Personalized Generation In Large Model Era: A Survey

Anonymous ACL submission

Abstract

In the era of large models, content generation is gradually shifting to Personalized Generation (PGen), tailoring content to individual preferences and needs. This paper presents the first comprehensive survey on PGen, investigating existing research in this rapidly growing field. We conceptualize PGen from a unified perspective, systematically formalizing its key components, core objectives, and abstract workflows. Based on this unified perspective, we propose a multi-level taxonomy, offering an in-depth review of technical advancements, commonly used datasets, and evaluation metrics across multiple modalities, personalized contexts, and tasks. Moreover, we envision the potential applications of PGen and highlight open challenges and promising directions for future exploration. By bridging PGen research across multiple modalities, this survey serves as a valuable resource for fostering knowledge sharing and interdisciplinary collaboration, ultimately contributing to a more personalized digital landscape.

1 Introduction

Recent advancements in large generative models have catalyzed a paradigm shift in content generation, moving from generic, one-size-fits-all generation to Personalized Generation (PGen) (Wang et al., 2023c; Xu et al., 2024c; Nguyen et al., 2024b). By crafting personalized content that resonates more deeply with individual preferences, PGen holds great potential to enhance user-centric services and foster more engaging, immersive user experiences across various domains, such as customized product images in e-commerce (Yang et al., 2024a), personalized advertisements in marketing campaigns (Tang et al., 2024a), and personalized AI assistants (Zhang et al., 2024a). Given its significant potential, PGen has attracted significant attention from both academia and industry.

Despite significant progress (Alaluf et al., 2025; Salemi et al., 2024b), research efforts in PGen have largely evolved independently within different communities, such as Natural Language Processing (NLP), Computer Vision (CV), and Information Retrieval (IR). There is no survey that specifically provides a cross-community overview of PGen research. Existing surveys related to PGen separately follow either a model-centric or task-centric perspective, offering only partial summaries of relevant studies. 1) Model-centric surveys focus on specific generative models for personalization, such as Multimodal Large Language Models (MLLMs) (Wu et al., 2024b), Large Language Models (LLMs) (Zhang et al., 2024j; Chen et al., 2024e; Li et al., 2024i), and Diffusion Models (DMs) (Zhang et al., 2024g); 2) Task-centric surveys summarize personalization techniques in specific applications, such as dialogue generation (Chen et al., 2024f), role-playing (Chen et al., 2024d; Tseng et al., 2024), and generative recommendation (Ayemowa et al., 2024). None of these offers a unified framework that comprehensively summarizes PGen research across communities.

A unified framework is crucial for systematically reviewing recent advances and emerging trends in PGen, providing a comprehensive, panoramic view of this field. Moreover, it can foster communication, knowledge sharing, and collaboration between various research communities, ultimately driving the development of a more advanced and personalized digital ecosystem. However, conducting such a unified survey is challenging, as different communities prioritize distinct modalities. For instance, the NLP and IR communities primarily focus on the text modality, while CV specializes in image, video, and 3D. Since each modality presents distinct data structures and challenges, these modality-specific differences introduce inherent technical divergences, making it difficult to unify PGen research into a cohesive framework.

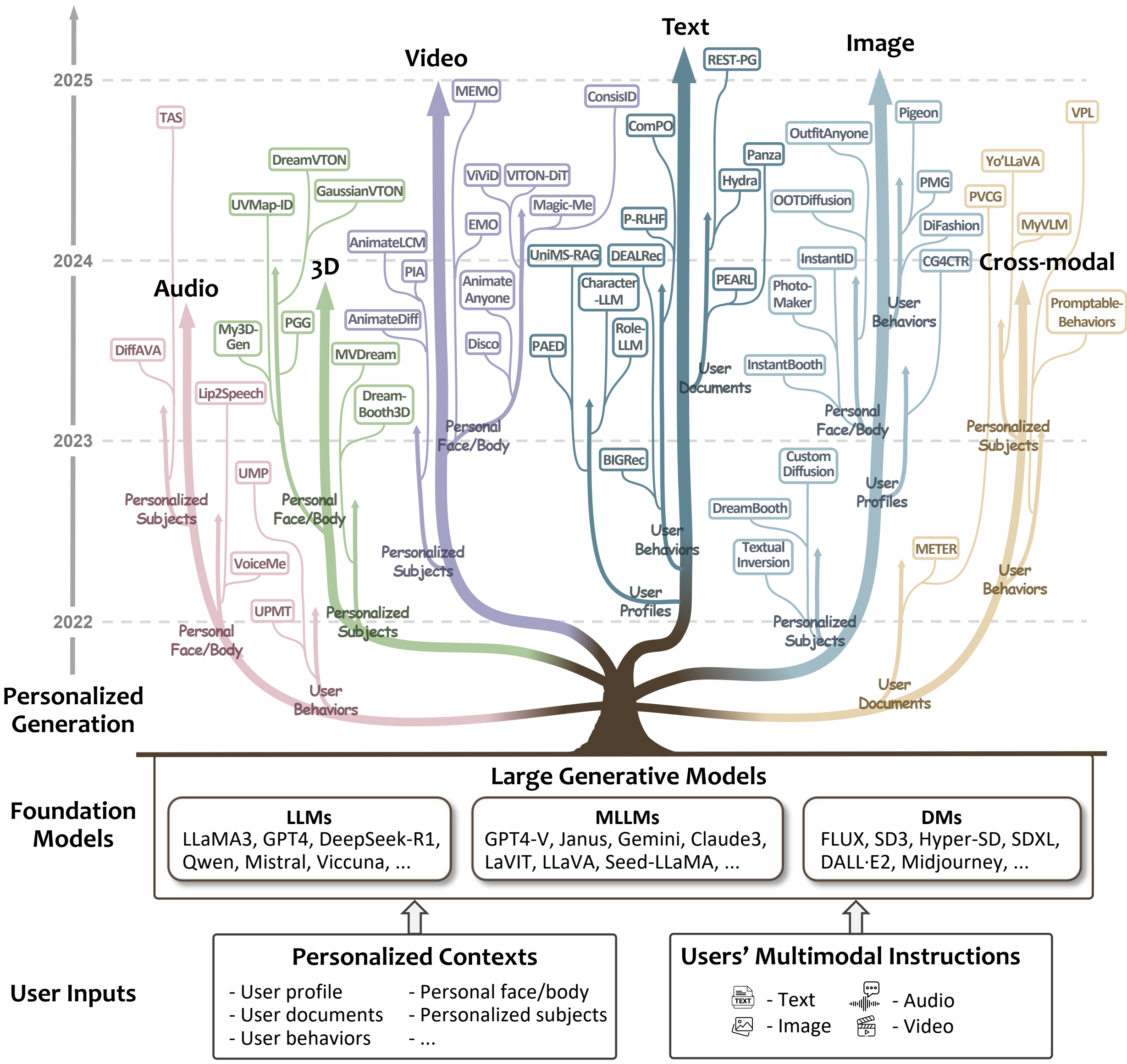

Figure 1: Overview of personalized generation across modalities, inspired by the figure in Yang et al. (2024b).

To address these challenges, it is essential to re-examine PGen from a high-level, modality-agnostic perspective. Fundamentally, PGen entails user modeling based on various personalized contexts and multimodal instructions, extracting personalized signals to guide the content generation process. As illustrated in Figure 1, existing PGen research essentially models various user inputs and leverages generative models for personalized content generation in multiple modalities.

To this end, we present the first survey dedicated to PGen. The structure of this survey and our key contributions are summarized as follows:

- **A unified user-centric perspective for PGen (Section 2).** We conceptualize PGen by formalizing the key components, core objectives, and general workflow, integrating studies across different modalities into a holistic framework.
- **A multi-level taxonomy for existing work in PGen (Section 3).** Building on the unified perspective, we introduce a novel taxonomy that systematically reviews PGen's technical advancements, commonly used datasets, and evaluation metrics across various modalities, personalized contexts, and tasks.
- **An outlook for potential applications of PGen in enhancing user-centric services (Section 4).** We categorize potential applications of PGen by content personalization stages, with a focus on the content creation and delivery processes.
- **An overview of key open problems in PGen for future research (Section 5).** We outline the critical open problems that need to be addressed to drive innovation in future research and advance the user-centric content ecosystem.

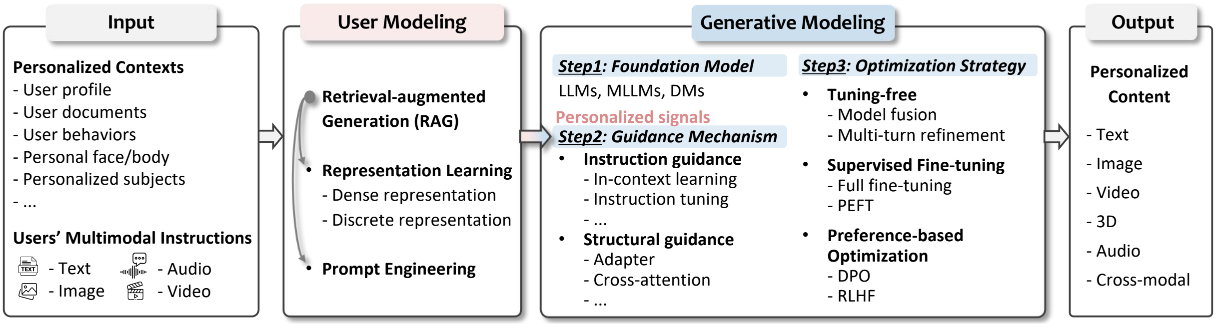

Figure 2: Personalized generation workflow.

2 A Unified User-centric Perspective for Personalized Generation

2.1 Task Formulation

PGen leverages generative models to synthesize content tailored to individual preferences and specific needs. As illustrated in Figure 1, it relies on two essential user inputs: 1) ***Personalized contexts*** that encapsulate user preferences; 2) ***Users' multimodal instructions***,which include textual prompts, voice commands, and other modality-specific inputs that explicitly convey their content needs. Generative models learn user preferences and personal characteristics from diverse personalized contexts and follow users' multimodal instructions to generate customized content across different modalities. The personalized contexts encompass the following dimensions:

- *User profiles*: A collection of demographic and personal attributes associated with a specific user, such as age, gender, occupation, and location.
- *User documents*: User-created textual content, such as comments, emails, and social media posts, that reflects personal creative preferences.
- *User behaviors*: User interactions captured during user engagement, such as searches, clicks, likes, comments, views, shares, and purchases.
- *Personal face/body*: Individual facial and bodily traits, including both static features (*e.g.,* facial structure and body shape) and dynamic features (*e.g.,* expressions, gestures, and motions). These are widely used in tasks like portrait generation, fashion virtual try-ons, and 3D modeling.
- *Personalized subjects*: User-specific concepts or entities, such as pets, personal items, and favorite objects that reflect unique tastes.

By integrating personalized contexts with users' multimodal instructions, generative models can create highly tailored content, aligning closely with individual preferences and addressing specific needs.

2.2 Objectives

Although PGen in each modality is shaped by unique data structures, specific challenges, and distinct tasks, three essential objectives and evaluation dimensions remain consistent across modalities:

- ***High quality***: Ensuring the generated content meets high standards of coherence, relevance, and aesthetics.
- ***Instruction alignment***: Requiring the generated content to accurately adhere to users' multimodal instructions and effectively address their needs.
- ***Personalization***: Guaranteeing that the generated content aligns with personalized contexts and caters to specific user preferences.

While text generation has consistently achieved high-quality outputs, challenges persist in other modalities, such as image, video, audio, and 3D generation. In these domains, generated content can sometimes appear chaotic or disjointed. Maintaining high-quality standards across all modalities is fundamental to achieving successful personalized generation. Furthermore, in certain domains where *factual accuracy* is particularly important, such as news, laws, policies, and expert knowledge, generative models must prioritize authenticity to ensure the reliability and trustworthiness of the content provided to users.

2.3 Workflow

As shown in Figure 2, the PGen workflow involves two key processes: 1) user modeling based on diverse user-specific data and 2) generative modeling across multiple modalities, ensuring high-quality, instruction-aligned, and personalized content.

User Modeling To effectively capture user preferences and specific content needs based on personalized contexts and users' multimodal instructions, three key techniques are commonly employed: 1) Representation learning, which encodes these inputs into dense embeddings (Ruiz et al., 2023; Tang et al., 2024b) or summarizes them into discrete representations (*e.g.,* texts) (Shen et al., 2024b); 2) Prompt engineering, which involves designing task-specific prompts to structure user-specific information for generative models (Chen et al., 2024g; Li et al., 2025); and 3) Retrieval-augmented generation (RAG), which enriches user-specific information by filtering out irrelevant information and integrating external relevant data (Salemi and Zamani, 2024; Mysore et al., 2024). By combining these techniques, user modeling establishes a robust foundation for PGen, extracting personalized signals to guide content personalization within the generative modeling process.

Generative Modeling To generate personalized content effectively, generative modeling follows a structured three-step process:

- ***Step1: Foundation model.*** In the era of large models, LLMs, MLLMs, and DMs serve as the backbone for content generation. Selecting an appropriate foundation model based on the target modality, task requirements, and user-specific data is crucial for achieving accurate and personalized content.
- ***Step2: Guidance mechanism.*** To effectively integrate personalized signals, two primary guidance mechanisms are employed: instruction guidance and structural guidance. Specifically, instruction guidance ensures models follow explicit user prompts and instructions using techniques such as in-context learning (Xu et al., 2023b; Chen et al., 2024g) and instruction tuning (Pi et al., 2024; Xu et al., 2024c). In contrast, structural guidance modifies the model architecture by incorporating additional modules, such as adapters (Ye et al., 2023) and cross-attention mechanisms (Wei et al., 2023), to embed personalized information.
- ***Step3: Optimization Strategy.*** Empowering large generative models with the capability of personalized generation involves three primary optimization strategies: 1) Tuning-free methods, which utilize pre-trained models for personalized generation without modifying model parameters. These methods often rely on model fusion to assemble multiple pre-trained models (Ding et al., 2024) or employ interactive generation processes that collect real-time user feedback for multi-turn refinement (Von Rütte et al., 2023), ensuring alignment with individual preferences. 2) Supervised fine-tuning which optimizes model parameters using explicit supervision signals, either through full fine-tuning (Xu et al., 2024b; Ruiz et al., 2023) or Parameter-Efficient Fine-Tuning (PEFT) techniques (Wu et al., 2024f; Tan et al., 2024; Zhang et al., 2024b). 3) Preference-based optimization, which incorporates user preference data to update model parameters. A key approach is Reinforcement Learning with Human Feedback (RLHF) (Li et al., 2024g; Zhang, 2024), which employs an explicit reward model to guide optimization. Alternatively, Direct Preference Optimization (DPO) offers a streamlined solution by directly aligning model parameters with pairwise user preferences (Zhang et al., 2024c; Huang et al., 2024b).

By integrating these advanced techniques and strategies, this workflow not only ensures adaptability to diverse personalized contexts and user instructions but also highlights the evolving landscape of large generative models, offering a scalable solution for PGen.

3 Personalized Generation Across Modalities

Based on the unified perspective, we present a multi-level taxonomy for PGen, systematically reviewing PGen research across various modalities, personalized contexts, and specific tasks, as outlined in Table 1. Studies on PGen are first categorized by modality, including text, image, video, audio, 3D, and cross-modal generation. Within each modality, we further classify research based on personalized contexts and examine corresponding tasks and techniques. A detailed review of image, video, and 3D modalities is provided in Appendix A. Additionally, we provide a comprehensive overview of commonly used evaluation metrics and datasets for each modality in Appendix B and summarize them in Table 2, Table 3, and Table 4.

3.1 Personalized Text Generation

Personalized text generation aims to provide textual content tailored to user preferences and needs, involving tasks ranging from information seeking to user simulation.

3.1.1 User Behaviors

User interactions with the system typically occur over time (Kelly and Belkin, 2002), allowing it to learn implicit preferences and behavioral patterns to enhance personalization and encourage long-term engagement (Shi et al., 2013). This personalized context is valuable for the following personalized text-based tasks.

Information Seeking A primary use case of personalized text generation is crafting responses to align with user preferences, enabling more engaging interactions. The system can leverage user feedback (e.g., thumbs up/down and selected best responses) to tailor its responses to user preferences. While personalization has been extensively studied in the context of information access and search (Kasela et al., 2024; Zeng et al., 2023; Eugene et al., 2013; Guo et al., 2021), which aims to select a response from predefined options, it remains relatively underexplored in generative scenarios. This is largely due to the lack of standardized metrics and benchmarks. However, recently, Li et al. (2024g) explored the use of personalized feedback to train LLMs to generate more tailored summaries for users, as a form of information seeking. Kumar et al. (2024b) extends this approach by collecting preference feedback from a group of users as a community to optimize the model's ability to respond to their collective information needs.

Recommendation While recommendation is not directly involved in content generation, it plays a crucial role in delivering personalized content, which has been explored across various scenarios (Hou et al., 2024; Harper and Konstan, 2015; Wan and McAuley, 2018; Wu et al., 2020a). The use of generative models in Recommendation Systems (RecSys) has been widely studied (Bao et al., 2023; Wu et al., 2024c; Yang et al., 2023a). Specifically, LLMs have been utilized either through prompting (Lyu et al., 2024) or by training them directly to perform recommendation tasks as a form of text generation (Lin et al., 2024a). Beyond transformer-based generative models, newer approaches like diffusion models have also been explored for recommendation, highlighting the versatility of generative methods in this domain (Wang et al., 2023e; Yang et al., 2024e).

Other work has also leveraged realistic user interaction to explore personalization for review generation (Ni et al., 2019; Li et al., 2020; Sun et al., 2020; Li et al., 2021) and news headline generation (Ao et al., 2021; Cai et al., 2023; Song et al., 2023). For example, Ao et al. (2021) presents a personalized headline generation benchmark by collecting user's click history on Microsoft News.

3.1.2 User Documents

In some cases, users may not interact with the system frequently but can provide valuable information for personalization, such as user-created documents (Salemi et al., 2024b).

Writing Assistant Personalization plays a critical role in enhancing text-based writing assistants, enabling tailored text generation across diverse formats and styles. For this purpose, the LaMP benchmark (Salemi et al., 2024b,a; Salemi and Zamani, 2024; Zhuang et al., 2024) focuses on short-form text generation, such as creating headlines for news articles or subject lines for emails. In contrast, the LongLaMP benchmark (Kumar et al., 2024a) targets longer-form tasks, such as writing product reviews from a user's perspective based on a rating, generating a post from its summary, or completing an email for a user (Salemi et al., 2025b). Additionally, the Personalized Linguistic Attributes Benchmark (Alhafni et al., 2024) is designed for text completion tasks, such as extending a post or review, leveraging user-written documents to extract stylistic features that guide LLMs in mimicking a user's unique writing style. In this domain, RAG is the dominant approach, retrieving relevant information from users' historically written documents to extract their writing preferences (Mysore et al., 2024; Nicolicioiu et al., 2024; Li et al., 2023a).

3.1.3 User Profiles

Generative models can infer user preferences from their profiles to guide personalized text generation.

Dialogue System In recent years, chatbots and conversational systems have been a central focus of personalized text generation (Kottur et al., 2017; Thompson et al., 2004; Kaiss et al., 2023; Kocaballi et al., 2019). The advent of advanced chat models like GPT-4 and Gemini have enabled more sophisticated and personalized interactions. By defining a persona or personality for these models based on user profiles, the system can tailor responses to individual preferences and needs (Pradhan and Lazar, 2021; Song et al., 2019). To support research in this domain, various dialogue datasets have been developed. For example, LiveChat (Gao et al., 2023) is a large-scale dataset created from live streaming interactions, FoCus (Jang et al., 2021) focuses on conversational information-seeking scenarios, and Pchatbot (Qian et al., 2021) compiles data from Weibo and judicial forums. Enhancing LLMs' ability to generate personalized responses involves approaches such as zero-shot prompting (Zhu et al., 2023b), in-context learning (Xu et al., 2023c), and training on limited persona datasets (Song et al., 2021; Wang et al., 2024d) or large-scale datasets (Zheng et al., 2020; Chen et al., 2023b). Additionally, chain-of-thought reasoning has proven effective in improving alignment with user preferences in dialogue systems (Li et al., 2025).

User Simulation Previous research demonstrates that LLMs excel at performing roles or personas assigned to them (Shanahan et al., 2023; Chen et al., 2024c), enabling user simulation based on their profiles to extract preferences and further personalize the system for them (Magee et al., 2024; Santurkar et al., 2023). In this domain, Shao et al. (2023) develops a test playground to interview trained agents and assess whether they effectively memorize their assigned characters and experiences. Additionally, Wang et al. (2024e) introduced a large-scale dataset for evaluating the role-playing abilities of LLMs, while Ng et al. (2024) presented a dataset specifically designed for assessing these capabilities within video game contexts.

3.2 Personalized Audio Generation

Personalized audio generation extracts users' auditory preferences to create tailored audio content, such as music and speech.

3.2.1 User behaviors

User behaviors on music, such as listening history and ratings, are important clues for inferring user preferences for personalization.

Music Generation Methods like UMP (Ma et al., 2022) and UP-Transformer (Hu et al., 2022) infer user preferences by analyzing listening histories and ratings. UIGAN (Wang et al., 2024j) adopts an interactive approach to collect user feedback, progressively refining the generated music. Ma et al. (2024b) construct a music community graph, where nodes represent users and songs, and edges capture relationships such as likes, subscriptions, and user interactions.

3.2.2 Personalized Subjects

In some cases, users provide audio clips and aim to manipulate them using textual prompts.

Text-to-Audio Generation Personalized text-to-audio generation explores methods for synthesizing tailored audio by aligning user preferences, textual descriptions, and contextual inputs. Mo et al. (2023) utilize Contrastive Language-Audio Pretraining (CLAP) to align features across audio and visual-text modalities in the spatiotemporal domain. Furthermore, Plitsis et al. (2024) adapt image-domain personalization techniques like Textual Inversion (Gal et al., 2023) and DreamBooth (Ruiz et al., 2023) to audio tasks, enhancing user-centric generation. Li et al. (2024j) introduce a Semantic-Aware Fusion (SAF) module to capture text-aware audio features, establishing nuanced perceptual relationships between text and audio inputs for more contextually aligned audio outputs.

3.2.3 Personal Face/Body

Users may provide their facial images or videos, enabling generative models to extract speaker-specific attributes for customized speech generation.

Face-to-Speech Generation VioceMe (van Rijn et al., 2022) employs SpeakerNet to derive speaker embeddings and incorporates Global Style Tokens (GST) for modeling speech styles. FR-PSS (Wang et al., 2022) enhances the extraction of speech features from facial characteristics using a residual guidance strategy, which integrates prior speech information for improved efficiency. Lip2Speech (Sheng et al., 2023) leverages a Variational Autoencoder (VAE) framework to disentangle speaker identity from linguistic content in videos, achieving fine-grained control over personalized speech synthesis via speaker embeddings.

3.3 Personalized Cross-modal Generation

Personalized cross-modal generation primarily aims to produce personalized textual responses (e.g., captions, answers, or robot actions) based on multimodal personalized contexts (e.g., images, videos, historical robot trajectories).

3.3.1 User behaviors

Based on user interactions, generative models can infer user preferences to tailor robotic behaviors.

Robotics Several studies have investigated inferring user preferences from historical trajectories and associated human feedback, enabling personalized robotic decision-making. Specifically, Poddar et al. (2024) proposed Variational Preference Learning, integrating variational inference into RLHF to model diverse user preferences and enable active learning. Hwang et al. (2024) introduced "Promptable Behaviors", using Multi-Objective Reinforcement Learning to dynamically customize policies

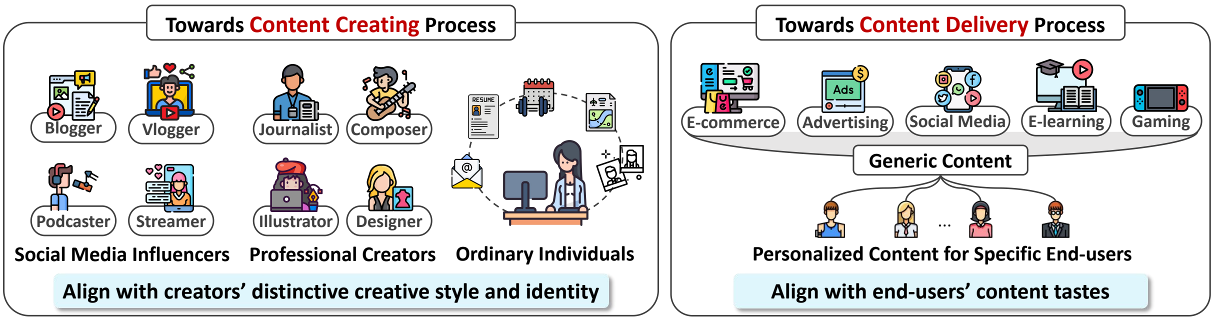

Figure 3: Applications of personalized generation.

via human demonstrations, trajectory feedback, and language instructions. Cui et al. (2024b) presented a framework with a RAG memory module to match individual users' preferences and driving styles.

3.3.2 User Documents

User-created documents, such as comments, reviews, and captions can be used to infer their writing style and preferences for personalization.

Caption/Comment Generation Several studies utilize user-created captions and comments to develop personalized captioning and comment systems (Long et al., 2020; Zhang et al., 2020c; Geng et al., 2022). To effectively model user preferences, Shin et al. (2018) engages users by soliciting answers to targeted questions during generation. Lin et al. (2024b); Wu et al. (2024e) utilize users' historical comments for personalized video comment generation. In addition, PV-LLM (Lin et al., 2024b) fine-tunes MLLMs using user-written comments, while PVCG (Wu et al., 2024e) learns a unique identifier for each user to enhance personalization.

3.3.3 Personalized subjects

In some cases, users may provide specific subject images, such as photos of their friends, for personalized visual question answering (e.g., *"Give me a birthday gift list for my friend Peter."*).

Cross-modal Dialogue Systems Given user-specific subject images and queries, the systems are expected to identify these subjects and infer user intents for personalized responses. Existing studies in this domain mainly fall into two groups:

- Memory-based methods store user-specific subjects and activate or retrieve them as needed. For example, CSMN (Chunseong Park et al., 2017; Park et al., 2018) stores image memory, user context memory, and word output memory, and the model generates personalized captions based on memory features. Hao et al. (2024) leverages RAG techniques to personalize LMMs, allowing LLMs to update their supported concepts without requiring additional training. PLVM (Pham et al., 2024) proposes a pre-trained Aligner to align referential concepts with the queried images.
- Optimization-based methods inject personalized information into the generation process via specific modules. For example, PerVL (Cohen et al., 2022), Yo'LLaVA (Nguyen et al., 2024b), and MC-LLaVA (An et al., 2024) learn to encode a personalized subject into a set of latent tokens based on several provided subject images. MyVLM (Alaluf et al., 2025) introduces learnable heads, each dedicated to recognizing a single user-specific subject. In addition, several studies have explored learning unique user embeddings Long et al. (2020); Zhang et al. (2020c); Xiong et al. (2020) or performing prefix tuning (Wang et al., 2023f) for personalization.

4 Applications

The prior section has highlighted the success of PGen across various modalities, demonstrating its potential to enhance user engagement and enrich experiences across diverse domains. As illustrated in Figure 3, applications of PGen can be categorized based on the stages of content personalization: 1) towards content creation process, which provides personalized tools and services for content creators of all levels to maintain their unique creative style while streamlining the creative workflows; and 2) towards content delivery process, which delivers multimodal content in a personalized manner tailored to individual preferences of end users. A more detailed discussion of PGen applications can be found in Appendix C.

5 Open Problems

Despite the significant progress made by PGen, several key challenges remain to be addressed.

5.1 Technical Challenges

- ***Scalability and Efficiency***. PGen relies on large generative models for content personalization, which often require extensive resource costs, limiting their deployment to real-time, large-scale user scenarios. Developing scalable and efficient algorithms for PGen remains a critical direction for future research (Yang et al., 2024c).
- ***Deliberative Reasoning for PGen***. In certain personalized scenarios that prioritize content quality over temporal efficiency – such as digital advertising, where advertisers typically serve only several ads per user each day – inference scaling presents significant opportunities for enhancing user satisfaction. Prior work mainly focuses on multi-turn refinement to progressively enhance content relevance and personalization (Nabati et al., 2024). Inspired by the great success of LLM reasoning (Guan et al., 2024; Guo et al., 2025), deliberative reasoning may emphasize extensive logical and contextual reasoning beforehand, enabling a thorough analysis of user preferences to drive more effective content personalization (Fang et al., 2025).
- ***Evolving User Preference***. As explored in traditional RecSys (Wang et al., 2023d), recognizing dynamic preferences from user behaviors is crucial to enhancing personalization. Adapting PGen to track and respond to user preference shifts remains a key research direction.
- ***Multi-modal Personalization***. Existing PGen research mainly focuses on single-modality generation, while multi-modal personalization remains underexplored, such as personalized social media posts that integrate both image and text. This challenge requires high-quality, instruction-aligned, and personalized output while ensuring consistency across multiple modalities.
- ***Synergy Between Generation and Retrieval***. Traditional personalized content delivery systems primarily focus on retrieval-based methods like RecSys. However, existing content may not fully meet users' content needs. Integrating PGen with retrieval-based approaches holds great promise for building more powerful personalized content delivery systems (Wang et al., 2023c).

5.2 Benchmarks and Metrics

A fundamental challenge in PGen is establishing robust metrics and benchmark datasets. Current evaluation methods primarily rely on traditional generation metrics (e.g., BLEU for text, CLIP-I for images), which do not fully capture how well the generated content aligns with user preferences. Future research should focus on developing more effective metrics for evaluating personalization.

5.3 Trustworthiness

Ensuring the trustworthiness of PGen is critical for fostering user confidence and promoting responsible deployment. Key considerations include:

- ***Privacy***. PGen relies on user-specific data for content personalization, raising concerns about privacy issues. Striking a balance between effective personalization and robust privacy protection is crucial for advancing this field.
- ***Fairness and Bias***. PGen may unintentionally reinforce biases and stereotypes present in training data, leading to skewed or discriminatory outcomes. Effective detection and mitigation strategies are essential to protect diverse user groups.
- ***Safety***. Establishing transparent governance protocols, reliable moderation mechanisms, and explainable generation processes is crucial to maintaining user trust and upholding safety standards.

6 Conclusion

In this work, we present the first comprehensive survey on PGen across multiple modalities, offering an in-depth review of recent advancements and emerging trends in the field. To unify existing research, we introduce a holistic framework that formalizes diverse user-specific data, core objectives, and general workflows for PGen, providing a structured foundation for future developments. We then propose a multi-level taxonomy to categorize PGen methods based on modality, user inputs, and specific tasks. Additionally, we summarize the commonly used datasets and evaluation metrics for each modality. Beyond technical aspects, we explore PGen's potential applications in both content creation and delivery, underscoring its significant economic and research value. Finally, we identify key research challenges that remain to be addressed. As a rapidly evolving field, PGen holds great potential to revolutionize the online content ecosystem, enabling more tailored and engaging user experiences. By unifying PGen research across multiple modalities, this survey serves as a valuable resource for fostering cross-modal knowledge sharing and collaboration in this field, contributing to a more personalized digital landscape.

Limitations

In this paper, we provide a comprehensive survey of personalized generation. However, the rapid evolution of this field makes it challenging to encompass all research efforts, as new methods, datasets, and evaluation metrics continue to emerge, requiring continuous updates to our taxonomy. Furthermore, the development of more effective and universally accepted benchmarks within different modalities remains an ongoing challenge.

References

Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 2023. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4552–4562.

Panos Achlioptas, Alexandros Benetatos, Iordanis Fostiropoulos, and Dimitris Skourtis. 2023. Stellar: Systematic evaluation of human-centric personalized text-to-image methods. *arXiv preprint arXiv:2312.06116*.

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10.

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 2025. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pages 73–91. Springer.

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Ruichuan An, Sihan Yang, Ming Lu, Kai Zeng, Yulin Luo, Ying Chen, Jiajun Cao, Hao Liang, Qi She, Shanghang Zhang, et al. 2024. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. Pens: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12.

Matthew O Ayemowa, Roliana Ibrahim, and Muhammad Murad Khan. 2024. Analysis of recommender system using generative artificial intelligence: A systematic literature review. *IEEE Access*.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738.

Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2.

Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23393–23402.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Chong Chen, Fuli Feng, and Qi Tian. 2023. A bi-step grounding paradigm for large language models in recommendation systems. *Preprint*, arXiv:2308.08434.

Ahmet Canberk Baykal, Abdul Basit Anees, Duygu Ceylan, Erkut Erdem, Aykut Erdem, and Deniz Yuret. 2023. Clip-guided stylegan inversion for text-driven real image editing. *ACM Transactions on Graphics*, 42(5):1–18.

David Beniaguev. 2022. Synthetic faces high quality (sfhq) dataset.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

M Akmal Butt and Petros Maragos. 1998. Optimum design of chamfer distance transforms. *IEEE Transactions on Image Processing*, 7(10):1477–1484.

Pengshan Cai, Kaiqiang Song, Sangwoo Cho, Hongwei Wang, Xiaoyang Wang, Hong Yu, Fei Liu, and Dong Yu. 2023. Generating user-engaging news headlines. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3265–3280.

Yufei Cai, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hu Han, and Wangmeng Zuo. 2024. Decoupled textual embeddings for customized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 909–917.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. 2022. Pergamo: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, volume 41, pages 293–304. Wiley Online Library.

Francesco Paolo Casale, Adrian Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. 2018. Gaussian process prior variational autoencoders. *Advances in neural information processing systems*, 31.

Daewon Chae, Nokyung Park, Jinkyu Kim, and Kimin Lee. 2023. Instructbooth: Instruction-following personalized text-to-image generation. *arXiv preprint arXiv:2312.03011*.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942.

Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. 2023. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*.

Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. 2024. Still-moving: Customized video generation without customized video data. *ACM Transactions on Graphics (TOG)*, 43(6):1–11.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024a. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*.

Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. 2023a. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*.

Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. 2024b. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3637–3646.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024c. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

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

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024e. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535.

Liang Chen, Hongru Wang, Yang Deng, Wai Chung Kwan, Zezhong Wang, and Kam-Fai Wong. 2023b. Towards robust personalized dialogue generation via order-insensitive representation regularization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7337–7345, Toronto, Canada. Association for Computational Linguistics.

Peng Chen, Xiaobao Wei, Ming Lu, Yitong Zhu, Naiming Yao, Xingyu Xiao, and Hui Chen. 2023c. Diffusiontalker: Personalization and acceleration for speech-driven 3d face diffuser. *arXiv preprint arXiv:2311.16565*.

Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2662–2670.

Wenhu Chen, Hexiang Hu, YANDONG LI, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. 2023d. Subject-driven text-to-image generation via apprenticeship learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024f. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665. ELRA and ICCL.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. 2024g. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7727–7736.

Jeongsoo Choi, Minsu Kim, Se Jin Park, and Yong Man Ro. 2024. Text-driven talking face synthesis by reprogramming audio-driven models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8065–8069. IEEE.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. 2025. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pages 206–235. Springer.

Schuhmann Christoph and Beaumont Romain. 2022. Laion-aesthetics.

Chih-Hsing Chu, I-Jan Wang, Jeng-Bang Wang, and Yuan-Ping Luh. 2017. 3d parametric human face modeling for personalized product design: Eyeglasses frame design case. *Advanced Engineering Informatics*, 32:202–223.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Joon Son Chung and Andrew Zisserman. 2017a. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer.

Joon Son Chung and Andrew Zisserman. 2017b. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. 2022. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pages 558–577. Springer.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10101–10111.

Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. 2024a. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8235–8239.

Can Cui, Zichong Yang, Yupeng Zhou, Juntong Peng, Sung-Yeon Park, Cong Zhang, Yunsheng Ma, Xu Cao, Wenqian Ye, Yiheng Feng, et al. 2024b. On-board vision-language models for personalized autonomous vehicle motion control: System design and real-world validation. *arXiv preprint arXiv:2411.11913*.

Ádám Tibor Czapp, Mátyás Jani, Bálint Domián, and Balázs Hidasi. 2024. Dynamic product image generation and recommendation at scale for personalized e-commerce. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 768–770.

Shuqi Dai, Xichu Ma, Ye Wang, and Roger B Dannenberg. 2022. Personalised popular music generation using imitation and structure. *Journal of New Music Research*, 51(1):69–85.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. 2020. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. 2024. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098.

Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. 2019a. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9026–9035.

Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. 2019b. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1161–1170.

Haoye Dong, Jun Liu, and Dong Huang. 2024. Df-vton: Dense flow guided virtual try-on network. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3175–3179. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Eugene, serdyukovpv, and Will Cukierski. 2013. Personalized web search challenge. https://kaggle.com/competitions/yandex-personalized-web-search-challenge. Kaggle.

Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. 2013. Random forests for real time 3d face analysis. *International journal of computer vision*, 101:437–458.

Yi Fang, Wenjie Wang, Yang Zhang, Fengbin Zhu, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Large language models for recommendation with deliberative user preference alignment. *arXiv preprint arXiv:2502.02061*.

Zhouxiang Fang, Min Yu, Zhendong Fu, Boning Zhang, Xuanwen Huang, Xiaoqi Tang, and Yang Yang. 2024a. How to generate popular post headlines on social media? *AI Open*, 5:1–9.

Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. 2024b. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*.

Najmeh Forouzandehmehr, Yijie Cao, Nikhil Thakurdesai, Ramin Giahi, Luyi Ma, Nima Farrokhsiar, Jianpeng Xu, Evren Korpeoglu, and Kannan Achan. 2023. Character-based outfit generation with vision-augmented style extraction via llms. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1–7. IEEE.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. 2022. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*.

Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2024. Lcm-lookahead for encoder-based text-to-image personalization. In *European Conference on Computer Vision*, pages 322–340. Springer.

Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. 2024. LLM blueprint: Enabling text-to-image generation with complex and detailed prompts. In *The Twelfth International Conference on Learning Representations*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)*, 41(6):1–12.

Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. 2019. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345.

Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard De Melo, and Yongfeng Zhang. 2022. Improving personalized explanation generation through visualization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 244–255.

Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4274–4282.

Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 2023. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7599–7607.

Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. 2020. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 369–385. Springer.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. 2024. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. 2023. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qian Guo, Wei Chen, and Huaiyu Wan. 2021. Aol4ps: A large-scale data set for personalized search. *Data Intelligence*, 3(4):548–567.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024a. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. 2024b. Pulid: Pure and lightning id customization via contrastive alignment. In *Advances in Neural Information Processing Systems*.

Xiaoyu Han, Shunyuan Zheng, Zonglin Li, Chenyang Wang, Xin Sun, and Quanling Meng. 2024. Shape-guided clothing warping for virtual try-on. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2593–2602.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552.

Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. 2024. Remember, retrieve and generate: Understanding infinite visual concepts as your personalized assistant. *arXiv preprint arXiv:2410.13360*.

Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. 2023. Vico: Plug-and-play visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Naomi Harte and Eoin Gillen. 2015. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*.

Junjie He, Yifeng Geng, and Liefeng Bo. 2024a. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. 2024b. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

Yutong He, Alexander Robey, Naoki Murata, Yiding Jiang, Joshua Williams, George J Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J Zico Kolter. 2024c. Automated black-box prompt engineering for personalized text-to-image generation. *arXiv preprint arXiv:2403.19103*.

Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. 2024d. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*.

Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. 2025. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In *European Conference on Computer Vision*, pages 123–139. Springer.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Yan Hong, Yuxuan Duan, Bo Zhang, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. 2025. Comfusion: Enhancing personalized generation by instance-scene compositing and fusion. In *European Conference on Computer Vision*, pages 1–18. Springer.

Alain Hore and Djemel Ziou. 2010. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE.

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

Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. 2024a. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763.

Junxing Hu, Hongwen Zhang, Yunlong Wang, Min Ren, and Zhenan Sun. 2023. Personalized graph generation for monocular 3d human pose and shape estimation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163.

Xinting Hu, Haoran Wang, Jan Eric Lenssen, and Bernt Schiele. 2025. PersonaHOI: Effortlessly improving personalized face with human-object interaction generation. *arXiv preprint*.

Yuke Hu, Jian Lou, Jiaqi Liu, Wangze Ni, Feng Lin, Zhan Qin, and Kui Ren. 2024b. Eraser: Machine unlearning in mlaas via an inference serving-aware approach. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*.

Zhejing Hu, Yan Liu, Gong Chen, and Yongxu Liu. 2022. Can machines generate personalized music? a hybrid favorite-aware method for user preference music transfer. *IEEE Transactions on Multimedia*, 25:2296–2308.

Jiancheng Huang, Mingfu Yan, Songyan Chen, Yi Huang, and Shifeng Chen. 2024a. Magicfight: Personalized martial arts combat video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10833–10842.

Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, and Jie Song. 2024b. Patchdpo: Patch-level dpo for finetuning-free personalized image generation. *arXiv preprint arXiv:2412.03177*.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910.

Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. 2024c. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36.

Zhichao Huang, Xintong Han, Jia Xu, and Tong Zhang. 2021. Few-shot human motion transfer by personalized geometry and texture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2297–2306.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024d. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C.K. Chan, and Ziwei Liu. 2024e. ReVersion: Diffusion-based relation inversion from images. In *SIGGRAPH Asia 2024 Conference Papers*.

Minyoung Hwang, Luca Weihs, Chanwoo Park, Kimin Lee, Aniruddha Kembhavi, and Kiana Ehsani. 2024. Promptable behaviors: Personalizing multi-objective rewards from human preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16216–16226.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339.

Yasamin Jafarian and Hyun Soo Park. 2021. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762.

Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. 2020. Sievenet: A unified framework for robust image-based virtual try-on. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2182–2190.

Yoonna Jang, Jung Hoon Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2021. Call for customized conversation: Customized conversation grounding persona and knowledge. In *AAAI Conference on Artificial Intelligence*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. 2022a. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10799–10808.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. 2024. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022b. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11.

Yuming Jiang, Nanxuan Zhao, Qing Liu, Krishna Kumar Singh, Shuai Yang, Chen Change Loy, and Ziwei Liu. 2025. Groupdiff: Diffusion-based group portrait editing. In *European Conference on Computer Vision*, pages 221–239. Springer.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer.

Wijdane Kaiss, Khalifa Mansouri, and Franck Poirier. 2023. Pre-Evaluation with a Personalized Feedback Conversational Agent Integrated in Moodle. *International Journal of Emerging Technologies in Learning*, 18(06):177–189.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405.

Tero Karras, Samuli Laine, and Timo Aila. 2021. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228.

Pranav Kasela, Marco Braga, Gabriella Pasi, and Raffaele Perego. 2024. Se-pqa: Personalized community question answering. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1095–1098, New York, NY, USA. Association for Computing Machinery.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157.

Diane Kelly and Nicholas J. Belkin. 2002. A user modeling system for personalized interaction and tailored retrieval in interactive ir. *Proceedings of the ASIST Annual Meeting*, 39:316–325.

Taekyung Ki and Dongchan Min. 2023. Stylelipsync: Style-based personalized lip-sync video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22841–22850.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2024a. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185.

Kangyeol Kim, Wooseok Seo, Sehyun Nam, Bodam Kim, Suhyeon Jeong, Wonwoo Cho, Jaegul Choo, and Youngjae Yu. 2024b. Layout-and-retouch: A dual-stage framework for improving diversity in personalized image generation. *arXiv preprint arXiv:2407.09779*.

Minchul Kim, Anil K Jain, and Xiaoming Liu. 2022. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663.

Jaehoon Ko, Kyusun Cho, Joungbin Lee, Heeji Yoon, Sangmin Lee, Sangjun Ahn, and Seungryong Kim. 2024. Talk3d: High-fidelity talking portrait synthesis via personalized 3d generative prior. *arXiv preprint arXiv:2403.20153*.

Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. 2019. The personalization of conversational agents in health care: Systematic review. *J Med Internet Res*, 21(11):e15360.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *Preprint*, arXiv:2302.14520.

Satwik Kottur, Xiaoyu Wang, and Vitor Carvalho. 2017. Exploring personalized neural conversational models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3728–3734.

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. 2023. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17089–17099.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024a. Longlamp: A benchmark for personalized long-form text generation. *Preprint*, arXiv:2407.11016.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith, and Hannaneh Hajishirzi. 2024b. Compo: Community preferences for language model personalization. *Preprint*, arXiv:2410.16027.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941.

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. 2022. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer.

Boyi Li, Jathushan Rajasegaran, Yossi Gandelsman, Alexei A Efros, and Jitendra Malik. 2024a. Synthesizing moving people with 3d control. *arXiv preprint arXiv:2401.10889*.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023a. Teach llms to personalize – an approach inspired by writing education. *Preprint*, arXiv:2308.07968.

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

Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. 2024c. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. *arXiv preprint arXiv:2411.17048*.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. 2023b. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 755–764.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601*.

Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. 2024d. UNIMO-G: Unified image generation through multimodal conditional diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6173–6188. Association for Computational Linguistics.

Xiang Li, Lei Meng, Lei Wu, Manyi Li, and Xiangxu Meng. 2024e. Dreamfont3d: personalized text-to-3d artistic font generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11.

Xiaoming Li, Xinyu Hou, and Chen Change Loy. 2024f. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2187–2196.

Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. 2022. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5904–5917.

Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. 2024g. Personalized language modeling from personalized human feedback. *Preprint*, arXiv:2402.05133.

Yameng Li, Shi Feng, Daling Wang, Yifei Zhang, and Xiaocui Yang. 2025. Paper: A persona-aware chain-of-thought learning framework for personalized dialogue response generation. In *Natural Language Processing and Chinese Computing*, pages 215–227, Singapore. Springer Nature Singapore.

Yang Li, Songlin Yang, Wei Wang, and Jing Dong. 2024h. Sefi-ide: Semantic-fidelity identity embedding for personalized diffusion-based generation. *arXiv preprint arXiv:2402.00631*.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024i. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Zhaojian Li, Bin Zhao, and Yuan Yuan. 2024j. Tas: Personalized text-guided audio spatialization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9029–9037.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024k. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650.

Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. 2018. Vmaf: The journey continues. *Netflix Technology Blog*, 25(1).

Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. 2021. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 653–661.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024a. Data-efficient fine-tuning for llm-based recommendation. *Preprint*, arXiv:2401.17197.

Xudong Lin, Ali Zare, Shiyuan Huang, Ming-Hsuan Yang, Shih-Fu Chang, and Li Zhang. 2024b. Personalized video comment generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16806–16820.

Fangfu Liu, Hanyang Wang, Weiliang Chen, Haowen Sun, and Yueqi Duan. 2025. Make-your-3d: Fast and consistent subject-driven 3d content generation. In *European Conference on Computer Vision*, pages 389–406. Springer.

Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Ying Shan, and Yujiu Yang. 2024a. Stylecrafter: Taming artistic video diffusion with reference-augmented adapter learning. *ACM Transactions on Graphics (TOG)*, 43(6):1–10.

Hanwen Liu, Zhicheng Sun, and Yadong Mu. 2024b. Countering personalized text-to-image generation with influence watermarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12257–12267.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jia Liu, Changlin Li, Qirui Sun, Jiahui Ming, Chen Fang, Jue Wang, Bing Zeng, and Shuaicheng Liu. 2024c. Ada-adapter: Fast few-shot style personlization of diffusion model with pre-trained image encoder. *arXiv preprint arXiv:2407.05552*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051.

Yexin Liu and Lin Wang. 2024. Mycloth: Towards intelligent and interactive online t-shirt customization based on user's preference. *arXiv preprint arXiv:2404.15801*.

Yilun Liu, Minggui He, Feiyu Yao, Yuhe Ji, Shimin Tao, Jingzhou Du, Duan Li, Jian Gao, Li Zhang, Hao Yang, et al. 2024d. What do you want? user-centric prompt generation for text-to-image synthesis via multi-turn guidance. *arXiv preprint arXiv:2408.12910*.

Yixin Liu, Ruoxi Chen, Xun Chen, and Lichao Sun. 2024e. Rethinking and defending protective perturbation in personalized diffusion models. *arXiv preprint arXiv:2406.18944*.

Zhilei Liu, Xiaoxing Liu, Sen Chen, Jiaxing Liu, Longbiao Wang, and Chongke Bi. 2024f. Multimodal fusion for talking face generation utilizing speech-related facial action units. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–24.

Cuirong Long, Xiaoshan Yang, and Changsheng Xu. 2020. Cross-domain personalized image captioning. *Multimedia Tools and Applications*, 79(45):33333–33348.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980.

Zhi Lu, Yang Hu, Yunchao Jiang, Yan Chen, and Bing Zeng. 2019. Learning binary code for personalized fashion recommendation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10562–10570.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. 2023. Notice of removal: Deltaedit: Exploring text-free training for text-driven image manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6894–6903.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2024a. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12.

Xichu Ma, Yuchen Wang, and Ye Wang. 2022. Content based user preference modeling in music generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2473–2482.

Xichu Ma, Yuchen Wang, and Ye Wang. 2024b. Symbolic music generation from graph-learning-based preference modeling and textual queries. *IEEE Transactions on Multimedia*.

Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*.

Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. 2024c. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*.

Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Chen Qian, Zhaoxiang Zhang, and Zhen Lei. 2024d. Diffspeaker: Speech-driven 3d facial animation with diffusion transformer. *arXiv preprint arXiv:2402.05712*.

Liam Magee, Vanicka Arora, Gus Gollings, and Norma Lam-Saw. 2024. The drama machine: Simulating character development with llm agents. *Preprint*, arXiv:2408.01725.

Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. 2020. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012a. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012b. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212.

Shentong Mo, Jing Shi, and Yapeng Tian. 2023. Diffava: Personalized text-to-audio generation with visual alignment. *arXiv preprint arXiv:2305.12903*.

Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. 2022. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *Preprint*, arXiv:2311.09180.

Ofir Nabati, Guy Tennenholtz, ChihWei Hsu, Moonkyung Ryu, Deepak Ramachandran, Yinlam Chow, Xiang Li, and Craig Boutilier. 2024. Personalized and sequential text-to-image generation. *arXiv preprint arXiv:2412.10419*.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027.

Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. 2024. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110.

Niranjan D Narvekar and Lina J Karam. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683.

Man Tik Ng, Hui Tung Tse, Jen-Tse Huang, Jingjing Li, Wenxuan Wang, and Michael R. Lyu. 2024. How well can llms echo us? evaluating ai chatbots' role-play ability with echo. *ArXiv*, abs/2404.13957.

Hung Nguyen, Quang Qui-Vinh Nguyen, Khoi Nguyen, and Rang Nguyen. 2024a. Swifttry: Fast and consistent video virtual try-on with diffusion models. *arXiv preprint arXiv:2412.10178*.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. 2024b. Yo'LLaVA: Your personalized language and vision assistant. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Armand Nicolicioiu, Eugenia Iofinova, Eldar Kurtic, Mahdi Nikdan, Andrei Panferov, Ilia Markov, Nir Shavit, and Dan Alistarh. 2024. Panza: A personalized text writing assistant via data playback and local fine-tuning. *Preprint*, arXiv:2407.10994.

Shuliang Ning, Duomin Wang, Yipeng Qin, Zirong Jin, Baoyuan Wang, and Xiaoguang Han. 2024. Picture: Photorealistic virtual try-on from unconstrained designs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6985.

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. 2022. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2203.17272*.

Takuto Onikubo and Yusuke Matsui. 2024. High-frequency anti-dreambooth: Robust defense against personalized image synthesis. *arXiv preprint arXiv:2409.08167*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Nadav Orzech, Yotam Nitzan, Ulysse Mizrahi, Dov Danon, and Amit H Bermano. 2024. Masked extended attention for zero-shot virtual try-on in the wild. *arXiv preprint arXiv:2406.15331*.

Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. 2024. Attndreambooth: Towards text-aligned personalized text-to-image generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Rishubh Parihar, Harsh Gupta, Sachidanand VS, and R Venkatesh Babu. 2024. Text2place: Affordance-aware text guided human placement. In *European Conference on Computer Vision*, pages 57–77.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2018. Towards personalized image captioning via multimodal memory networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):999–1012.

Junseo Park, Beomseok Ko, and Hyeryung Jang. 2024. Text-to-image synthesis for any artistic styles: Advancements in personalized artistic image generation via subdivision and dual binding. *arXiv preprint arXiv:2404.05256*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024a. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14554–14562.

Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. 2024b. λ-ECLIPSE: Multi-concept personalized text-to-image diffusion models by leveraging CLIP latent space. *Transactions on Machine Learning Research*.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. 2024. Personalized large vision-language models. *arXiv preprint arXiv:2412.17610*.

Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. 2024. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*.

Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis. 2024. Investigating personalization methods in text to music generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1081–1085. IEEE.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Alisha Pradhan and Amanda Lazar. 2021. Hey google, do you have a personality? designing personality and personas for conversational agents. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA. Association for Computing Machinery.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 484–492. Association for Computing Machinery.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*.

Luchao Qi, Jiaye Wu, Shengze Wang, and Soumyadip Sengupta. 2023a. My3dgen: Building lightweight personalized 3d generative model. *arXiv preprint arXiv:2307.05468*.

Zipeng Qi, Xulong Zhang, Ning Cheng, Jing Xiao, and Jianzong Wang. 2023b. Difftalker: Co-driven audio-image diffusion for talking faces via intermediate landmarks. *arXiv preprint arXiv:2309.07509*.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2470–2477, New York, NY, USA. Association for Computing Machinery.

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

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2349–2359.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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.

Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. 2020. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10):2586–2606.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 752–762, New York, NY, USA. Association for Computing Machinery.

Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025a. Expert: Effective and explainable evaluation of personalized long-form text generation. *Preprint*, arXiv:2501.14956.

Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, Tao Chen, Zhuowan Li, Michael Bendersky, and Hamed Zamani. 2025b. Reasoning-enhanced self-training for long-form personalized text generation. *Preprint*, arXiv:2501.04167.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392. Association for Computational Linguistics.

Alireza Salemi and Hamed Zamani. 2024. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. *Preprint*, arXiv:2409.09510.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In

Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-play with large language models. *Preprint*, arXiv:2305.16367.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. 2024a. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2023. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991.

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

Zheng-Yan Sheng, Yang Ai, and Zhen-Hua Ling. 2023. Zero-shot personalized lip-to-speech synthesis with face image based voice control. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552.

Lei Shi, Alexandra Ioana Cristea, Malik Shahzad Kaleem Awan, Craig D. Stewart, and Maurice Hendrix. 2013. Towards understanding learning behavior patterns in social adaptive personalized e-learning systems. In *Americas Conference on Information Systems*.

Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. 2023a. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12469–12480.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2023b. Mvdream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*.

Veronika Shilova, Ludovic Dos Santos, Flavian Vasile, Gaëtan Racic, and Ugo Tanielian. 2023. Adbooster: Personalized ad creative generation using stable diffusion outpainting. *arXiv preprint arXiv:2309.11507*.

Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Customized image narrative generation via interactive visual question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8925–8933.

Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662.

Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, Yuan Hao, Glenn Entis, Irina Blok, and Daniel Castro Chin. 2023. Styledrop: Text-to-image synthesis of any style. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5190–5196. International Joint Conferences on Artificial Intelligence Organization.

Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. 2025. Moma: Multimodal llm adapter for fast personalized image generation. In *European Conference on Computer Vision*, pages 117–132. Springer.

Wenfeng Song, Xuan Wang, Shi Zheng, Shuai Li, Aimin Hao, and Xia Hou. 2024a. Talkingstyle: Personalized speech-driven 3d facial animation with style preservation. *IEEE Transactions on Visualization and Computer Graphics*.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. 2024b. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8048–8058.

Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. 2023. General then personal: Decoupling and pre-training for personalized headline generation. *Transactions of the Association for Computational Linguistics*, 11:1588–1607.

K Soomro. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Ke Sun, Jian Cao, Qi Wang, Linrui Tian, Xindi Zhang, Lian Zhuo, Bang Zhang, Liefeng Bo, Wenbo Zhou, Weiming Zhang, et al. 2024. Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person. *arXiv preprint arXiv:2407.16224*.

Peijie Sun, Le Wu, Kun Zhang, Yanjie Fu, Richang Hong, and Meng Wang. 2020. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. In *Proceedings of The Web Conference 2020*, pages 837–847.

Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. 2024. Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. In *Interspeech 2024*, pages 1380–1384.

Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. 2025. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*.

Brian Jay Tang, Kaiwen Sun, Noah T Curran, Florian Schaub, and Kang G Shin. 2024a. Genai advertising: Risks of personalizing ads with llms. *arXiv preprint arXiv:2409.15436*.

Yihong Tang, Bo Wang, Dongming Zhao, Jinxiaojia Jinxiaojia, Zhangjijun Zhangjijun, Ruifang He, and Yuexian Hou. 2024b. MORPHEUS: Modeling role from personalized dialogue history by exploring and utilizing latent space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7664–7676, Miami, Florida, USA. Association for Computational Linguistics.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11.

Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley. 2004. A personalized system for conversational recommendations. *J. Artif. Int. Res.*, 21(1):393–428.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. 2025. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631. Association for Computational Linguistics.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. 2023. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. 2023. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127.

Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter Harrison, Elisabeth André, and Nori Jacoby. 2022. Voiceme: Personalized voice generation in tts. In *Interspeech 2022*, pages 2588–2592.

Shanu Vashishtha, Abhinav Prakash, Lalitesh Morishetti, Kaushiki Nag, Yokila Arora, Sushant Kumar, and Kannan Achan. 2024. Chaining text-to-image and large language model: A novel approach for generating personalized e-commerce banners. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5825–5835.

Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617.

Dimitri Von Rütte, Elisabetta Fedele, Jonathan Thomm, and Lukas Wolf. 2023. Fabric: Personalizing diffusion models with iterative feedback. *arXiv preprint arXiv:2307.10159*.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*.

Cong Wan, Yuhang He, Xiang Song, and Yihong Gong. 2024. Prompt-agnostic adversarial perturbation for customized diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

Siqi Wan, Yehao Li, Jingwen Chen, Yingwei Pan, Ting Yao, Yang Cao, and Tao Mei. 2025. Improving virtual try-on with garment-focused diffusion models. In *European Conference on Computer Vision*, pages 184–199. Springer.

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018a. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024a. Learning personalized alignment for evaluating open-ended text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13274–13292, Miami, Florida, USA. Association for Computational Linguistics.

Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. 2024b. Animatelcm: Computation-efficient personalized style video generation without personalized video data. In *SIGGRAPH Asia 2024 Technical Communications*, pages 1–5.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018b. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274.

Haotian Wang, Yuzhe Weng, Yueyan Li, Zilu Guo, Jun Du, Shutong Niu, Jiefeng Ma, Shan He, Xiaoyan Wu, Qiming Hu, et al. 2024c. Emotivetalk: Expressive talking head generation through audio information decoupling and emotional video diffusion. *arXiv preprint arXiv:2411.16726*.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024d. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *Preprint*, arXiv:2401.13256.

Jianrong Wang, Zixuan Wang, Xiaosheng Hu, Xuewei Li, Qiang Fang, and Li Liu. 2022. Residual-guided personalized speech synthesis based on face image. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4743–4747. IEEE.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024e. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

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

Quan Wang, Sheng Li, Xinpeng Zhang, and Guorui Feng. 2023a. Rethinking neural style transfer: Generating personalized and watermarked stylized images. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6928–6937.

Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023b. Disco: Disentangled control for referring human dance generation in real world. *arXiv e-prints*, pages arXiv–2307.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018c. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1152–1164. Curran Associates Inc.

Weijie Wang, Jichao Zhang, Chang Liu, Xia Li, Xingqian Xu, Humphrey Shi, Nicu Sebe, and Bruno Lepri. 2024g. Uvmap-id: A controllable and personalized uv map generative model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10725–10734.

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

Wenjie Wang, Xinyu Lin, Liuhui Wang, Fuli Feng, Yunshan Ma, and Tat-Seng Chua. 2023d. Causal disentangled recommendation against user preference shifts. *ACM Transactions on Information Systems*, 42(1):1–27.

Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. 2023e. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 832–841, New York, NY, USA. Association for Computing Machinery.

X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. 2024h. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*.

Xianquan Wang, Likang Wu, Shukang Yin, Zhi Li, Yanjiang Chen, Hufeng Hufeng, Yu Su, and Qi Liu. 2024i. 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.

Xuan Wang, Guanhong Wang, Wenhao Chai, Jiayu Zhou, and Gaoang Wang. 2023f. User-aware prefix-tuning is a good learner for personalized image captioning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 384–395. Springer.

Yanan Wang, Yan Pei, Zerui Ma, and Jianqiang Li. 2024j. A user-guided generation framework for personalized music synthesis using interactive evolutionary computation. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1762–1769.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023g. Automated evaluation of personalized text generation using large language models. *Preprint*, arXiv:2310.11593.

Yuanbin Wang, Weilun Dai, Long Chan, Huanyu Zhou, Aixi Zhang, and Si Liu. 2024k. Gpd-vvto: Preserving garment details in video virtual try-on. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7133–7142.

Zhao Wang, Aoxue Li, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. 2024l. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*.

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

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.

Fanyue Wei, Wei Zeng, Zhenyang Li, Dawei Yin, Lixin Duan, and Wen Li. 2025a. Powerful and flexible: Personalized text-to-image generation via reinforcement learning. In *European Conference on Computer Vision*, pages 394–410. Springer.

Huawei Wei, Zejun Yang, and Zhisheng Wang. 2024a. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. 2024b. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. 2024c. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*.

Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hongzhi Zhang, Lei Zhang, and Wangmeng Zuo. 2025b. Masterweaver: Taming editability and face identity for personalized text-to-image generation. In *European Conference on Computer Vision*, pages 252–271. Springer.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953.

Matthias Wright and Björn Ommer. 2022. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pages 560–576. Springer.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020a. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020b. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3597–3606.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023a. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154.

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. 2024a. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, et al. 2024b. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024c. A survey on large language models for recommendation. *Preprint*, arXiv:2305.19860.

Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. 2024d. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023b. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105.

Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. 2025a. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. In *European Conference on Computer Vision*, pages 279–296. Springer.

Yihan Wu, Ruihua Song, Xu Chen, Hao Jiang, Zhao Cao, and Jin Yu. 2024e. Understanding human preferences: Towards more personalized video to text generation. In *Proceedings of the ACM on Web Conference 2024*, pages 3952–3963.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. 2025b. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Yujia Wu, Yiming Shi, Jiwei Wei, Chengwei Sun, Yuyang Zhou, Yang Yang, and Heng Tao Shen. 2024f. Difflora: Generating personalized low-rank adaptation weights with diffusion. *arXiv preprint arXiv:2408.06740*.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. 2024a. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. 2024b. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*.

Zhujun Xiao, Jenna Cryan, Yuanshun Yao, Yi Hong Cheo, Yuanchao Shu, Stefan Saroiu, Ben Y Zhao, and Haitao Zheng. 2023. "my face, my rules": Enabling personalized protection against unacceptable face editing. *Proceedings on Privacy Enhancing Technologies*, 2023(3).

Zhenyu Xie, Haoye Dong, Yufei Gao, Zehua Ma, and Xiaodan Liang. 2024. Dreamvton: Customizing 3d virtual try-on with personalized diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10784–10793.

Kun Xiong, Liu Jiang, Xuan Dang, Guolong Wang, Wenwen Ye, and Zheng Qin. 2020. Towards personalized aesthetic image caption. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitris Tzionas, and Michael J Black. 2024. Puzzleavatar: Assembling 3d avatars from personal albums. *ACM Transactions on Graphics (TOG)*, 43(6):1–15.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023a. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023b. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023c. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398, Toronto, Canada. Association for Computational Linguistics.

Yifei Xu, Xiaolong Xu, Honghao Gao, and Fu Xiao. 2024a. Sgdm: An adaptive style-guided diffusion model for personalized text to image generation. *IEEE Transactions on Multimedia*.

Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. 2024b. Diffusion models for generative outfit recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1350–1359.

Yiyan Xu, Wenjie Wang, Yang Zhang, Tang Biao, Peng Yan, Fuli Feng, and Xiangnan He. 2024c. Personalized image generation with large multimodal models. *arXiv preprint arXiv:2410.14170*.

Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. 2024d. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*.

Zhengze Xu, Mengting Chen, Zhao Wang, Linyu Xing, Zhonghua Zhai, Nong Sang, Jinsong Lan, Shuai Xiao, and Changxin Gao. 2024e. Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3199–3208.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024f. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490.

Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. 2023. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023a. Palr: Personalization aware llms for recommendation. *Preprint*, arXiv:2305.07622.

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

Jianan Yang, Haobo Wang, Yanming Zhang, Ruixuan Xiao, Sai Wu, Gang Chen, and Junbo Zhao. 2023b. Controllable textual inversion for personalized text-to-image generation. *arXiv preprint arXiv:2304.05265*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024b. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200.

Wanting Yang, Zehui Xiong, Song Guo, Shiwen Mao, Dong In Kim, and Merouane Debbah. 2024c. Efficient multi-user offloading of personalized diffusion models: A drl-convex hybrid solution. *arXiv preprint arXiv:2411.15781*.

Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. 2024d. Loracomposer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*.

Zhengyi Yang, Jiancan Wu, Zhicai Wang, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024e. Generate what you prefer: reshaping sequential recommendation via guided diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zhijing Yang, Mingliang Yang, Siyuan Peng, Jieming Xie, and Chao Long. 2024f. Animated clothing fitting: Semantic-visual fusion for video customization virtual try-on with diffusion models. In *2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*, pages 675–680. IEEE.

Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. 2022. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*.

Zebin Yao, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2024. Concept conductor: Orchestrating multiple personalized concepts in text-to-image synthesis. *arXiv preprint arXiv:2408.03632*.

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

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. 2024. Stylemaster: Stylize your video with artistic generation and translation. *arXiv preprint arXiv:2412.07744*.

Yu-Ying Yeh, Jia-Bin Huang, Changil Kim, Lei Xiao, Thu Nguyen-Phuoc, Numair Khan, Cheng Zhang, Manmohan Chandraker, Carl S Marshall, Zhao Dong, et al. 2024. Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4304–4314.

Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*.

Cong Yu, Yang Hu, Yan Chen, and Bing Zeng. 2019. Personalized fashion design. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9046–9055.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. 2023. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814.

Zhengyang Yu, Zhaoyuan Yang, and Jing Zhang. 2024. Dreamsteerer: Enhancing source image conditioned editability using personalized diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. 2023. Inserting anybody in diffusion models via celeb basis. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. 2024. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*.

Dongxu Yue, Maomao Li, Yunfei Liu, Qin Guo, Ailing Zeng, Tianyu Yang, and Yu Li. 2025. Addme: Zero-shot group-photo synthesis by inserting people into scenes. In *European Conference on Computer Vision*, pages 222–239. Springer.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. 2019. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. 2021. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pages 726–747. PMLR.

Hansi Zeng, Surya Kallumadi, Zaid Alibadi, Rodrigo Nogueira, and Hamed Zamani. 2023. A personalized dense retrieval framework for unified information access. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 121–130, New York, NY, USA. Association for Computing Machinery.

Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, HsiangTao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. 2023a. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105.

Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. 2017. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200.

Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 2021a. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1438–1449.

Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Xiaohu Guo, and Jiashi Feng. 2023b. Dreamtalk: diffusion-based realistic emotional audio-driven method for single image talking face generation. *arXiv preprint arXiv:2312.13578*.

Jiarui Zhang. 2024. Guided profile generation improves personalization with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4005–4016.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024a. Llm-based medical assistant personalization with short-and long-term memory coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2386–2398.

Kai Zhang, Lizhi Qing, Yangyang Kang, and Xiaozhong Liu. 2024b. Personalized llm response generation with parameterized memory injection. *arXiv preprint arXiv:2404.03565*.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Mozhi Zhang, Pengyu Wang, Chenkun Tan, Mianqiu Huang, Dong Zhang, Yaqian Zhou, and Xipeng Qiu. 2024c. Metaalign: Align large language models with diverse preferences during inference time. *arXiv preprint arXiv:2410.14184*.

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

Shufang Zhang, Minxue Ni, Shuai Chen, Lei Wang, Wenxin Ding, and Yuhong Liu. 2024d. A two-stage personalized virtual try-on framework with shape control and texture guidance. *IEEE Transactions on Multimedia*.

Tianshu Zhang, Buzhen Huang, and Yangang Wang. 2020a. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhang, Yue Ying, Pan Lu, and Hongyuan Zha. 2020c. Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9571–9578.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023c. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081.

Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. 2024e. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*.

Xulu Zhang, Xiao-Yong Wei, Jinlin Wu, Tianyi Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. 2024f. Compositional inversion for stable diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7350–7358.

Xulu Zhang, Xiao-Yong Wei, Wengyu Zhang, Jinlin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. 2024g. A survey on personalized content synthesis with diffusion models. *arXiv preprint arXiv:2405.05538*.

Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. 2024h. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7747–7756.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. 2024i. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024j. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021b. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670.

Haoyu Zhao, Tianyi Lu, Jiaxi Gu, Xing Zhang, Qingping Zheng, Zuxuan Wu, Hang Xu, and Yu-Gang Jiang. 2025. Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In *European Conference on Computer Vision*, pages 205–221. Springer.

Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2018. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 792–800.

Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. 2024a. Viton-dit: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint arXiv:2405.18326*.

Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. 2024b. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9693–9700.

Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. 2021. Mv-ton: Memory-based video virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 908–916.

Yong Zhong, Min Zhao, Zebin You, Xiaofeng Yu, Changwang Zhang, and Chongxuan Li. 2025. Posecrafter: One-shot personalized video synthesis following flexible pose control. In *European Conference on Computer Vision*, pages 243–260. Springer.

Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. 2019. Iou loss for 2d/3d object detection. In *2019 international conference on 3D vision (3DV)*, pages 85–94. IEEE.

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

Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. 2023. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*.

Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Li Xiu. 2024. Multibooth: Towards generating all your concepts in an image from text. *arXiv preprint arXiv:2404.14239*.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. 2023a. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615.

Luyao Zhu, Wei Li, Rui Mao, Vlad Pandelea, and Erik Cambria. 2023b. PAED: Zero-shot persona attribute extraction in dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9771–9787, Toronto, Canada. Association for Computational Linguistics.

Yizhe Zhua, Chunhui Zhanga, Qiong Liub, and Xi Zhoub. 2023. Audio-driven talking head video generation with diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. HYDRA: Model factorization framework for black-box LLM personalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Personalized Generation Across Modalities

A.1 Personalized Image Generation

Personalized image generation aims to synthesize images that reflect individual preferences and requirements. By incorporating various personalized contexts, existing studies have made significant strides in enhancing the capability of generative models to produce images tailored to specific needs, ranging from general-purpose generation to more specialized tasks.

A.1.1 User Behaviors

User interactions serve as a key source for inferring visual preferences, guiding the personalized image generation process. Based on user behaviors such as historical engagements and real-time feedback, existing methods have explored various approaches to enhance personalization.

General-purpose Image Generation This task involves generating tailored images across various scenarios. For instance, PMG (Shen et al., 2024b), I-AM-G (Wang et al., 2024i), and Pigeon (Xu et al., 2024c) utilize historically interacted images of users to infer their visual tastes, enabling personalized generation across various scenarios, including stickers, movie posters, fashion designs, and news posters. In specific, PMG converts historical interacted images into textual descriptions, allowing LLMs to distill user preferences effectively; I-AM-G introduces an interest rewrite strategy to address preference ambiguity and leverages retrieval-augmented generation (RAG) to enrich semantic information; and Pigeon leverages MLLMs with specialized modules to manage noisy historical data and multimodal instructions, ensuring precise user modeling. Similarly, SGDM (Xu et al., 2024a) enhances personalization by introducing a style extraction module that captures user-specific style preferences to guide image generation. Personalized PR (Chen et al., 2024g) proposes a personalized prompt-rewriting method, leveraging historical user query-image pairs to enhance text-to-image (T2I) personalization.

Beyond inferring user preferences from interaction history, some studies (Von Rütte et al., 2023; Liu et al., 2024d; Nabati et al., 2024) have explored interactive, personalized image generation by incorporating real-time user feedback through multi-turn interactions, thereby progressively refining the generated outputs.

Fashion Design Generation This task involves personalized fashion design with inferring personal style preferences from user behaviors. Yu et al. (2019) employs GANs to learn user preference vectors from interaction history, generating fashion images compatible with provided images. DiFashion (Xu et al., 2024b) adopts diffusion models to extract user preferences from interaction history for personalized outfit generation and recommendation.

E-commerce Product Image Generation This task aims to create customized, eye-catching visuals for e-commerce products to attract target consumers. Based on user behaviors, AdBooster (Shilova et al., 2023) utilizes the Stable Diffusion outpainting model, conditioning on individual user interest to generate appealing product images. Vashishtha et al. (2024) leverages LLMs to generate text prompts for diffusion models to craft engaging banners. Czapp et al. (2024) employs a contextual bandit algorithm to select prompts from a pool, generating personalized product backgrounds.

A.1.2 User Profiles

Some studies utilize users' demographic attributes to infer preferences or categorize them into groups for personalized image generation.

Fashion Design Generation Base on user attributes (*e.g.,* age, gender, interests in characters), LVA-COG (Forouzandehmehr et al., 2023) utilizes LLMs to extract user preferences to guide fashion design generation for recommendation.

E-commerce Product Image Generation By categorizing users into distinct groups based on their attributes, CG4CTR (Yang et al., 2024a) proposes a self-cyclic generation pipeline to produce tailored product images for each user group.

A.1.3 Personalized Subjects

This is a primary focus of the computer vision community, which aims to capture the subject representation from a limited set of subject images and follow users' instructions for subject-driven text-to-image (T2I) generation. For instance, given a few images of a user's pet, the model generates new images featuring the pet in different contexts or environments while preserving its identity.

Subject-driven T2I Generation Existing research in this area can be broadly categorized into two branches:

- Optimization-based methods introduce a learnable unique identifier in the embedding space to encapsulate the semantics and visual details of each subject. Specifically, Textual Inversion (Gal et al., 2023) optimizes a pseudo-word identifier, denoted as S^*, to encode a personalized representation that guides T2I generation in diffusion models. DreamBooth (Ruiz et al., 2023) combines a unique identifier with a subject class name (*e.g.,* "A S^* cat") to leverage the class-specific prior knowledge embedded in the model, leveraging the model's prior knowledge of the class. Building upon these pioneering works, recent efforts have aimed to improve efficiency, subject fidelity, and instruction alignment, addressing both single-subject generation (Voynov et al., 2023; Alaluf et al., 2023; Tewel et al., 2023; Pang et al., 2024; Cai et al., 2024; Hong et al., 2025) and multi-subject generation (Avrahami et al., 2023; Kumari et al., 2023; Gu et al., 2024; Yao et al., 2024; Zhang et al., 2024f).
- Encoder-based methods utilize a pre-trained image encoder to extract subject-specific features, which are then incorporated into text prompts or directly injected into the generator through dedicated cross-attention mechanisms or adapters. For instance, ELITE (Wei et al., 2023), IP-Adapter (Ye et al., 2023), Subject-Diffusion (Ma et al., 2024a) and SSR-Encoder (Zhang et al., 2024i) employ CLIP (Radford et al., 2021) as the feature extractor, each adopting unique encoding and injection strategies to seamlessly incorporate subject features into the image generation process. MoMA (Song et al., 2025) takes advantage of a pre-trained MLLM, LLaVA (Liu et al., 2023b), to extract subject features and refine them based on the target prompt, producing contextualized image features that are injected into the generator via cross-attention layers. In addition, encoder-based methods have been extended to support multi-subject generation (Patel et al., 2024b; Li et al., 2024d; Zhu et al., 2024; Wang et al., 2024h).

Furthermore, other research has explored diverse techniques for personalized subject-driven generation. These include prompt engineering(He et al., 2024c), which formulates structured prompts to guide generation; instruction tuning(Zhou et al., 2024; Hu et al., 2024a), which fine-tunes models on personalized instructions for improved alignment; and reinforcement learning (Chen et al., 2023d; Chae et al., 2023; Huang et al., 2024b; Wei et al., 2025a), which optimizes models through user feedback to refine subject representation.

Moreover, beyond capturing user preferences for concrete subjects like objects, some studies have focused on more abstract concepts, such as specified relations or styles, to guide personalized T2I generation (Huang et al., 2024e; Sohn et al., 2023; Wang et al., 2023a; Liu et al., 2024c; Park et al., 2024).

A.1.4 Personal Face/Body

Personal face and body images have become popular for personalized image generation, due to their high relevance to individual identity. By leveraging these images, generative models can create highly tailored and realistic images that reflect users' distinct identities (IDs) while adhering to user-specific requirements. Existing studies primarily focus on face generation and virtual try-on applications.

Face Generation Generative models utilize personal face images to create high-fidelity portraits or avatars that preserve individual face IDs while adhering to users' multimodal instructions, such as modifying expressions, actions, backgrounds, and styles. Early GAN-based work mainly encodes face images into the latent space of StyleGAN (Karras et al., 2019) for face manipulation (Xia et al., 2021; Patashnik et al., 2021; Lyu et al., 2023; Baykal et al., 2023). To achieve more precise identity preservation and more flexible control, DM-based methods usually integrate a separate image encoder to convert face images into ID representations, which are then combined with user instructions to guide the face generation process. For instance, FastComposer (Xiao et al., 2024b), Face0 (Valevski et al., 2023), PhotoMaker (Li et al., 2024k), Infinite-ID (Wu et al., 2025a), Master-Weaver (Wei et al., 2025b), and AddMe (Yue et al., 2025) utilize a pre-trained CLIP image encoder or face recognition model to extract ID embeddings for identity preservation. In contrast, SeFi-IDE (Li et al., 2024h) directly optimizes one ID representation as multiple per-stage tokens to enhance semantic control. Except for ID representations, some studies have explored additional features or conditions to enhance style control (Yan et al., 2023), spatial control (Wang et al., 2024f; He et al., 2024d,a; Jiang et al., 2025), especially specific human-object interaction (Guo et al., 2024b; Hu et al., 2025), and scene affordance (Kulal et al., 2023; Parihar et al., 2024).

However, the rapid development of personalized face generation techniques has raised concerns about potential misuse and privacy risks. Recent research has investigated unlearning-related methods (Wu et al., 2025b; Hu et al., 2024b), adversarial attack-based methods (Van Le et al., 2023; Xiao et al., 2023; Wan et al., 2024; Onikubo and Matsui, 2024; Liu et al., 2024e) and watermark-based methods (Liu et al., 2024b) to protect user privacy.

Virtual Try-on This task aims to synthesize a photorealistic image of a dress person by combining their body and face images with specified garments. Early GAN-based works (Wang et al., 2018a; Dong et al., 2019a; Men et al., 2020; Choi et al., 2021; Lee et al., 2022) mainly follow a two-stage process. Initially, a dedicated warping module is employed to align garment images with the person's body shape. Subsequently, the reshaped garment is seamlessly blended with the person's image to generate the final try-on result. With the great success of DMs in various tasks, recent research has deployed their applications in virtual try-ons (Zhang et al., 2024e; Ning et al., 2024; Kim et al., 2024a; Wan et al., 2025). For example, LaDI-VTON (Morelli et al., 2023) and DCI-VTON (Gou et al., 2023) explicitly warp clothes to match the person's body and then utilize DMs for blending. In contrast, TryOnDiffusion (Zhu et al., 2023a) proposes a Parallel-UNet architecture, which performs implicit warping and blending in a unified process. Similarly, several subsequent studies (Choi et al., 2025; Xu et al., 2024d; Sun et al., 2024; Shen et al., 2024a) utilize parallel UNets for garment feature extraction and enhance blending via self-attention and cross-attention mechanisms.

A.2 Personalized Video Generation

Personalized video generation aims to produce tailored video content that reflects individual preferences, traits, and specific needs.

A.2.1 Personalized Subjects

In some cases, users may provide one or several images of a personalized subject, such as an object or concept, along with a specified text prompt, requiring generative models to perform subject-driven text-to-video (T2V) generation. This task is conceptually similar to subject-driven T2I generation.

Subject-driven T2V Generation Given the great success of various personalized models in subject-driven T2I generation, methods such as AnimateDiff (Guo et al., 2024a), PIA (Zhang et al., 2024h), and Still-Moving (Chefer et al., 2024) adapt these models for T2V generation by incorporating motion and temporal dynamics through additional modules. MagDiff (Zhao et al., 2025) enhances subject-driven video generation with three types of alignments. VideoBooth (Jiang et al., 2024) proposes a coarse-to-fine manner to encode subject images into the generator. CustomCrafter (Wu et al., 2024d) integrates a plug-and-play module with a dynamic weighted video sampling strategy to maintain motion generation and conceptual combination abilities during subject learning. Other studies introduce additional conditions, such as motion control (Wu et al., 2024a; Wei et al., 2024b,c) and depth control (He et al., 2023), enabling more flexible subject customization. Besides, some studies have explored multi-subject T2V generation (Chen et al., 2023a, 2024b; Wang et al., 2024l). Beyond these, methods such as StyleCrafter(Liu et al., 2024a) and StyleMaster (Ye et al., 2024) integrate specified style images as subjects, allowing for stylized T2V generation that tailors the video aesthetic to the desired style.

A.2.2 Personal Face/Body

Similarly, users may provide one or more personal face and body images, enabling generative models to synthesize personalized videos that preserve their identities while following multimodal instructions. These tasks include ID-preserving T2V generation, talking head generation, pose-guided video generation, and video virtual try-on.

ID-preserving T2V Generation This task focuses on creating personalized videos that align with personal face IDs and specified text prompts. For example, Magic-Me (Ma et al., 2024c) builds upon Textual Inversion (Gal et al., 2023) to learn ID-specific representations to guide T2V generation, requiring separate training for each ID. ID-Animator (He et al., 2024b) employs a face adapter to encode ID-related information and incorporates it into the generator via cross-attention. ConsisID (Yuan et al., 2024) decomposes facial information into frequency-aware features, which are integrated into Diffusion Transformers (DiT) for video generation. PersonalVideo (Li et al., 2024c) applies direct supervision on T2V-generated videos, aligning model tuning with the inference process.

Talking Head Generation This task aims to synthesize lip-synchronized talking videos, typically driven by personal face images and corresponding audio clips. Recent research has explored diverse approaches, such as Neural Radiance Fields (NeRFs) (Yao et al., 2022; Li et al., 2023b) and different backbone networks, including GANs (Yi et al., 2020; Ki and Min, 2023; Guan et al., 2023) and DMs (Zhang et al., 2023b; Zhua et al., 2023; Shen et al., 2023; Liu et al., 2024f; Wei et al., 2024a; Tian et al., 2025; Tan et al., 2025; Wang et al., 2024c; Zheng et al., 2024b). Beyond audio-driven generation, some studies have also investigated video-driven (Zhang et al., 2023a) and text-driven (Choi et al., 2024) methods for talking head generation.

Pose-guided Video Generation Recent studies have explored adapting personal face and body images to match specific pose sequences through various condition mechanisms for video generation (Wang et al., 2023b; Chang et al., 2023; Karras et al., 2023; Xu et al., 2024f; Hu, 2024; Zhong et al., 2025). Beyond single-person scenarios, Magic-Fight (Huang et al., 2024a) tackles the complexities of two-person martial arts combat video generation, addressing challenges such as identity confusion and action mismatches.

Video Virtual Try-on This task seeks to seamlessly transfer a specified garment onto a person in a source video while preserving their motion and identity. Early GAN-based work (Dong et al., 2019b; Zhong et al., 2021; Jiang et al., 2022a) primarily follows a two-stage workflow, warping the specified garment and blending it with the target person by a GAN generator. Recent studies utilize DMs for video try-ons, incorporating specialized modules for garment and pose encoding, along with dedicated condition mechanisms, such as Tunnel Try-on (Xu et al., 2024e), ACF (Yang et al., 2024f), GPD-VVTO (Wang et al., 2024k), VITON-DiT (Zheng et al., 2024a), ViViD (Fang et al., 2024b), WildVidFit (He et al., 2025), and SwiftTry (Nguyen et al., 2024a).

A.3 Personalized 3D Generation

Personalized 3D generation involves transforming users' personalized visual or textual contexts (e.g., body shapes, facial features, images, and prompts) into 3D assets.

A.3.1 Personalized Subjects

The most common paradigm for personalized 3D generation involves users providing some image-based personalized subjects, and then generating the corresponding 3D assets.

Image-to-3D Generation Personalized image-to-3D generation focuses on creating 3D assets that accurately capture the geometry and appearance of given personalized subjects. 3DAvatarGAN (Abdal et al., 2023) introduces a cross-domain adaptation framework that aligns features from 2D-GANs with those of 3D-GANs. PuzzleAvatar (Xiu et al., 2024) utilizes an enhanced Score Distillation Sampling (SDS) technique to optimize the geometry and texture of 3D portraits. TextureDreamer (Yeh et al., 2024) integrates geometric information using ControlNet, proposing a Geometry-Aware Personalized Score Distillation (PGSD) approach.

Some methods further ensure alignment with textual prompts during the 3D generation process. MVDream (Shi et al., 2023b) employs a multi-view diffusion model to generate consistent multi-view images based on text prompts. DreamBooth3D (Raj et al., 2023) combines DreamFusion and DreamBooth models within a three-stage optimization framework to enhance detail preservation and consistency through multi-view pseudo-data generation. Wonder3D (Long et al., 2024) introduces a cross-domain diffusion model leveraging multi-view cross-attention to produce detailed normal and color maps. DreamFont3D (Li et al., 2024e) utilizes NeRF as the 3D representation, optimizing font geometry and texture with multi-view mask constraints and progressive weight adjustments. Make-your-3D (Liu et al., 2025) implements a joint optimization framework that combines identity-aware and subject-prior optimizations, aligning a 2D personalization model with a multi-view diffusion model for accurate 3D generation.

A.3.2 Personal Face/Body

In some cases, users may provide personal face and body inputs in the form of images or monocular videos, aiming to generate identity-preserving 3D assets.

3D Face Generation For 3D face generation, (Zhang et al., 2021a) introduces PoseGAN, a module designed to generate dynamic head poses. (Gao et al., 2022) proposes a linear blend model based on multi-level voxel fields, representing expressions as neural radiance field bases. My3DGen (Qi et al., 2023a) adapts the pre-trained EG3D model using Low-Rank Adaptation (LoRA) for parameter-efficient training on a limited set of personalized images. DiffusionTalker (Chen et al., 2023c) employs contrastive learning to map speech features to personalized speaker identity. (Wang et al., 2024g) integrates a face fusion module into a fine-tuned text-to-image diffusion model for identity-driven customization. (Ko et al., 2024) utilizes VIVE3D to fine-tune the EG3D generator by inverting key frames from monocular videos. (Song et al., 2024a) applies Cross-Modal Aggregation to blend style and facial motion features, ensuring alignment between generated facial expressions, speech, and styles. DiffSpeaker (Ma et al., 2024d) proposes a diffusion model-based Transformer architecture to enhance speech-to-facial-animation mapping.

3D Human Pose generation For 3D human pose generation, (Huang et al., 2021) combines source image shape information with 2D key points to generate a personalized UV map. PGG (Hu et al., 2023) introduces a geometry-aware graph constructed from intermediate human mesh predictions, enabling personalized and dynamic pose generation. 3DHM (Li et al., 2024a) and DreamWaltz (Huang et al., 2024c) enable animating people from a single image or textual prompts.

3D Virtual Try-on 3D Virtual Try-on enables the creation of high-quality, customized 3D models from minimal inputs, such as user images, clothing images, and textual prompts. (Chu et al., 2017) addresses the precision requirements of personalized facial modeling for applications like eyeglass frame design, utilizing parametric modeling techniques for 3D face creation. Pergamo (Casado-Elvira et al., 2022) addresses the challenge of reconstructing 3D clothing from 2D images, employing semantic segmentation, normal prediction, and a parameterized clothing model to optimize coarse geometry and fine details through differentiable rendering. DreamVTON (Xie et al., 2024) incorporates Multi-Concept LoRA and Normal Style LoRA into Stable Diffusion, enabling the generation of pose-consistent, detail-rich clothing models.

B Evaluation Metrics

B.1 Personalized Text Generation

Evaluating personalized text generation is challenging because only the target user can accurately determine whether the generated content aligns with their preferences and needs. One approach to evaluating personalized text generation is through human judgment, where individuals assess the quality and relevance of the generated content. Automatic evaluation of personalized text generation can be conducted using reference-based and reference-free approaches. In reference-based evaluation, it is assumed that a reference output is available for comparison. This comparison can be performed using term-matching metrics such as accuracy, ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) or semantic-matching methods using models like BERTScore (Zhang et al., 2020b), GEMBA (Kocmi and Federmann, 2023), or G-Eval (Liu et al., 2023c). Recently, ExPerT (Salemi et al., 2025a) was introduced for evaluating personalized text generation in reference-based settings. It segments both the generated text and the reference output into atomic facts and scores them based on content and writing style similarity. In reference-free evaluation, an LLM can assess the generated content directly, assigning scores based on various aspects, including how well the content aligns with the user profile or preferences (Wang et al., 2023g, 2024a), offering a more dynamic and personalized evaluation framework.

B.2 Personalized Audio Generation

To quantify personalization and stylistic alignment in tasks like music transfer and text-to-audio generation, existing studies commonly use similarity metrics such as CLAP scores, Pattern Similarity (PS), and Embedding Distance are commonly applied to assess how well the generated content aligns with target musical styles or speaker characteristics (Sheng et al., 2023; Plitsis et al., 2024).

Audio quality assessment often involves perceptual metrics like Short-Time Objective Intelligibility (STOI) and Extended STOI (ESTOI) for speech intelligibility, Perceptual Evaluation of Speech Quality (PESQ) for clarity, and Fréchet Audio Distance (FAD) for measuring realism and diversity. For music generation tasks, metrics such as Precision, Recall, Density, and Coverage (P&R&D&C) are frequently employed to capture both creativity and output diversity (Wang et al., 2024j; Mo et al., 2023; Hu et al., 2022).

Subjective evaluations remain crucial, particularly for tasks where user experience and personalization are central. User studies are often conducted to measure factors such as perceived musicality, naturalness, emotional impact, and how closely the generated content reflects user preferences (Ma et al., 2022; Dai et al., 2022).

B.3 Personalized Cross-modal Generation

To quantify the degree of personalization in text generation tasks such as personal assistant and comment generation, existing studies commonly employ (1) *term-matching metrics*, such as ROUGE, BLEU, Meteor, CIDEr (Cohen et al., 2022; Alaluf et al., 2025; Nguyen et al., 2024b; Pi et al., 2024; An et al., 2024); (2) *semantic matching metrics*, such as CLIPScore (Cohen et al., 2022); (3) *recall, precision and F1-score* to validate whether the user-specific concept appears in the generated caption (Cohen et al., 2022; Alaluf et al., 2025; Nguyen et al., 2024b; Pi et al., 2024; An et al., 2024); (4) *human evaluation* to determine alignment with ground truths in terms of emotion, style, and relevance.

To measure whether agents align with diverse human preferences, studies employ success rate, success weighted by path length (SPL), distance to goal, and episode length (Poddar et al., 2024; Hwang et al., 2024). Human evaluation is also an effective validation method for determining whether an agent has successfully completed a personalized task (Hwang et al., 2024; Cui et al., 2024b).

B.4 Personalized Image Generation

To assess the alignment between generated images and personalized contexts, as well as adherence to users' multimodal instructions, most studies rely on similarity metrics like Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) and Structural Similarity Index Measure (SSIM) (Wang et al., 2004). Additionally, pre-trained models such as CLIP (Radford et al., 2021), DINO (Oquab et al., 2024), and various face recognition models (Schroff et al., 2015; Deng et al., 2019) are often used to extract image features for computing cosine similarity, enabling a more contextual evaluation of personalization and instruction alignment. Moreover, Stellar (Achlioptas et al., 2023) introduces specialized metrics designed for subject-driven image generation and human generation.

To quantify the quality and coherence of generated images, conventional metrics such as Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) are also commonly employed. Beyond quantitative evaluation, most studies present case studies and conduct human evaluations to assess the personalization and instruction alignment of the generated images. In e-commerce scenarios, product image generation often incorporates online tests to evaluate model performance in real-world scenarios.

B.5 Personalized Video Generation

To assess personalization and instruction alignment, similar to personalized image generation, existing studies commonly rely on similarity metrics such as LPIPS, SSIM, and Peak Signal-to-Noise Ratio (PSNR) (Hore and Ziou, 2010). Additionally, pre-trained image encoders like CLIP and DINO are frequently used to extract frame-level features and compute cosine similarity for quantitative evaluation. There are some task-specific metrics, such as SyncNet score (Casale et al., 2018), which evaluates audio-visual synchronization quality for audio-driven talking head generation, and face similarity metrics for ID-preserving human generation, which are based on backbone face recognition models like ArcFace (Deng et al., 2019) and CurricularFace (Huang et al., 2020).

For overall video quality assessment, standard metrics include frame-level evaluations like FID and KID, as well as video-level metrics such as VFID (Wang et al., 2018c), FID-VID (Balaji et al., 2019), FVD (Unterthiner et al., 2018), and KVD (Unterthiner et al., 2018). Additionally, some studies leverage CLIP-based cosine similarity between consecutive video frames to assess temporal consistency.

Beyond quantitative metrics, many studies complement their evaluations with qualitative case studies and human assessments to better capture personalization, instruction alignment, and overall video quality.

B.6 Personalized 3D Generation

To quantify personalization and instruction alignment in 3D generation, existing studies commonly use similarity metrics such as LPIPS, SSIM, PSNR, and CLIP scores, similar to those in image and video generation. Additionally, some task-specific scores can be evaluated through pre-trained models, such as facial attribute classifiers (Abdal et al., 2023; Qi et al., 2023a).

For 3D geometric quality, commonly used metrics include Chamfer Distance (CD) (Gao et al., 2022; Xiu et al., 2024) and Point-to-Surface Distance (P2S) (Xiu et al., 2024) for shape fidelity, as well as Normal Consistency and Volume IoU for surface detail and volumetric overlap (Xie et al., 2024). Task-specific metrics such as Mean Per Joint Position Error (MPJPE) and Mean Per Vertex Error (MPVE) are frequently used in the human body and pose estimation tasks (Hu et al., 2023).

Beyond objective metrics, qualitative assessments like user studies are often conducted to evaluate subjective aspects of 3D generation (Zhang et al., 2021a; Qi et al., 2023a; Xie et al., 2024; Huang et al., 2021), including realism, texture photorealism, and shape-texture consistency.

C Applications

C.1 Towards Content Creation Process

Generative models have transformed the realm of content creation, pushing the boundaries of productivity and creativity. By incorporating personalization techniques, PGen can further empower content creators across various domains, allowing them to achieve higher efficiency while preserving their distinctive creative style and identity.

For social media influencers, such as bloggers, vloggers, and podcasters, PGen can analyze their past content to offer tailored suggestions for compelling headlines (Fang et al., 2024a) and introductions, or even generate new content that aligns seamlessly with their established brand. This not only streamlines the creative process but also maintains their unique style, fostering deeper connections with their audiences.

For professional creators, such as journalists, designers, illustrators, and music composers, personal style serves as their creative fingerprint, crucial for building reputation and recognition within competitive creative industries. By analyzing creators' previous work, PGen can identify and adapt to their unique stylistic traits to provide tailored ideas, drafts, or modifications, striking a perfect balance between personal style and external demands.

Ordinary individuals can also benefit from PGen for routine tasks, such as personalized email drafting, resume creation, travel planning, workout scheduling, and portrait generation.

C.2 Towards Content Delivery Process

In the era of information overload, personalized content delivery is becoming increasingly essential for helping individuals navigate through vast amounts of multimodal content on the internet. By integrating PGen into the content delivery process, generic content can be adapted into diverse, personalized formats to engage different audiences, catering to their unique tastes and content needs. Below are representative application scenarios of PGen for personalized content delivery:

Marketing and Advertising. PGen can assist organizations and brands in creating targeted marketing strategies and dynamic advertisements that resonate deeply with specific audiences, ultimately driving higher click-through and conversion rates.

Retail and E-commerce. Through personalized product descriptions and images, customized manuals, and virtual try-ons, PGen empowers retailers to attract consumers and enhance engagement, delivering unique and tailored shopping experiences.

Entertainment and Media. On digital content platforms such as Flipboard, Twitter, Netflix, and YouTube, personalized content plays a crucial role in attracting and retaining users. Examples include personalized news, posts, movie posters, video thumbnails, and other tailored media assets that can enhance user loyalty to the platform.

Education and E-learning. Generative models have shown significant promise in education, exemplified by platforms like Google Learn About [1]. PGen can further enhance personalized educational experiences by offering customized learning roadmaps and materials, dynamically adapting to individual learning styles, goals, and progress.

Gaming. Integrating PGen into the gaming industries enables the creation of dynamic storylines, customized tasks, scalable difficulty levels, and interactive characters that adapt to players' preferences and behaviors, fostering more immersive and engaging gaming experiences.

Personalized AI Assistant. PGen can be incorporated into AI assistants to provide specialized support, such as legal assistance, medical advice, and financial guidance, ensuring precision and user-specific customization.

[1] https://learning.google.com/experiments/learn-about.

Table 1: Overview of personalized generation.

Modality	Personalized Contexts	Tasks	Representative Works
Text (Section 3.1)	**User behaviors**	Recommendation	LLM-Rec (Lyu et al., 2024), DEALRec (Lin et al., 2024a), BigRec (Bao et al., 2023), DreamRec (Yang et al., 2024e)
		Information seeking	P-RLHF (Li et al., 2024g), ComPO (Kumar et al., 2024b)
	User documents	Writing Assistant	REST-PG (Salemi et al., 2025b), RSPG (Salemi et al., 2024a), Hydra (Zhuang et al., 2024), PEARL (Mysore et al., 2024), Panza (Nicolicioiu et al., 2024)
	User profiles	Dialogue System	PAED (Zhu et al., 2023b), BoB (Song et al., 2021), UniMS-RAG (Wang et al., 2024d), ORIG (Chen et al., 2023b)
		User Simulation	Drama Machine (Magee et al., 2024), Character-LLM (Shao et al., 2023), RoleLLM (Wang et al., 2024e)
Image (Section A.1)	**User behaviors**	General-purpose generation	PMG (Shen et al., 2024b), Pigeon (Xu et al., 2024c), PASTA (Nabati et al., 2024)
		Fashion design	DiFashion (Xu et al., 2024b), Yu et al. (2019)
		E-commerce product image	AdBooster (Shilova et al., 2023), Vashishtha et al. (2024), Czapp et al. (2024)
	User profiles	Fashion design	LVA-COG (Forouzandehmehr et al., 2023)
		E-commerce product image	CG4CTR (Yang et al., 2024a)
	Personalized subjects	Subject-driven T2I generation	Textual Inversion (Gal et al., 2023), DreamBooth (Ruiz et al., 2023), Custom Diffusion (Kumari et al., 2023)
	Personal face/body	Face generation	PhotoMaker (Li et al., 2024k), InstantBooth (Shi et al., 2024), InstantID (Wang et al., 2024f)
		Virtual try-on	IDM-VTON (Choi et al., 2025), OOTDiffusion (Xu et al., 2024d), OutfitAnyone (Sun et al., 2024)
Video (Section A.2)	**Personalized subjects**	Subject-driven T2V generation	AnimateDiff (Guo et al., 2024a), AnimateLCM (Wang et al., 2024b), PIA (Zhang et al., 2024h)
	Personal face/body	ID-preserving T2V generation	Magic-Me (Ma et al., 2024c), ID-Animator (He et al., 2024b), ConsisID (Yuan et al., 2024)
		Talking head generation	DreamTalk (Ma et al., 2023), EMO (Tian et al., 2025), MEMO (Zheng et al., 2024b)
		Pose-guided video generation	Disco (Wang et al., 2023b), AnimateAnyone (Hu, 2024), MagicAnimate (Xu et al., 2024f)
		Video virtual try-on	ViViD (Fang et al., 2024b), VITON-DiT (Zheng et al., 2024a), WildVidFit (He et al., 2025)
3D (Section A.3)	**Personalized subjects**	Image-to-3D generation	MVDream (Shi et al., 2023b), DreamBooth3D (Raj et al., 2023), Wonder3D (Long et al., 2024)
	Personal face/body	3D face generation	PoseGAN (Zhang et al., 2021a), My3DGen (Qi et al., 2023a), DiffSpeaker (Ma et al., 2024d)
		3D human pose generation	FewShotMotionTransfer (Huang et al., 2021), PGG (Hu et al., 2023), 3DHM (Li et al., 2024a), DreamWaltz (Huang et al., 2024c)
		3D virtual try-on	Pergamo (Casado-Elvira et al., 2022), DreamVTON (Xie et al., 2024)
Audio (Section 3.2)	**Personal face**	Face-to-speech generation	VioceMe (van Rijn et al., 2022), FR-PSS (Wang et al., 2022), Lip2Speech (Sheng et al., 2023)
	User behaviors	Music generation	UMP (Ma et al., 2022), UP-Transformer (Hu et al., 2022), UIGAN (Wang et al., 2024j)
	Personalized subjects	Text-to-audio generation	DiffAVA (Mo et al., 2023), TAS (Li et al., 2024j)
Cross-Modal (Section 3.3)	**User behaviors**	Robotics	VPL (Poddar et al., 2024), Promptable Behaviors (Hwang et al., 2024)
	User documents	Caption/Comment generation	PV-LLM (Lin et al., 2024b), PVCG (Wu et al., 2024e), METER (Geng et al., 2022)
	Personalized subjects	Cross-modal dialogue systems	MyVLM (Alaluf et al., 2025), Yo'LLaVA (Nguyen et al., 2024b), MC-LLaVA (An et al., 2024)

Table 2: Datasets for personalized generation.

Modality	Personalized Contexts	Tasks	Datasets
Text (Section 3.1)	**User behaviors**	Recommendation	Amazon (Hou et al., 2024; Ni et al., 2019), MovieLens (Harper and Konstan, 2015), MIND (Wu et al., 2020a), Goodreads (Wan and McAuley, 2018; Wan et al., 2019)
		Information seeking	SE-PQA (Kasela et al., 2024), PWSC (Eugene et al., 2013), AOL4PS (Guo et al., 2021)
	User documents	Writing Assistant	LaMP (Salemi et al., 2024b), LongLaMP (Kumar et al., 2024a), PLAB (Alhafni et al., 2024)
	User profiles	Dialogue System	LiveChat (Gao et al., 2023), FoCus (Jang et al., 2021), Pchatbot (Qian et al., 2021)
		User Simulation	OpinionsQA (Santurkar et al., 2023), 3 RoleBench (Wang et al., 2024e)
Image (Section A.1)	**User behaviors**	General-purpose generation	Pinterest (Geng et al., 2015), MovieLens (Harper and Konstan, 2015), MIND (Wu et al., 2020b), POG (Chen et al., 2019), PASTA (Nabati et al., 2024), FABRIC (Von Rütte et al., 2023), DialPrompt (Liu et al., 2024d), PIP (Chen et al., 2024g)
		Fashion design	POG (Chen et al., 2019), Polyvore-U (Lu et al., 2019)
		E-commerce product image	-
	User profiles	Fashion design	-
		E-commerce product image	-
	Personalized subjects	Subject-driven T2I generation	Dreambench (Ruiz et al., 2023), Dreambench++ (Peng et al., 2024), CustomConcept101 (Kumari et al., 2023), ConceptBed (Patel et al., 2024a), Textual Inverison (Gal et al., 2023), ViCo (Hao et al., 2023), DreamMatcher (Nam et al., 2024), Break-A-Scene (Avrahami et al., 2023), Mix-of-Show (Gu et al., 2024), Concept Conductor (Yao et al., 2024), LoRA-Composer (Yang et al., 2024d), StyleDrop (Sohn et al., 2023)
	Personal face/body	Face generation	CelebA-HQ (Karras et al., 2018), FFHQ (Karras et al., 2021), SFHQ (Beniaguev, 2022), LV-MHP-v2 (Zhao et al., 2018), Stellar (Achlioptas et al., 2023), AddMe-1.6M (Yue et al., 2025), FFHQ-FastComposer (Xiao et al., 2024a), LAION-Face (Zheng et al., 2022), PPR10K (Liang et al., 2021), LCM-Lookahead (Gal et al., 2024), CelebRef-HQ (Li et al., 2022), CelebV-T (Yu et al., 2023), FaceForensics++ (Rossler et al., 2019), VGGFace2 (Cao et al., 2018)
		Virtual try-on	VITON (Han et al., 2018), VITON-HD (Choi et al., 2021), DressCode (Morelli et al., 2022), StreetTryOn (Cui et al., 2024a), DeepFashion (Ge et al., 2019), Deepfashion-Multimodal (Jiang et al., 2022b), MPV (Dong et al., 2019a), IGPair (Shen et al., 2024a), SHHQ (Fu et al., 2022)
Video (Section A.2)	**Personalized subjects**	Subject-driven T2V generation	WebVid-10M (Bain et al., 2021), UCF101 (Soomro, 2012), AnimateBench (Zhang et al., 2024h), VideoBooth (Jiang et al., 2024), StyleCrafter (Liu et al., 2024a), Datasets for subject-driven T2I generation...
	Personal face/body	ID-preserving T2V generation	ID-Animator (He et al., 2024b), ConsisID (Yuan et al., 2024)
		Talking head generation	LRW (Chung and Zisserman, 2017a), VoxCeleb (Nagrani et al., 2020), VoxCeleb2 (Chung et al., 2018), TCD-TIMIT (Harte and Gillen, 2015), LRS2 (Son Chung et al., 2017), HDTF (Zhang et al., 2021b), MEAD (Wang et al., 2020), GRID (Cooke et al., 2006), MultiTalk (Sung-Bin et al., 2024)
		Pose-guided video generation	FashionVideo (Zablotskaia et al., 2019), TikTok (Jafarian and Park, 2021), TED-talks (Siarohin et al., 2021), Everybody-dance-now (Chan et al., 2019)
		Video virtual try-on	VVT (Dong et al., 2019b), ViViD (Fang et al., 2024b), FashionVideo (Zablotskaia et al., 2019), TikTok (Jafarian and Park, 2021), TikTokDress (Nguyen et al., 2024a)
3D (Section A.3)	**Personalized subjects**	Image-to-3D generation	Dreambench (Ruiz et al., 2023), Objaverse (Deitke et al., 2023)
	Personal face/body	3D face generation	Mystyle (Nitzan et al., 2022), BIWI (Fanelli et al., 2013), VOCASET (Cudeiro et al., 2019)
		3D human pose generation	Human3.6M (Ionescu et al., 2013), 3DPW (Von Marcard et al., 2018), 3DOH50K (Zhang et al., 2020a)
		3D virtual try-on	BUFF (Zhang et al., 2017), DreamVTON (Xie et al., 2024)
Audio (Section 3.2)	**Personal face**	Face-to-speech generation	Voxceleb2 (Chung et al., 2018), LibriTTS (Zen et al., 2019), VGGFace2 (Cao et al., 2018), GRID (Cooke et al., 2006), MultiTalk (Sung-Bin et al., 2024)
	User behaviors	Music generation	Echo (Bertin-Mahieux et al., 2011), MAESTRO (Hawthorne et al., 2019)
	Personalized subjects	Text-to-audio generation	TASBench (Li et al., 2024j), AudioCaps (Kim et al., 2019), AudioLDM (Liu et al., 2023a)
Cross-Modal (Section 3.3)	**User behaviors**	Robotics	D4RL (Fu et al., 2020), Ravens (Zeng et al., 2021), Habitat-Rearrange (Puig et al., 2023), RoboTHOR (Deitke et al., 2020)
	User documents	Caption/Comment generation	TripAdvisor (Geng et al., 2022), Yelp (Geng et al., 2022), PerVidCom (Lin et al., 2024b)
	Personalized subjects	Cross-modal dialogue systems	Yo'LLaVA (Nguyen et al., 2024b), MyVLM (Alaluf et al., 2025)

Table 3: Evaluation metrics for personalized text and image generation.

Text (Section 3.1)	Metrics	1	2	3	4	5	6	Evaluation Dimensions	Representative Works
1. Recommendation **2. Information Seeking** **3. Content Generation** **4. Writing Assistant** **5. Dialogue System** **6. User Simulation**	NDCG (Järvelin and Kekäläinen, 2002)	✓	✓					Overall	BIGRec (Bao et al., 2023), DEALRec (Lin et al., 2024a), AOL4PS (Guo et al., 2021)
	Hit Rate	✓	✓					Overall	BIGRec (Bao et al., 2023)
	Precision	✓	✓					Overall	LLM-Rec (Lyu et al., 2024), AOL4PS (Guo et al., 2021)
	Recall	✓	✓					Overall	LLM-Rec (Lyu et al., 2024), DEALRec (Lin et al., 2024a), AOL4PS (Guo et al., 2021)
	win-rate		✓					Overall	Personalized RLHF (Li et al., 2024g)
	ROUGE (Lin, 2004)			✓	✓	✓	✓	Overall	LaMP (Salemi et al., 2024b), RSPG (Salemi et al., 2024a), Hydra (Zhuang et al., 2024)
	BLEU (Papineni et al., 2002)			✓	✓	✓	✓	Overall	AuthorPred (Li et al., 2023a)
	BERTScore (Zhang et al., 2020b)			✓	✓	✓	✓	Overall	LongLaMP (Kumar et al., 2024a)
	GEMBA (Kocmi and Federmann, 2023)			✓	✓	✓	✓	Overall	REST-PG (Salemi et al., 2025b)
	G-Eval (Liu et al., 2023c)			✓	✓	✓	✓	Overall	REST-PG (Salemi et al., 2025b)
	ExPerT (Salemi et al., 2025a)			✓	✓			Personalization	ExPerT (Salemi et al., 2025a)
	AuPEL (Wang et al., 2023g)			✓	✓			Personalization	AuPEL (Wang et al., 2023g)
	PERSE (Wang et al., 2024a)			✓	✓			Personalization	PERSE (Wang et al., 2024a)

Image (Section A.1)	Metrics	1	2	3	4	5	6	Evaluation Dimensions	Representative Works
1. General-purpose generation **2. Fashion design** **3. E-commerce product image** **4. Subject-driven T2I generation** **5. Face generation** **6. Virtual try-on**	CLIP-I (Radford et al., 2021)	✓	✓		✓	✓	✓	Personalization	Textual Inversion (Gal et al., 2023), Custom Diffuison (Kumari et al., 2023), DreamBooth (Ruiz et al., 2023)
	DINO-I (Caron et al., 2021; Oquab et al., 2024)	✓	✓		✓	✓		Personalization	DreamBooth (Ruiz et al., 2023), BLIP-Diffusion (Li et al., 2024b), ELITE (Wei et al., 2023)
	LPIPS (Zhang et al., 2018)	✓	✓		✓		✓	Personalization	DreamSteerer (Yu et al., 2024), DiFashion (Xu et al., 2024b), PMG (Shen et al., 2024b)
	PSNR (Hore and Ziou, 2010)					✓	✓	Personalization	GroupDiff (Jiang et al., 2025), MYCloth (Liu and Wang, 2024), SCW-VTON (Han et al., 2024)
	SSIM (Wang et al., 2004)	✓	✓		✓	✓	✓	Personalization	DreamSteerer (Yu et al., 2024), PMG (Shen et al., 2024b), OOTDifffusion (Xu et al., 2024d)
	MS-SSIM (Wang et al., 2003)	✓	✓		✓		✓	Personalization	DreamSteerer (Yu et al., 2024), Pigeon (Xu et al., 2024c), SieveNet (Jandial et al., 2020)
	DreamSim (Fu et al., 2023)				✓		✓	Personalization	IMPRINT (Song et al., 2024b), MaX4Zero (Orzech et al., 2024)
	Face similarity (Deng et al., 2019; Schroff et al., 2015; Kim et al., 2022; Wang et al., 2018b)					✓		Personalization	Infinite-ID (Wu et al., 2025a), PhotoMaker (Li et al., 2024k), ProFusion (Zhou et al., 2023)
	Face detection rate (Deng et al., 2019; Zhang et al., 2016)					✓		Personalization	SeFi-IDE (Li et al., 2024h), Celeb Basis (Yuan et al., 2023), W_+ Adapter (Li et al., 2024f)
	CLIP-T (Radford et al., 2021)	✓	✓	✓	✓	✓		Instruction Alignment	Textual Inversion (Gal et al., 2023), Custom Diffuison (Kumari et al., 2023), DreamBooth (Ruiz et al., 2023)
	ImageReward (Xu et al., 2023a)				✓	✓	✓	Instruction Alignment	InstructBooth (Chae et al., 2023), DiffLoRA (Wu et al., 2024f), IMAGDressing-v1 (Shen et al., 2024a)
	PickScore (Kirstain et al., 2023)				✓			Instruction Alignment	InstructBooth (Chae et al., 2023), FABRIC (Von Rütte et al., 2023), Stellar (Achlioptas et al., 2023)
	HPSv1 (Wu et al., 2023b)				✓	✓		Instruction Alignment	Stellar (Achlioptas et al., 2023)
	HPSv2 (Wu et al., 2023b)				✓	✓		Instruction Alignment	Stellar (Achlioptas et al., 2023)
	R-precision (Xu et al., 2018)				✓			Instruction Alignment	COTI (Yang et al., 2023b)
	PAR score (Gani et al., 2024)			✓				Instruction Alignment	Vashishtha et al. (2024)
	FID (Heusel et al., 2017)	✓	✓		✓	✓	✓	Content Quality	COTI (Yang et al., 2023b), IMPRINT (Song et al., 2024b), DiFashion (Xu et al., 2024b)
	KID (Bińkowski et al., 2018)				✓	✓	✓	Content Quality	Custom Diffuison (Kumari et al., 2023), OOTDiffusion (Xu et al., 2024d), LaDI-VTON (Morelli et al., 2023)
	IS (Salimans et al., 2016)				✓		✓	Content Quality	PE-VTON (Zhang et al., 2024d), DF-VTON (Dong et al., 2024), Layout-and-Retouch (Kim et al., 2024b)
	LAION-Aesthetics (Christoph and Romain, 2022)				✓	✓		Content Quality	BLIP-Diffusion (Li et al., 2024b), UniPortrait (He et al., 2024a)
	TOPIQ (Chen et al., 2024a)				✓			Content Quality	DreamSteerer (Yu et al., 2024)
	MUSIQ (Ke et al., 2021)				✓		✓	Content Quality	DreamSteerer (Yu et al., 2024), PE-VTON (Zhang et al., 2024d)
	MANIQA (Yang et al., 2022)						✓	Content Quality	PE-VTON (Zhang et al., 2024d)
	LIQE (Zhang et al., 2023c)				✓			Content Quality	DreamSteerer (Yu et al., 2024)
	QS (Gu et al., 2020)				✓			Content Quality	AddMe (Yue et al., 2025)
	BRISQUE (Mittal et al., 2012a)			✓				Content Quality	Vashishtha et al. (2024)
	CTR			✓				Overall	CG4CTR (Yang et al., 2024a), Czapp et al. (2024)
	Stellar metrics (Achlioptas et al., 2023)				✓	✓		Overall	Stellar (Achlioptas et al., 2023)
	CAMI (Shen et al., 2024a)						✓	Overall	IMAGDressing-v1 (Shen et al., 2024a)

Table 4: Evaluation metrics for personalized generation across video, 3D, audio, and cross-modal domains.

Video (Section A.2)	Metrics	1	2	3	4	5	-	Evaluation Dimensions	Representative Works
1. Subject-driven T2V generation **2. ID-preserving T2V generation** **3. Talking head generation** **4. Pose-guided video generation** **5. Video virtual try-on**	CLIP-I (Radford et al., 2021)	✓	✓		✓			Personalization	PIA (Zhang et al., 2024h), PoseCrafter (Zhong et al., 2025), ID-Animator (He et al., 2024b)
	DINO-I (Caron et al., 2021; Oquab et al., 2024)	✓	✓					Personalization	DisenStudio (Chen et al., 2024b), DreamVideo (Wei et al., 2024b), Magic-Me (Ma et al., 2024c)
	SSIM (Wang et al., 2004)			✓	✓	✓		Personalization	AnimateAnyone (Hu, 2024), VITON-DiT (Zheng et al., 2024a), ViViD (Fang et al., 2024b)
	PSNR (Hore and Ziou, 2010)			✓	✓	✓		Personalization	AnimateAnyone (Hu, 2024), Yi et al. (2020), Zhua et al. (2023)
	LPIPS (Zhang et al., 2018)			✓	✓	✓		Personalization	AnimateAnyone (Hu, 2024), DiffTalk (Shen et al., 2023), DisCo (Wang et al., 2023b)
	VGG (Johnson et al., 2016)				✓			Personalization	DreamPose (Karras et al., 2023)
	L1 error				✓			Personalization	DisCo (Wang et al., 2023b), DreamPose (Karras et al., 2023), MagicAnimate (Xu et al., 2024f)
	AED				✓			Personalization	DisCo (Wang et al., 2023b), DreamPose (Karras et al., 2023)
	Face similarity (Deng et al., 2019; Huang et al., 2020; Kim et al., 2022)		✓	✓	✓			Personalization	ID-Animator (He et al., 2024b), MagicPose (Chang et al., 2023), ConsisID (Yuan et al., 2024)
	CLIP-T (Radford et al., 2021)	✓	✓		✓			Instruction Alignment	PIA (Zhang et al., 2024h), ConsisID (Yuan et al., 2024), PoseCrafter (Zhong et al., 2025)
	UMT score (Liu et al., 2022)	✓						Instruction Alignment	StyleMaster (Ye et al., 2024)
	AKD (Siarohin et al., 2021)				✓			Instruction Alignment	MagicAnimate (Xu et al., 2024f)
	MKR (Siarohin et al., 2021)				✓			Instruction Alignment	MagicAnimate (Xu et al., 2024f)
	MSE-P				✓			Instruction Alignment	PoseCrafter (Zhong et al., 2025)
	SyncNet score (Chung and Zisserman, 2017b)			✓				Instruction Alignment	DreamTalk (Ma et al., 2023), EMO (Tian et al., 2025), MEMO (Zheng et al., 2024b)
	LMD (Chen et al., 2018)			✓				Instruction Alignment	DFA-NeRF (Yao et al., 2022), DreamTalk (Ma et al., 2023), Yi et al. (2020)
	LSE-C (Prajwal et al., 2020)			✓				Instruction Alignment	StyleLipSync (Ki and Min, 2023), DiffTalker (Qi et al., 2023b), Choi et al. (2024)
	LSE-D (Prajwal et al., 2020)			✓				Instruction Alignment	StyleLipSync (Ki and Min, 2023), DiffTalker (Qi et al., 2023b), Choi et al. (2024)
	PD (Baldrati et al., 2023)					✓		Instruction Alignment	ACF (Yang et al., 2024f)
	FID (Heusel et al., 2017)		✓	✓	✓	✓		Content Quality	EMO (Tian et al., 2025), ConsisID (Yuan et al., 2024), DisCo (Wang et al., 2023b)
	KID (Bińkowski et al., 2018)					✓		Content Quality	WildVidFit (He et al., 2025)
	ArtFID (Wright and Ommer, 2022)	✓						Content Quality	StyleMaster (Ye et al., 2024)
	VFID (Wang et al., 2018c)					✓		Content Quality	SwiftTry (Nguyen et al., 2024a), VITON-DiT (Zheng et al., 2024a), ViViD (Fang et al., 2024b)
	FVD (Unterthiner et al., 2018)	✓	✓	✓	✓			Content Quality	PersonalVideo (Li et al., 2024c), MotionBooth (Wu et al., 2024a), AnimateAnyone (Hu, 2024)
	FID-VID (Balaji et al., 2019)				✓			Content Quality	DisCo (Wang et al., 2023b), MagicAnimate (Xu et al., 2024f), MagicPose (Chang et al., 2023)
	KVD (Unterthiner et al., 2018)	✓						Content Quality	Animate-A-Story (He et al., 2023)
	E-FID (Tian et al., 2025)			✓				Content Quality	EMO (Tian et al., 2025), EmotiveTalk (Wang et al., 2024c)
	NIQE (Mittal et al., 2012b)				✓			Content Quality	MagicFight (Huang et al., 2024a)
	CPBD (Narvekar and Karam, 2011)			✓				Content Quality	DreamTalk (Ma et al., 2023)
	Temporal consistency (Radford et al., 2021)	✓	✓					Content Quality	AnimateDiff (Guo et al., 2024a), Magic-Me (Ma et al., 2024c), DreamVideo (Wei et al., 2024b)
	Dynamic degree (Huang et al., 2024d)	✓	✓					Content Quality	StyleMaster (Ye et al., 2024), ID-Animator (He et al., 2024b), PersonalVideo (Li et al., 2024c)
	Video IS (Saito et al., 2020)	✓				✓		Content Quality	MagDiff (Zhao et al., 2025)
	Flow error (Shi et al., 2023a)	✓						Content Quality	MotionBooth (Wu et al., 2024a)
	Stitch score	✓						Content Quality	VideoDreamer (Chen et al., 2023a)
	Dover score (Wu et al., 2023a)		✓					Content Quality	ID-Animator (He et al., 2024b)
	Motion score (Li et al., 2018)		✓					Content Quality	ID-Animator (He et al., 2024b)
3D (Section A.3)	**Metrics**	**1**	**2**	**3**	**4**	**-**	**-**	**Evaluation Dimensions**	**Representative Works**
1. Image-to-3D generation **2. 3D face generation** **3. 3D human pose generation** **4. 3D virtual try-on**	LPIPS (Zhang et al., 2018)	✓	✓	✓				Personalization	Wonder3D (Long et al., 2024), PuzzleAvatar (Xiu et al., 2024), My3DGen (Qi et al., 2023a)
	PSNR (Hore and Ziou, 2010)	✓	✓	✓				Personalization	Wonder3D (Long et al., 2024), PuzzleAvatar (Xiu et al., 2024), My3DGen (Qi et al., 2023a)
	SSIM (Wang et al., 2004)	✓	✓	✓				Personalization	Wonder3D (Long et al., 2024), PuzzleAvatar (Xiu et al., 2024), My3DGen (Qi et al., 2023a)
	Chamfer Distances (Butt and Maragos, 1998),	✓						Personalization	Wonder3D (Long et al., 2024), PuzzleAvatar (Xiu et al., 2024)
	CLIP (Radford et al., 2021)				✓			Personalization	DreamVTON (Xie et al., 2024)
	Volume IoU (Zhou et al., 2019),	✓						Personalization	Wonder3D (Long et al., 2024)
	Lip Vertex Error (LVE) (Ma et al., 2024d)		✓					Personalization	DiffSpeaker (Ma et al., 2024d)
	Facial Dynamics Deviation (FDD) (Ma et al., 2024d)		✓					Personalization	DiffSpeaker (Ma et al., 2024d), DiffsuionTalker (Chen et al., 2023c)
	FReID (Huang et al., 2021)			✓				Personalization	FewShotMotionTransfer (Huang et al., 2021)
	CLIP-T (Radford et al., 2021)	✓						Instruction Alignment	MVDream (Shi et al., 2023b), DreamBooth3D (Raj et al., 2023), MakeYour3D (Liu et al., 2025)
	FID (Heusel et al., 2017)	✓			✓			Content Quality	MVDream (Shi et al., 2023b), 3DAvatarGAN (Abdal et al., 2023), TextureDreamer (Yeh et al., 2024), DreamV-TON (Xie et al., 2024)
	IS (Salimans et al., 2016)	✓						Content Quality	MVDream (Shi et al., 2023b)
Audio (Section 3.2)	**Metrics**	**1**	**2**	**3**	**-**	**-**	**-**	**Evaluation Dimensions**	**Representative Works**
1. Face-to-speech generation **2. Music generation** **3. Text-to-audio generation**	CLAP (Elizalde et al., 2023)			✓				Personalization	DB&TI (Plitsis et al., 2024)
	Embedding Distance	✓	✓					Personalization	UMP (Ma et al., 2022), FR-PSS (Wang et al., 2022)
	FAD (Kilgour et al., 2018)	✓	✓	✓				Personalization	UIGAN (Wang et al., 2024j), DiffAVA (Mo et al., 2023), DB&TI (Plitsis et al., 2024)
	IS (Salimans et al., 2016)			✓				Content Quality	DiffAVA (Mo et al., 2023)
	STOI, ESTOI, PESQ (Sheng et al., 2023)	✓						Content Quality	Lip2Speech (Sheng et al., 2023)
Cross-modal (Section 3.3)	**Metrics**	**1**	**2**	**3**	**-**	**-**	**-**	**Evaluation Dimensions**	**Representative Works**
1. Robotics **2. Caption/Comment generation** **3. Multimodal dialogue systems**	BLEU,Meteor		✓	✓				Overall	PVCG (Wu et al., 2024e), METER (Geng et al., 2022), PV-LLM (Lin et al., 2024b)
	Recall, Precision, F1			✓				Overall	MyVLM (Alaluf et al., 2025), Yo'LLaVA (Nguyen et al., 2024b)
	success rate	✓						Overall	VPL (Poddar et al., 2024), Promptable Behaviors (Hwang et al., 2024)