Beyond Single-Concept Personalization: A Systematic Survey of Multi-Concept Methods in Text-to-Image Diffusion Models

Muhammad Rifqi Fauzi[1]★, Gembong Edhi Setyawan[1], and Novanto Yudistira[1]

Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia
{mrifqifauzi@student.ub.ac.id, gembong@ub.ac.id, yudistira@ub.ac.id}

Abstract. Text-to-image (T2I) generative diffusion models have made remarkable progress in single-concept personalization. However, extending customization to multiple concepts simultaneously introduces critical challenges, including identity bleeding (where visual features of distinct subjects merge) and attribute leakage (where properties bind to incorrect subjects). These failures stem from global cross-attention mechanisms that lack spatial constraints. Despite the rapid proliferation of methods addressing multi-concept personalization, a comprehensive survey that systematically categorizes these approaches is still missing in the literature. This paper presents a systematic survey of the state-of-the-art in multi-concept personalization. We propose a unified taxonomy that classifies existing methodologies along four primary dimensions: (1) personalization paradigm and concept representation space (optimization-based, tuning-free, and hybrid), (2) disentanglement and generation strategy (weight merging, attention routing, layout control, and score fusion), (3) target personalization granularity (full-object, part-level, relation-aware, and identity-specific), and (4) generative backbone architecture (U-Net, Diffusion Transformer, and Autoregressive models). Furthermore, we systematically examine how these methods integrate with modern backbones and granularities, highlighting key trade-offs between parameter tuning efficiency and zero-shot generalization. We analyze standard evaluation protocols, covering benchmark datasets and quantitative metrics for subject fidelity, text alignment, and spatial composition. Finally, we identify key open challenges such as strict spatial isolation, hierarchical subject-object binding, and zero-shot scaling, offering a clear, structured roadmap for future research in this rapidly evolving field.

Keywords: Multi-concept personalization · Text-to-image synthesis · Diffusion models · Concept disentanglement · Identity preservation

1 Introduction

Text-to-Image (T2I) diffusion models [28] have revolutionized visual content generation by synthesizing high-fidelity images conditioned on natural language prompts. However, generic models cannot generate specific, user-provided instances (e.g., a specific

★ Corresponding author.

© The Author(s) 2026
M. Murtadha et al. (eds.), *Proceedings of the 2026 11th International Conference on Sustainable Information Engineering and Technology (SIET 2026)*, Atlantis Highlights in Computer Sciences 26,
https://doi.org/10.2991/978-94-6239-828-3_2

personal pet) because natural language lacks the bandwidth to express fine-grained individual details. Single-subject personalization methods were developed to address this limitation. Textual Inversion [8] learns a new pseudo-token v^* in the embedding space of the text encoder, DreamBooth [30] fine-tunes the entire generative network, and Low-Rank Adaptation (LoRA) [12] updates low-rank residual attention projection matrices for parameter efficiency.

In practical applications, users frequently need to compose multiple distinct personalized subjects within a single scene. However, naively applying single-subject methods to multi-concept settings leads to severe compositional failures: (1) *identity bleeding*, where visual features of distinct subjects are mixed or swapped; (2) *attribute leakage*, where descriptive modifiers bind to incorrect subjects; and (3) *spatial conflict*, which causes scale anomalies, unnatural overlaps, or complete subject omission. These failures stem directly from unconstrained global cross-attention mechanisms, where image query tokens interact with text conditioning keys across the entire spatial canvas without boundary isolation.

To address these limitations, a substantial body of literature has emerged across optimization-based, tuning-free, and hybrid approaches. To organize this rapidly expanding field, we present a unified taxonomy structured along four primary dimensions: (1) *Personalization Paradigm and Concept Representation Space*, (2) *Disentanglement and Generation Strategy*, (3) *Target Personalization Granularity*, and (4) *Generative Backbone Architecture* (§3), as illustrated in Fig. 1.

To guide our systematic analysis, this survey is structured around three core Research Questions (RQs):

- **RQ1 (Paradigms & Representation):** What are the dominant personalization paradigms and concept representation spaces used to encode multi-subject identities, and how do they balance parameter efficiency with visual fidelity?
- **RQ2 (Disentanglement Strategies):** How do modern disentanglement strategies (weight merging, attention routing, layout control, score fusion) isolate multi-concept features to prevent identity bleeding and attribute leakage?
- **RQ3 (Granularities, Backbones & Bottlenecks):** How do generative backbones (U-Net, DiT, Autoregressive) and target granularities impact spatial isolation, zero-shot scaling, and benchmark standardization?

1.1 Systematic Search Protocol

To ensure complete and reproducible coverage, we conducted a systematic literature search following standard survey protocols. We queried major electronic databases and digital libraries, including arXiv, IEEE Xplore, ACM Digital Library, Google Scholar, CVPR, ICCV, ECCV, NeurIPS, and AAAI, covering literature published or posted between 2023 and early 2025. Search query strings combined primary keywords: *"multi-concept personalization"*, *"multi-subject diffusion"*, *"concept disentanglement"*, *"tuning-free personalization"*, and *"multi-LoRA composition"*.

We applied predefined inclusion and exclusion criteria: included works must specifically address multi-concept or multi-subject image generation using diffusion or transformer-based backbones and provide empirical evaluations (quantitative or qualitative

metrics). We excluded single-subject-only methods, non-diffusion text-to-image approaches, and non-peer-reviewed blog posts lacking technical rigor. From an initial pool of 154 identified candidate records, 60 primary papers were selected for detailed analysis and taxonomical synthesis in this survey.

2 Preliminaries and Problem Definition

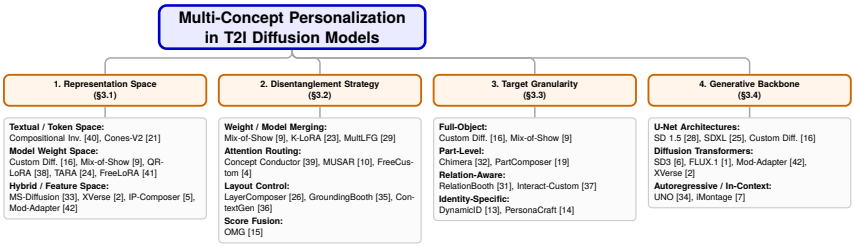


Fig. 1. Unified 4-dimensional taxonomy of multi-concept personalization methods in text-to-image diffusion models

LDMs [28] operate in a lower-dimensional latent space via an autoencoder with encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Input I is mapped to latents $z_0 = \mathcal{E}(I)$. The model is trained to denoise latent representations conditioned on text prompts.

Single-subject personalization associates an identity with a pseudo-token v^*, given reference images of a category C. Textual Inversion [8] learns a new token embedding while keeping model weights frozen. DreamBooth [30] fine-tunes model parameters to bind identities to textual concepts. LoRA [12] injects trainable low-rank adapters into attention heads.

Multi-concept personalization extends the task to N different subjects $\{S_i\}_{i=1}^{N}$ with tokens $\{v_i^*\}_{i=1}^{N}$ and a composite prompt y_{joint}. Because global cross-attention mechanisms lack spatial boundaries, naive compositions lead to three major failure modes (Fig. 2): **Identity Bleeding**, where visual features of distinct subjects merge; **Attribute Leakage**, where descriptive attributes bind to incorrect subjects; and **Spatial Conflicts**, where concepts fail to activate, resulting in scale distortion or subject omission.

3 Taxonomy of Multi-Concept Personalization

Our taxonomy organizes multi-concept personalization literature along four primary dimensions: (1) Personalization Paradigm and Concept Representation Space, (2) Disentanglement and Generation Strategy, (3) Target Personalization Granularity, and (4) Generative Backbone Architecture.

3.1 Personalization Paradigms and Concept Representation Spaces

The concept representation space dictates where and how custom identities are encoded and injected into the generative architecture.

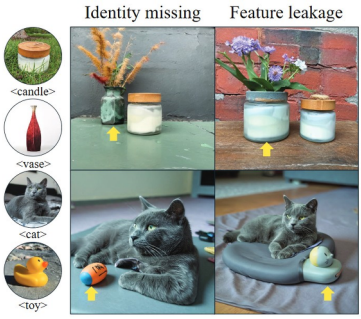


Fig. 2. Compositional failure modes (adapted from [24]): (Left) Identity missing; (Right) Feature leakage/identity bleeding

Textual and Token-Space Methods. These methods represent custom concepts as unique pseudo-tokens v_i^* in the word embedding lookup table of a frozen text encoder. To stabilize multi-concept optimization, Cones-V2 [21] decomposes embeddings into category-level base embeddings and concept-specific residual updates ($v_i^* = v_{\text{base}} + \Delta v_i$). Compositional Inversion [40] optimizes pseudo-words while regularizing joint feature representations. While token-space methods preserve base model parameters, their visual fidelity is constrained by the expressiveness of the text encoder lookup table.

Model Weight-Space Methods. To capture intricate visual details, weight-space methods fine-tune network parameters or low-rank adapters. Custom Diffusion [16] updates key-value projection matrices in cross-attention while freezing query projections. Mix-of-Show [9] trains single-concept LoRAs independently and merges them at inference. QR-LoRA [38] factorizes low-rank updates using QR decomposition ($\Delta W_i^* = \Delta W_i(I - P_{\text{shared}})$) to enforce subspace orthogonality and eliminate linear merging interference. TARA [24] proposes token-aware LoRA projection, while FreeLoRA [41] and Cached Multi-LoRA [43] dynamically scale and mask attention outputs at inference.

Hybrid and Feature-Space Methods. Hybrid approaches combine textual tokens with visual features extracted from pre-trained vision encoders (e.g., CLIP, DINOv2). Decoupled cross-attention layers (e.g., IP-Composer [5]) process text and visual prompts independently to prevent text override. In Diffusion Transformers, modulation layers scale and shift block-level latents. Mod-Adapter [42] predicts personalized scale-shift parameters directly from multimodal prompts. MS-Diffusion [33] and XVerse [2] leverage visual prompts to enable zero-shot multi-subject composition.

3.2 Disentanglement and Generation Strategies

This axis classifies methodologies by the computational strategy used to isolate multiple concepts during generation and prevent identity bleeding.

Model and Weight Merging. Model merging combines independently trained concept modules (e.g., subject-specific LoRAs) at inference without joint retraining. Linear weight addition ($W_{\text{merged}} = W_0 + \sum_{i=1}^{N} \lambda_i \Delta W_i$) frequently causes parameter conflicts

and feature distortion. Mix-of-Show [9] resolves conflicts via gradient-guided weight fusion. K-LoRA [23] provides a training-free fusion mechanism that dynamically selects Top-K parameter activations from subject and style LoRAs at attention time using timestep-dependent scaling factors. MultLFG [29] filters out conflicting frequency bands. FreeLoRA [41] and Cached Multi-LoRA [43] apply subject-aware token routing during sampling.

Attention Routing and Modulation. Attention modulation restricts textual conditioning to specific spatial coordinates during inference. In cross-attention, query activations Q and key projections K_i of tokens v_i^* are modulated by spatial masks M_i:

$$A_i^*(x) = \frac{\exp(Q(x)K_i^T/\sqrt{d}) \cdot M_i(x)}{\sum_j \exp(Q(x)K_j^T/\sqrt{d}) \cdot M_j(x)}, \tag{1}$$

where $A_i^*(x)$ is the modulated cross-attention weight at pixel location x. By addressing identity bleeding directly in attention maps, training-free routing (e.g., Concept Conductor [39], MUSAR [10], FreeCustom [4]) scales dynamically to multiple concepts and integrates seamlessly with pre-trained single-concept LoRAs.

Layout and Spatial Constraints. Layout control methods (e.g., LayerComposer [26], GroundingBooth [35], ContextGen [36], Create Anything Anywhere [17]) utilize explicit inputs such as bounding boxes or layered canvases, steering diffusion latents via loss gradients to enforce spatial boundaries.

Score and Noise-Level Fusion. Score fusion approaches (e.g., OMG [15]) merge multiple concepts during sampling by blending noise predictions from single-concept models within spatial masks.

3.3 Target Personalization Granularity

The complexity of multi-concept synthesis depends heavily on the target granularity of the customized elements.

Full-Object Personalization. The standard setup generates distinct whole subjects (e.g., a specific dog and a specific cat) composed in a single scene, focusing on identity preservation across separate entities.

Part-Level Personalization. Part-level methods extract components from distinct source images to assemble a novel composite object (e.g., a specific toy character wearing armor from another image). Chimera [32] and PartComposer [19] achieve this via localized spatial cross-attention routing that binds part key vectors to sub-regions of the same entity.

Relation-Aware Personalization. Relation-aware customization synthesizes subjects engaged in complex physical interactions or semantic relationships where spatial bounding boxes overlap. RelationBooth [31] and Interact-Custom [37] employ localized self-attention gating and interaction tokens to model physical contact and spatial dependencies.

Identity-Specific Personalization. For human-centric generation, face identity preservation is paramount. DynamicID [13] and PersonaCraft [14] integrate dedicated face encoders to inject identity embeddings into key-value projection layers via decoupled cross-attention.

3.4 Generative Backbone Architectures and Dynamics

The underlying generative architecture fundamental governs multi-concept composition dynamics.

U-Net Architectures. U-Net backbones (e.g., Stable Diffusion 1.5 [28], SDXL [25]) offer parameter efficiency but suffer from multi-resolution boundary mismatches caused by factorized cross-attention layers across downsampling blocks.

Diffusion Transformers (DiTs). DiT models (e.g., SD3 [6], FLUX.1 [1]) enhance prompt alignment via joint self-attention over text and image patch tokens. However, this global token interaction amplifies identity leakage risks, requiring specialized modulation adapters (e.g., Mod-Adapter [42], XVerse [2]).

Autoregressive Models. Autoregressive paradigms (e.g., UNO [34], iMontage [7]) frame personalization as visual in-context next-token prediction, achieving zero-shot scalability at the expense of fine-grained detail loss due to vision encoder tokenization limits.

3.5 Comparative Analysis and Method Trade-Offs

To provide actionable insights beyond individual paper descriptions, we compare the relative strengths, weaknesses, and operational trade-offs of key method groups:

Tuning-Free vs. Optimization-Based Paradigms. Optimization-based methods (e.g., DreamBooth, Custom Diffusion) achieve superior identity fidelity and fine-grained attribute matching, but require high computational overhead during inference-time training (5–15 minutes per concept pair) and are prone to overfitting. Conversely, *tuning-free methods* (e.g., MS-Diffusion, XVerse) enable instantaneous zero-shot generation and interactive workflow integration, but are constrained by the embedding capacity of pretrained vision encoders, occasionally resulting in identity smoothing under complex prompts.

Attention Routing vs. Layout Control Strategies. Attention routing mechanisms (e.g., Concept Conductor, MUSAR) operate dynamically without requiring user layout annotations, making them ideal for unconstrained text prompts. However, they remain susceptible to soft-boundary leakage when subjects are spatially adjacent. *Layout-controlled approaches* (e.g., LayerComposer, ContextGen) guarantee strict spatial isolation and exact bounding box placement, but depend on manual or automated layout inputs and struggle when handling overlapping physical interactions.

Table 1 summarizes representative multi-concept methods structured across all four taxonomy dimensions.

Table 1. Taxonomical comparison of representative multi-concept personalization methods across all four proposed dimensions

Method	Paradigm (Representation)	Disentanglement Strategy	Target Granularity	Generative Backbone
Textual Inversion [8]	Optimization (Token)	Concept Fusion	Full-Object	U-Net
DreamBooth [30]	Optimization (Weight)	Concept Fusion	Full-Object	U-Net
Custom Diffusion [16]	Optimization (Weight)	Model Merging	Full-Object	U-Net
Mix-of-Show [9]	Optimization (Weight)	Model Merging	Full-Object	U-Net
Cones-V2 [21]	Optimization (Token)	Spatial Control	Full-Object	U-Net
OMG [15]	Tuning-Free (Feature)	Score Fusion	Full-Object	U-Net
Concept Conductor [39]	Tuning-Free (Token)	Attention Routing	Full-Object	U-Net
MS-Diffusion [33]	Tuning-Free (Hybrid)	Attention Routing	Full-Object	U-Net/DiT
Mod-Adapter [42]	Tuning-Free (Hybrid)	Attention Routing	Full-Object	DiT
XVerse [2]	Tuning-Free (Hybrid)	Attention Routing	Full-Object	DiT
iMontage [7]	Tuning-Free (Hybrid)	Attention Routing	Full-Object	DiT (AR)
Chimera [32]	Hybrid (Feature)	Attention Routing	Part-Level	U-Net
PartComposer [19]	Hybrid (Feature)	Attention Routing	Part-Level	U-Net
RelationBooth [31]	Optimization (Weight)	Attention Routing	Relation-Aware	U-Net
Interact-Custom [37]	Tuning-Free (Hybrid)	Attention Routing	Relation-Aware	U-Net

3.6 Evaluation Protocols

Evaluating multi-concept personalization requires assessing identity preservation, text alignment, and spatial composition simultaneously.

Benchmark Datasets. Multi-concept evaluation relies on three standard benchmarks: (1) DreamBench (Adapted) [30] (30 concept pairs in composite scenes); (2) CustomConcept101 [16] (101 concept pairs covering category and style variations); and (3) MUSAR-Gen [10] (60+ prompts with layout annotations).

Quantitative Metrics. Evaluation protocols employ distinct metric suites: (1) subject fidelity uses CLIP-I [27] for global visual similarity, DINOv2 [22] for fine structure/texture similarity, and ArcFace/InsightFace [3] for facial identity preservation; (2) text editability relies on CLIP-T [27] prompt alignment scores; (3) compositional correctness applies Grounding DINO IoU [20] for bounding box precision and VLM evaluation (LLaVA) [18] for attribute binding verification; and (4) image quality is measured via FID [11].

4 Open Challenges and Conclusion

Despite significant progress, multi-concept personalization faces four key open challenges. First, tuning-free layout models (e.g., MS-Diffusion [33]) suffer from identity smoothing under dense text conditions, while autoregressive frameworks (e.g., UNO [34]) omit fine details due to encoder compression. Bridging the fidelity gap between tuning-free and optimization-based models remains an open challenge. Second, soft attention masking still allows identity bleeding in adjacent spatial regions; guaranteeing strict spatial isolation requires mathematical bounds such as asymptotically negative attention penalties in self-attention layers. Third, attribute binding degrades sharply beyond 3–5 concepts, requiring hierarchical token-to-region routing architectures. Fourth, transitioning from U-Nets to Diffusion Transformers necessitates re-architecting cross-attention modulation layers for unified MMDiT blocks, compounded by a lack of standardized multi-subject benchmarks.

In conclusion, this systematic survey reviewed the multi-concept personalization landscape in text-to-image models, introducing a unified taxonomy across four dimensions: personalization paradigm, disentanglement strategy, target granularity, and generative backbone. While the field is rapidly adopting tuning-free and DiT-based architectures, resolving strict spatial isolation, multi-subject scaling, and standardized evaluation remains the critical frontier. We hope this survey provides a clear roadmap for future research in personalized image synthesis.

Acknowledgments. This research was supported by the Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Black Forest Labs, Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
2. Chen, B., Sun, H., Chen, L., Wang, X., Du, D., Wu, X., et al.: Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. Advances in Neural Information Processing Systems **37**, 7437–7459 (2024)
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
4. Ding, G., Zhao, C., Wang, W., Yang, Z., Liu, Z., Chen, H., Shen, C.: Freecustom: Tuning-free customized image generation for multi-concept composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9089–9098 (2024)
5. Dorfman, S., Cohen-Bar, D., Gal, R., Cohen-Or, D.: Ip-composer: Semantic composition of visual concepts. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning. vol. 235, pp. 12606–12633. PMLR (2024)
7. Fu, Z., Zeng, X., et al.: imontage: Unified, versatile, highly dynamic many-to-many image generation. arXiv preprint arXiv:2511.20635 (2025)
8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
9. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. Advances in Neural Information Processing Systems **36**, 15890–15902 (2023)

10. Guo, Z., Zhang, P., Wu, Y., Mou, C., Zhao, S., He, Q.: Musar: Exploring multi-subject customization from single-subject dataset via attention routing. arXiv preprint arXiv:2505.02823 (2025)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6626–6637 (2017)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
13. Hu, X., Wang, J., Chen, H., Zhang, W., Wang, B., Li, Y., Nan, H.: Dynamicid: Zero-shot multi-id image personalization with flexible facial editability. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10549–10559 (2025)
14. Kim, G., Jeon, S.Y., Lee, S., Chun, S.Y.: Personacraft: Personalized and controllable full-body multi-human scene generation using occlusion-aware 3d-conditioned diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12034–12044 (2025)
15. Kong, Z., Zhang, Y., Yang, T., Wang, T., Zhang, K., Wu, B., Chen, G., Liu, W., Luo, W.: Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In: European Conference on Computer Vision. pp. 253–270. Springer (2024)
16. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
17. Li, W., Li, H., Peng, Y., Wu, S., Zhang, Y., Sun, X.: Create anything anywhere: Layout-controllable personalized diffusion model for multiple subjects. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)
18. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (2024)
19. Liu, J., Jones, R.K., Ritchie, D.: Partcomposer: Learning and composing part-level concepts from single-image examples. pp. 1–11. Association for Computing Machinery (ACM) (12 2025). https://doi.org/10.1145/3757377.3763947
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, L., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with transformers for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
21. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Customizable image synthesis with multiple subjects. Advances in neural information processing systems **36**, 57500–57519 (2023)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Vasiljevic, V., Saliou, P., Sandmeier, A., Dupont, S., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
23. Ouyang, Z., Li, Z., Hou, Q.: K-lora: Unlocking training-free fusion of any subject and style loras. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13041–13050. IEEE (2025)
24. Peng, Y., Zheng, L., Yang, Y., Huang, Y., Yan, M., Liu, J., Chen, S.: Tara: Token-aware lora for composable personalization in diffusion models (2025)
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Doherty, P.L., Hartmann, A., Hilsenstein, V., Deuel, D., Levi, Y., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
26. Qian, G.G., Zhang, R., Chen, T.S., Dalva, Y., Goyal, A.A., Menapace, W., Skorokhodov, I., Dong, M., Sahni, A., Ostashev, D., Hu, J., Tulyakov, S., Wang, K.C.J.: Layercomposer: Multi-human personalized generation via layered canvas (11 2025), http://arxiv.org/abs/2510.20820

27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
29. Roy, A., Suin, M., Shah, K., Chellappa, R.: Multlfg: Training-free multi-lora composition using frequency-domain guidance. arXiv preprint arXiv:2505.20525 (2025)
30. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
31. Shi, Q., Qi, L., Wu, J., Bai, J., Wang, J., Tong, Y., Li, X.: Dreamrelation: Bridging customization and relation generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15723–15732 (2025)
32. Singh, S., Chen, Y., Chatterjee, A., Raj, A., Hays, J., Yang, Y., Baral, C.: Chimera: Compositional image generation using part-based concepting (10 2025), http://arxiv.org/abs/2510.18083
33. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In: International Conference on Learning Representations. vol. 2025, pp. 95118–95146 (2025)
34. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18682–18692 (2025)
35. Xiong, Z., Xiong, W., Shi, J., Zhang, H., Song, Y., Jacobs, N.: Groundingbooth: Grounding text-to-image customization. arXiv preprint arXiv:2409.08520 (2024)
36. Xu, R., Zhou, D., Ma, F., Yang, Y.: Contextgen: Contextual layout anchoring for identity-consistent multi-instance generation (2 2025), http://arxiv.org/abs/2510.11000
37. Xu, Z., Wang, Z., Peng, Y., Liu, Y.: Interact-custom: Customized human object interaction image generation. In: MM 2025 - Proceedings of the 33rd ACM International Conference on Multimedia, Co-Located with MM 2025. pp. 9500–9508. Association for Computing Machinery, Inc (10 2025). https://doi.org/10.1145/3746027.3754911
38. Yang, J., Ma, Y., Di, D., Cui, J., Li, H., Chen, W., Xie, Y., Yang, X., Zuo, W.: Qr-lora: Efficient and disentangled fine-tuning via qr decomposition for customized generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17587–17597 (2025)
39. Yao, Z., Feng, F., Li, R., Wang, X.: Concept conductor: Orchestrating multiple personalized concepts in text-to-image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9427–9435 (2025)
40. Zhang, X., Wei, X.Y., Wu, J., Zhang, T., Zhang, Z., Lei, Z., Li, Q.: Compositional inversion for stable diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7350–7358 (2024)
41. Zheng, P., Wang, Y., Ma, R., Wu, Z.: Freelora: Enabling training-free lora fusion for autoregressive multi-subject personalization (7 2025), http://arxiv.org/abs/2507.01792
42. Zhong, W., Yang, H., Liu, Z., He, H., He, Z., Niu, X., Zhang, D., Li, G.: Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter (2 2025), http://arxiv.org/abs/2505.18612
43. Zou, X., Shen, M., Bouganis, C.S., Zhao, Y.: Cached multi-lora composition for multi-concept image generation. arXiv preprint arXiv:2502.04923 (2025)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

