a bowtie, sitting at a table with a cup of hot chocolate.

A chibi character with purple hair, blowing a kiss to the viewer.

Figure 4: **Comparison of user study results for Meme Creation.** We compare Chinese and English memes, and our observations indicate that our method achieves competitive performance.

	T2I	Personalized Generation		
Model	CLIP-T	CLIP-T	CLIP-I	DINO
Exp1	31.97	29.379	85.37	63.77
Exp2	31.908	29.485	86.082	66.37
Exp3	**32.611**	**29.517**	**86.14**	**67.55**

Table 1: **Ablation study.** The effectiveness of the M^{3}OE Module.

	CLIP-T	CLIP-I	DINO
DreamLLM [Dong *et al.*, 2023]	27.423	75.762	58.09
Dreambooth [Ruiz *et al.*, 2023]	22.457	83.319	66.65
Textual Inversion [Gal *et al.*, 2022]	22.457	81.229	65.47
IP-Adapter [Ye *et al.*, 2023]	29.354	85.216	67.52
Ours	**29.517**	**86.14**	**67.55**

(a) Comparison for personalized generation.

	Informativeness	Relevance	Creativity	Humorous
DreamLLM [Dong *et al.*, 2023]	66.4	81	42.6	51
CLot [Zhong *et al.*, 2024]	75.4	74.6	**70.6**	74
LLaVA-v1.6 [Liu *et al.*, 2024]	65	**95**	60	80.6
Qwen2-VL [Wang *et al.*, 2024a]	70.4	87.6	63	66.6
Ours	**77.6**	81.4	65.4	**83.4**

(b) Comparison for caption generation.

	Effective -ness	Read -ability
Random	21.67	13.3
Zero-shot	20	8.3
Few-shot	91.7	78.3
Ours	**93.3**	**86.7**

(c) Comparison for layout generation.

Table 2: **Quantitative comparisons** on various metrics. The best results are marked in **bold**.

Humorous and Relevance undermine its viability for meme creation. In contrast, IterMeme delivers captions with superior Humorous and Informativeness while balancing Relevance and Creativity, showcasing its exceptional ability to enhance memes' emotional expressiveness. For layout generation, the results in table 1c indicate that IterMeme achieves

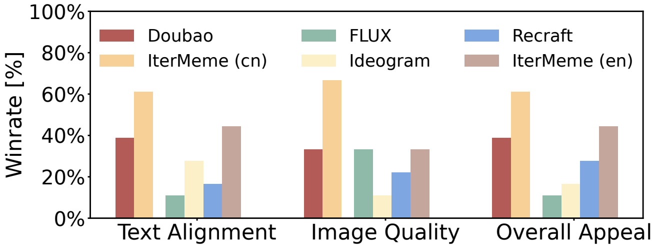


Figure 5: Results of user study.

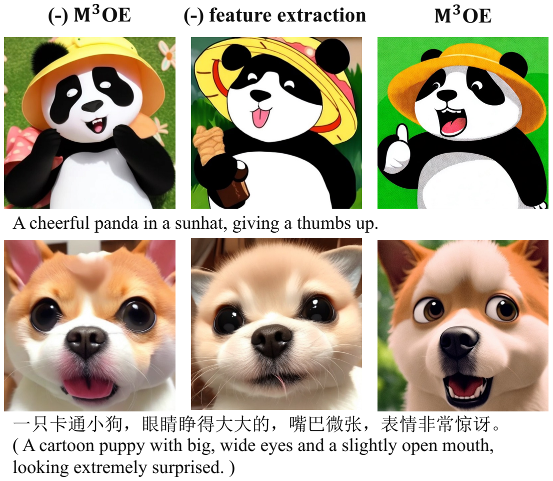
the highest scores in Effectiveness and Readability, demonstrating superior control over caption layouts. Notably, all baselines use the same IterMeme-generated captions, with random layouts constrained to have midpoint coordinates within image boundaries and font sizes set to $[H/20, H/8]$, where H represents the image height. Similarly, the personalized generation results in table 1a show IterMeme outperforming all baselines, with a 10.378% CLIP-I and 2.094% CLIP-T improvement over DreamLLM. This highlights the effectiveness of the proposed M^{3}OE framework in guaranteeing IP-consistent images and enhanced text-image coherence.

Qualitative comparison. In fig. 3, we present a qualitative comparison of personalized generation results between IterMeme and the baseline models. The proposed model significantly outperforms the baselines in terms of image generation quality, exhibits superior IP consistency, and achieves stronger alignment with textual prompts. Additional qualitative results are provided in the appendix for further reference.

User study. Our user study involves 24 users, each evaluating 10 randomly selected cases (5 in Chinese and 5 in English). Due to language limitations, Doubao is evaluated solely on Chinese data. Since the baseline models cannot automatically generate meme captions, we manually insert text matching our model's outputs into their prompts to ensure a fair comparison. In total, we collect 720 votes, with the results shown in fig. 5. The results show that IterMeme significantly outperforms Doubao in all three dimensions for Chinese. For English, while IterMeme matches FLUX in image quality, its superior text alignment leads to higher overall appeal. Additionally, fig. 4 provides a series of comparative results for further evaluation. From these comparisons, it is observed that Doubao generates flawed text, FLUX lacks text generation capability, and Ideogram and Recraft produce sticker-like outputs with poor caption-visual alignment.

4.3 Ablation Study

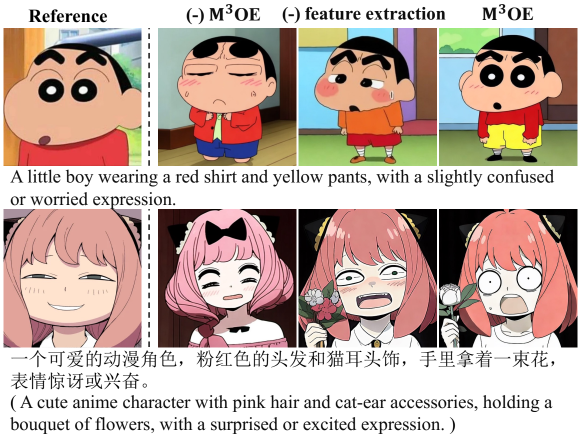
We conduct an ablation study on the proposed M^{3}OE module through two comparative experiments. The first experiment, denoted as Exp 1, evaluates the necessity of the M^{3}OE module by removing it from the model, and to enable effective training, we also optimize the parameters of the learnable queries. The second experiment, denoted as Exp 2, assesses the effectiveness of the feature extraction module within M^{3}OE by omitting it during training on T2I data. Our method is denoted as Exp 3. The experimental results, presented in table 1, demonstrate a significant performance drop across all metrics in both experiments. Additionally, the qualitative results in fig. 7 confirm that incorporating M^{3}OE equips the MLLM with personalized generation capabilities. Applying the feature extraction module of M^{3}OE to T2I data further yields higher-quality outputs as shown in fig. 6, along with improved text alignment and enhanced IP consistency.



A cheerful panda in a sunhat, giving a thumbs up.

一只卡通小狗，眼睛睁得大大的，嘴巴微张，表情非常惊讶。
(A cartoon puppy with big, wide eyes and a slightly open mouth, looking extremely surprised.)

Figure 6: Ablation study on text-to-image generation confirms the effectiveness of the M^{3}OE module.



A little boy wearing a red shirt and yellow pants, with a slightly confused or worried expression.

一个可爱的动漫角色，粉红色的头发和猫耳头饰，手里拿着一束花，表情惊讶或兴奋。
(A cute anime character with pink hair and cat-ear accessories, holding a bouquet of flowers, with a surprised or excited expression.)

Figure 7: Ablation study on personalized generation confirms the effectiveness of the M^{3}OE module.

5 Conclusion

In this paper, we propose IterMeme, a framework leveraging a unified MLLM to address incomplete component integration in meme creation. The proposed MemeCap dataset facilitates precise caption-layout control, while the expert-guided M^{3}OE module enables fine-grained IP extraction and semantic alignment. A two-stage training strategy further achieves exceptional caption-layout generation and interactive personalization, filling the gaps in existing approaches.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (62425114, 62121002, U23B2028, 62232006). We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. We also thank the USTC supercomputing center for providing computational resources for this project.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Amin and Burghardt, 2020] Miriam Amin and Manuel Burghardt. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, 2020.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[ByteDance, 2024] ByteDance. Doubao. https://www.doubao.com, 2024.

[Chen *et al.*, 2024] Yuyan Chen, Songzhou Yan, Zhihong Zhu, Zhixu Li, and Yanghua Xiao. Xmecap: Meme caption generation with sub-image adaptability. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3352–3361, 2024.

[Dong *et al.*, 2023] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

[Gal *et al.*, 2022] 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.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Hsu *et al.*, 2023] Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6018–6026, 2023.

[Huang *et al.*, 2024] Jiehui Huang, Xiao Dong, Wenhui Song, Hanhui Li, Jun Zhou, Yuhao Cheng, Shutao Liao, Long Chen, Yiqiang Yan, Shengcai Liao, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024.

[Ideogram, 2024] Ideogram. Ideogram. https://ideogram.ai, 2024.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

[Labs, 2024] Black Forest Labs. Announcing black forest labs, 2024.

[Liu *et al.*, 2024] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[Loshchilov, 2017] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2024] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

[Recraft, 2024] Recraft. Recraft. https://www.recraft.ai/, 2024.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Ruiz *et al.*, 2023] 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.

[Team, 2024] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[Wang *et al.*, 2024a] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wang *et al.*, 2024b] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[Xie *et al.*, 2024] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[Yang *et al.*, 2024] Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Posterllava: Constructing a unified multi-modal layout generator with llm. *arXiv preprint arXiv:2406.02884*, 2024.

[Ye *et al.*, 2023] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[Yu *et al.*, 2016] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 516–520, 2016.

[Zhang *et al.*, 2024a] Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Dynamic equilibrium between realism and compositionality improves text-to-image diffusion models. *arXiv preprint arXiv:2402.12908*, 2024.

[Zhang *et al.*, 2024b] Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. *arXiv preprint arXiv:2410.07171*, 2024.

[Zhong *et al.*, 2024] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257, 2024.

[Zhou *et al.*, 2024a] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[Zhou *et al.*, 2024b] Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. Customization assistant for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9182–9191, 2024.