Match-and-Fuse: Consistent Generation from Unstructured Image Sets

Kate Feingold Omri Kaduri Tali Dekel

Weizmann Institute of Science

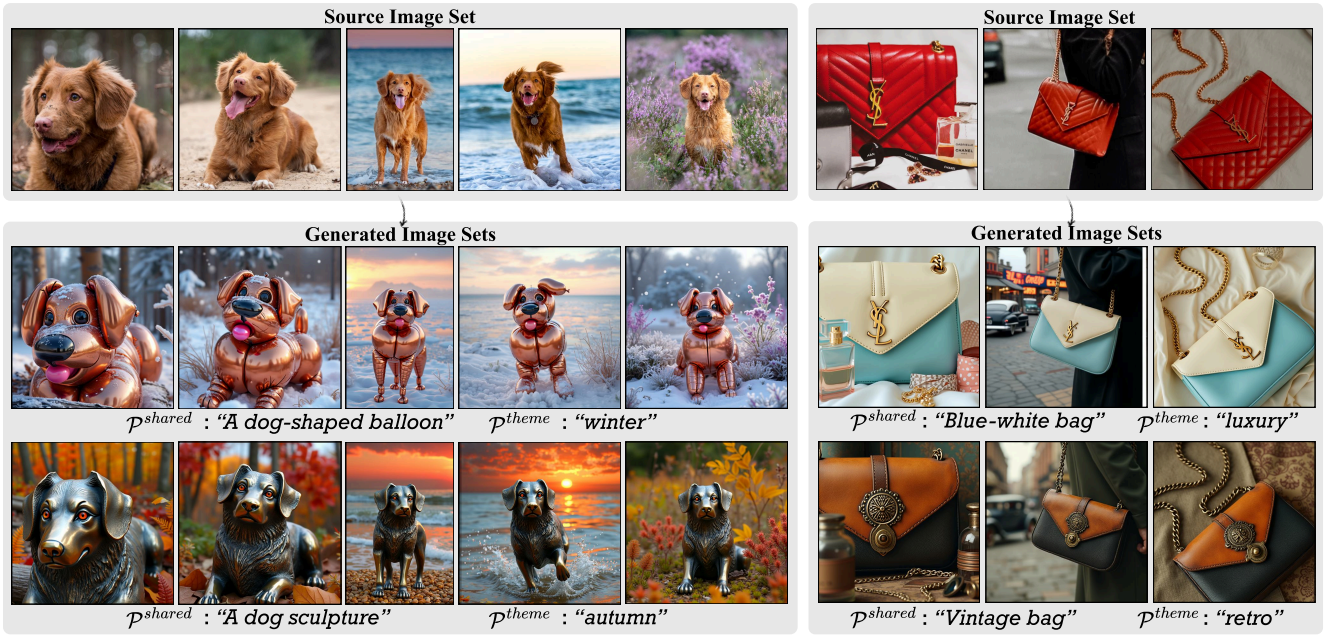


Figure 1. Given a source set of images (top), depicting shared objects in varied settings (*e.g.*, pose, environment, viewpoint), our method, **Match-and-Fuse**, jointly generates an output set in which the consistency among the shared content is preserved (bottom). The output adheres to the user-provided prompts that describe the target shared content ($\mathcal{P}^{shared}$), and the scene's style/theme ($\mathcal{P}^{theme}$).

Abstract

We present Match-and-Fuse – a zero-shot, training-free method for consistent controlled generation of unstructured image sets – collections that share a common visual element, yet differ in viewpoint, time of capture, and surrounding content. Unlike existing methods that operate on individual images or densely sampled videos, our framework performs set-to-set generation: given a source set and user prompts, it produces a new set that preserves cross-image consistency of shared content. Our key idea is to model the task as a graph, where each node corresponds to an image and each edge triggers a joint generation of image pairs. This formulation consolidates all pairwise generations into a unified framework, enforcing local consistency while ensuring global coherence across the entire set. This is achieved by fusing internal features across image pairs, guided by dense input correspondences, without requiring masks or manual supervision, and by leveraging an emergent prior in text-to-image models that encourages coherent generation when multiple views share a single canvas. Match-and-Fuse achieves state-of-the-art consistency and visual quality, and unlocks new capabilities for content creation from image collections. Code and data: project page.

1. Introduction

Much of our visual experience—how we capture, organize, and interpret the world—is structured not around single images but around sets of them. Photo albums, real-estate listings, product catalogs, and historical archives all offer multiple perspectives on shared content captured at different times or viewpoints. Such sets are richer than individual images yet more flexible than continuous video. Despite major progress in single-image and video generation, Generative AI remains largely underexplored for this fundamental unit of visual communication.

Given a source image set and a user prompt describing the desired shared content, our method produces an output set that adheres to the prompt while preserving the source layout and the cross-image consistency of shared elements,

as illustrated in Fig. 1. These shared elements—semantic regions appearing across multiple images—are kept coherent in identity, appearance, and geometry. Importantly, non-shared regions (e.g., backgrounds) are not forced to align and may vary according to a separate thematic prompt. This capability enables creative workflows of transforming fixed multi-view layouts into coherent edits for product ads, character concept art, film set design, and more.

Achieving consistency across an image set is nontrivial. Unlike videos, which benefit from dense temporal sampling and continuity, image sets challenge this assumption. Without temporal cues such as continuous motion, enforcing visual consistency becomes difficult. The problem is further compounded when the content is deformable – e.g., varying human poses, expressions, or environments – where no stable 3D structure can be inferred. Consequently, existing generative methods lack mechanisms to model shared content without strong spatial or temporal cues.

Our method builds on a pre-trained text-to-image diffusion model in a zero-shot, training-free manner. Our key idea is to model an image set as a graph whose nodes are images and whose edges represent joint pairwise generation. This formulation consolidates all pairwise generations into a joint framework, achieving coherence both locally between image pairs and globally across the entire set. Crucially, this formulation allows us to: (i) harness an emergent prior in text-to-image diffusion models, which exhibit coherent behavior when multiple images are composed within a shared generation canvas; (ii) enhance consistency by incorporating dense pixel correspondences computed from the source images, guiding alignment without requiring masks or manual supervision; (iii) flexibly handle sets of varying sizes without sacrificing resolution. We demonstrate the effectiveness of our framework in producing high-quality results from diverse collections of 2–15 images, ranging from real photographs to hand-drawn sketches.

To summarize, we make the following contributions:

- The first method for unstructured *set-to-set generation*, moving beyond image pairs to entire collections.
- A flexible, automated, training-free mask-free approach, requiring only simple text prompts as input.
- A new metric for evaluating fine-grained cross-image consistency that aligns strongly with human judgments.

2. Related work

Text-to-image controlled generation. T2I models [12, 28, 30, 32, 41] have made a remarkable progress. Numerous methods extended T2I models to go beyond input text and condition the generation on various signals, such as style (e.g., [26, 42]) or various spatial controls: pose, depth, or edge maps [44]. A growing line of work tackles image editing rather than conditioning, including trainable editing models [6, 8, 23] that learn to modify an image according to text or paired examples, as well as training-free editing approaches [18, 37, 46] that manipulate diffusion trajectories at inference time. However, all these methods operate on individual images and lack the set-level consistency mechanisms required for our task.

Beyond single-image generation.

Storyboard generation. This line of work aims to produce image sequences depicting consistent recurring characters across frames [1, 14, 36, 42], or emphasizing narrative coherence across prompts [17, 20, 25, 45]. Zero-shot methods manipulate attention features through injected correspondences or masks [36, 45], whereas training-based approaches fine-tune diffusion models for improved consistency [17, 20, 25]. Aside from the limited pose control offered by [36], these models rely solely on text and lack spatial control mechanisms necessary for set-to-set generation.

Consistent image-set generation. Related to our work, Edicho [2] addresses pairwise consistent editing, first modifying one image and then transferring the edit to its partner using 2D correspondences to warp intermediate features. Their method is similar to ours in leveraging explicit matches and operating on unstructured image sets, but it remains strictly pairwise: edits are propagated only from a single reference image, causing coherence to degrade for views farther from that anchor. In contrast, [20] enforces consistency by generating all images simultaneously on a single canvas, where a LoRA module [19] is fine-tuned on the multi-image prompt. This design ties the method to a fixed grid, limiting its scalability to larger image sets.

3D and video-based approaches. Distinct from these methods, 3D editing techniques [9, 16, 39, 40] enable manipulation of static multi-view sets through rendered views but operate on 3D representations rather than directly editing 2D images. This imposes restrictive assumptions that limit their applicability to image sets with unknown camera parameters, articulated objects, and varying backgrounds. Similarly, video editing and generation methods (*e.g.*, [15, 38, 43]) are unsuitable for our task, as they assume temporal continuity and rely on motion coherence that does not hold across unordered image sets.

3. Method

The input to our method is an *unstructured* set of N images along with user-provided prompts: $\mathcal{P}^{\text{shared}}$ and $\mathcal{P}^{\text{theme}}$, describing the target shared content and general style or theme, respectively. Our method outputs N images that preserve the source semantic layout while ensuring visual consistency across shared elements.

We build on a pre-trained, frozen, depth-conditioned T2I model. Although designed for single-image generation, these models have been shown to produce image grids when prompted with joint layouts (e.g., "Side-by-side views of..."), establishing cross-image relationships as demonstrated in recent work [20, 33]. However, this emerged capability, which we refer to as the *grid prior*, exhibits several key limitations: (i) it provides only partial consistency in

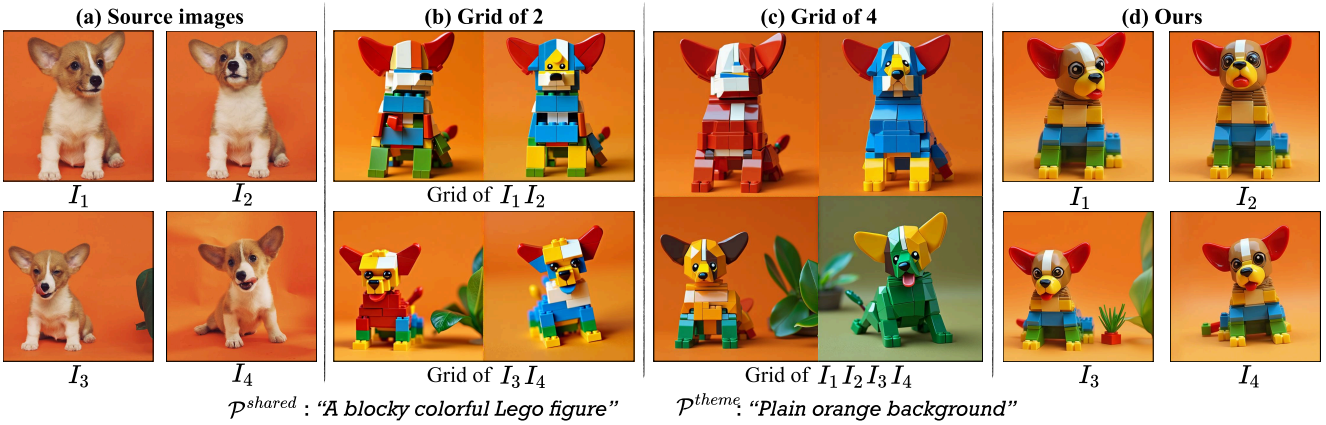


Figure 2. **Image grid generation vs. our method.** (a) Source image set. (b) Joint generation of two-image grid results in partial consistency, where several regions often remain inconsistent in appearance or semantic meaning (e.g., a dog's face). (c) Extending this to more images further reduces consistency. (d) Our method leverages this prior yet overcomes its limitations of consistency and scale.

appearance, shape, and semantics (Fig. 2b); (ii) the consistency deteriorates rapidly as more images are composed (Fig. 2b); and (iii) generating a single canvas is bounded by the model's native resolution, limiting scalability.

Our method leverages the grid prior while overcoming its core limitations (Fig. 2d). Specifically, we model the image set as a **Pairwise Consistency Graph**, comprising all possible two-image grid generations (Sec. 3.2). This allows us to exploit the strong inductive bias of the grid prior, while eliminating its scale limitation. To enhance visual consistency both within each image grid and across grids, we perform joint feature manipulation across all pairwise generations. To this end, we utilize dense 2D correspondences from the source set to automatically identify shared regions – without requiring object masks – and enforce fine-grained alignment. We find that feature-space similarity along these matches correlates strongly with visual coherence, motivating the use of **Multiview Feature Fusion** (Sec. 3.3). We further refine details via **Feature Guidance**, using a feature-matching objective (Sec. 3.4). Our full pipeline is illustrated in Fig. 3.

3.1. Prompt Composition

Given $\mathcal{P}^{\text{shared}}$ and $\mathcal{P}^{\text{theme}}$, we automatically generate detailed *per-image* captions $\mathcal{P}_i^{non-shared}$ using a Vision-Language Model (VLM). The VLM is instructed to identify shared and non-shared elements, integrate the user-provided ideas while respecting the underlying structure, and describe per-image pose variations. More details in Supplementary Materials (SM).

3.2. Pairwise Consistency Graph

We define a graph $G = (V, E)$, where nodes $V = \{I_i\}_{i=1}^N$ represent images, and edges $E = \{i, j \mid i \neq j,\ I_i, I_j \in V\}$ connect all distinct image pairs. During the generation, each node is associated with a noisy latent z_i^t, and each edge with latent of a two-image grid $z_{ij}^t = \text{concat}(z_i^t, z_j^t)$. Edges are further assigned concatenated control depth maps and grid prompts: $\mathcal{P}_{ij}$ = *"Image grid of [$\mathcal{P}^{shared}$]. Left: [$\mathcal{P}_i^{non\text{-}shared}$]. Right: [$\mathcal{P}_j^{non\text{-}shared}$]."*

At each generation step, $\{z_{ij}^t\}_{e\in E}$ are constructed from $\{z_i^t\}_{i=1}^N$ and denoised jointly with Multiview Feature Fusion (Sec. 3.3) that encourages consistency across the entire graph by enforcing pre-computed 2D matches M_{ij}:

$$\{z_{ij}^{t-1}\}_{e\in E} = \text{denoise}_{\text{MFF}}(\{z_{ij}^t\}_{e\in E}, \{\mathcal{P}_{ij}\}_{e\in E}, \{M_{ij}\}_{e\in E}) \quad (1)$$

Since every node z_i^t participates in multiple adjacent edges, a graph denoising step yields multiple versions of $z_{i|ij}^{t-1}$. We thus consolidate pairwise latents $\{z_{ij}^{t-1}\}_{e\in E}$ back into image latents $\{z_i^{t-1}\}_{i=1}^N$ by extracting and averaging all $z_{i|ij}^{t-1}$, following [3]. In practice, full graph connectivity is not required, as analyzed in Sec. 4.4.

3.3. Multiview Feature Fusion

The grid prior alone is insufficient for fine-grained alignment within an image pair (Fig. 2). Moreover, denoising separate edges can yield noticeably different appearances, even when a shared image latent is used. To promote pairwise and global consistency, we follow prior work (Sec. 2) and directly manipulate the model's internal features.

A natural choice for our task is to leverage off-the-shelf 2D correspondences extracted from the source images, which reliably capture shared content through confidence-based filtering. We analyze the model's feature space and observe that cosine similarity at these matched locations strongly correlates with generation consistency (Fig. 5). Hence, we propose to promote it by increasing the similarity of matched features.

Matches between each image pair are defined as a partial feature coordinate mapping $M_{ij} : \mathcal{C}_i \to \mathcal{C}_j$, where a coordinate $\mathbf{c} \in \mathcal{C}_i$ corresponds to $\mathbf{c}' = M_{ij}(\mathbf{c}) \in \mathcal{C}_j$ and $\mathcal{C}_i$ contains all matched points (Fig. 4b). We first address pairwise consistency by considering a two-node graph with a single edge. Let $\mathbf{f}_{ij} \in \mathbb{R}^{H\times 2W\times D}$ be a feature map from a model's forward pass on a grid, interpreted as $\mathbf{f}_{ij} = \text{concat}(\mathbf{f}_i, \mathbf{f}_j)$,

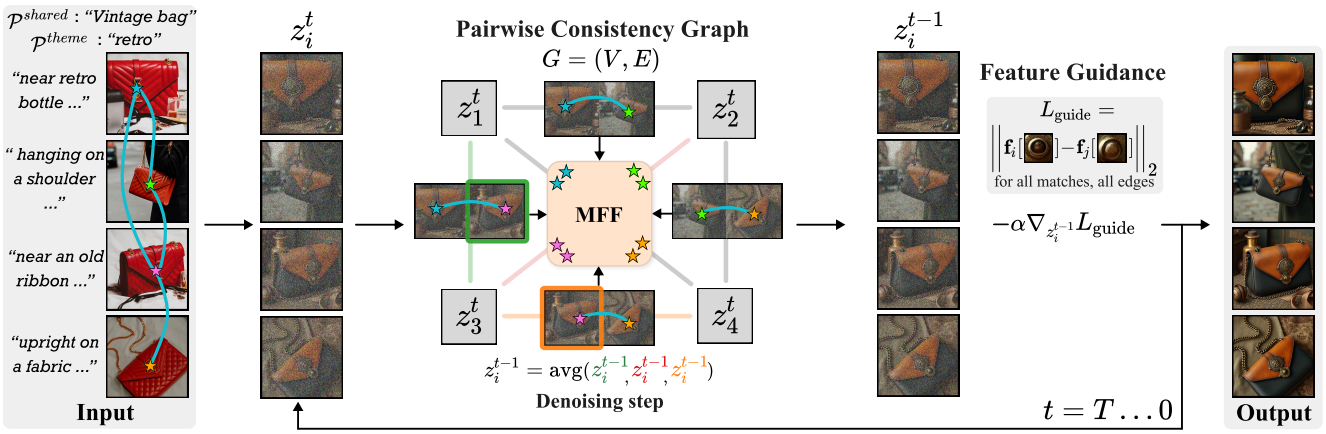


Figure 3. **Match-and-Fuse pipeline.** Example for 4 images. In pre-processing, pairwise matches are computed between all inputs, and per-image prompts are generated from the set-level prompts. At each denoising step, noisy image latents form a Pairwise Consistency Graph, whose edges z_{ij}^t are jointly denoised with Multiview Feature Fusion (MFF) and aggregated back into per-image latents z_i^{t-1} by averaging over adjacent edges. The latents are further refined with Feature Guidance via a feature-level matching objective.

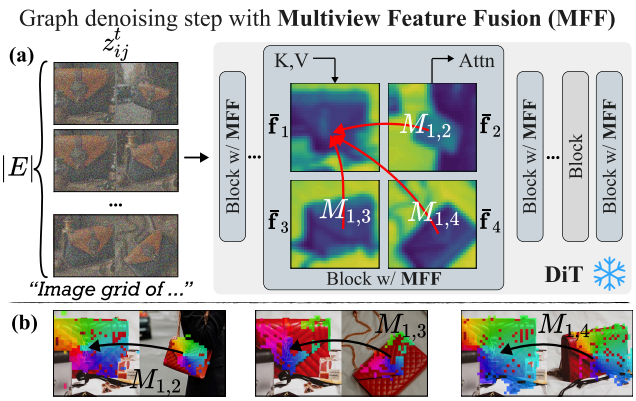


Figure 4. **MFF Denoising step.** (a) Two-image grids on all edges are denoised with a frozen DiT. Selected blocks average K,V along adjacent edges into $\bar{\mathbf{f}}_i$, which are then fused via source matches (b). Images are fused jointly, illustrated by arrows for i=1.

with $\mathbf{f}_i, \mathbf{f}_j \in \mathbb{R}^{H \times W \times D}$. Given a pooling operator $\mathbf{f}[\mathbf{c}]$ extracting the vector at $\mathbf{c}$, Multiview Feature Fusion (MFF) fuses matched coordinates:

$$\mathbf{f}_i[\mathbf{c}] \leftarrow \tfrac{1}{2}\left(\mathbf{f}_i[\mathbf{c}] + \mathbf{f}_j[M_{ij}(\mathbf{c})]\right), \quad \forall \mathbf{c} \in \mathcal{C}_i. \tag{2}$$

For an N-node graph, Eq. (2) is generalized by aggregating features of each image over its incident edges, followed by joint fusion across all images:

$$\bar{\mathbf{f}}_i = \frac{1}{|\delta(i)|} \sum_{e \in \delta(i)} \mathbf{f}_i^e, \quad \delta(i) = \{e \in E \mid i \in e\} \tag{3}$$

$$\mathbf{f}_i[\mathbf{c}] \leftarrow \tfrac{1}{N}\left(\bar{\mathbf{f}}_i[\mathbf{c}] + \sum_{j \neq i} \bar{\mathbf{f}}_j[M_{ij}(\mathbf{c})]\right) \tag{4}$$

This stage is illustrated in Fig. 4. See SM for further details.

3.4. Feature Guidance

To align small details, we further perform feature guidance [7, 11, 31], completing each generation step (Fig. 3, right). Specifically, we define an optimization objective over all

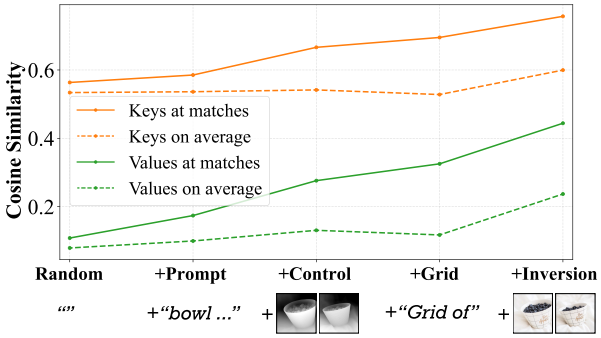


Figure 5. **Matched feature similarity vs. visual consistency.** We consider increasingly consistent generation (left to right): random images → adding descriptive prompts → adding control signals → generating in a grid → DDIM [35] inversion which reconstructs fully consistent source images. Keys and values differ in scale but follow the same pattern: cosine similarity at matched locations rises with consistency. Dashed lines show the baseline all-to-all similarity of feature maps. Points are averaged over 40 image pairs, all correspondences, blocks, and timesteps.

edges as the distance between matched features, and refine all node latents via gradient descent with $\nabla_{z_i^{t-1}} L_{\text{guide}}$ of:

$$L_{\text{guide}} = \frac{1}{|E|} \sum_{\{i,j\} \in E} \frac{1}{|M_{ij}|} \sum_{\mathbf{c} \in M_{ij}} ||\mathbf{f}_i[\mathbf{c}] - \mathbf{f}_j[M_{ij}(\mathbf{c})]||_2 \tag{5}$$

The feature maps $\mathbf{f}$ are extracted from an additional DiT forward pass. MFF can be viewed as the analytical solution to this optimization problem operating during inference, while guidance corrects remaining inconsistencies directly in the latent space. Using guidance as the primary driver would be expensive and risk pushing latents off-distribution, so we apply it only as a light refinement. The exact schedule is provided in SM. Because gradients propagate through the model, updates have a wider receptive field than the discrete match locations, improving robustness under sparse correspondences (see Sec. 4.4).

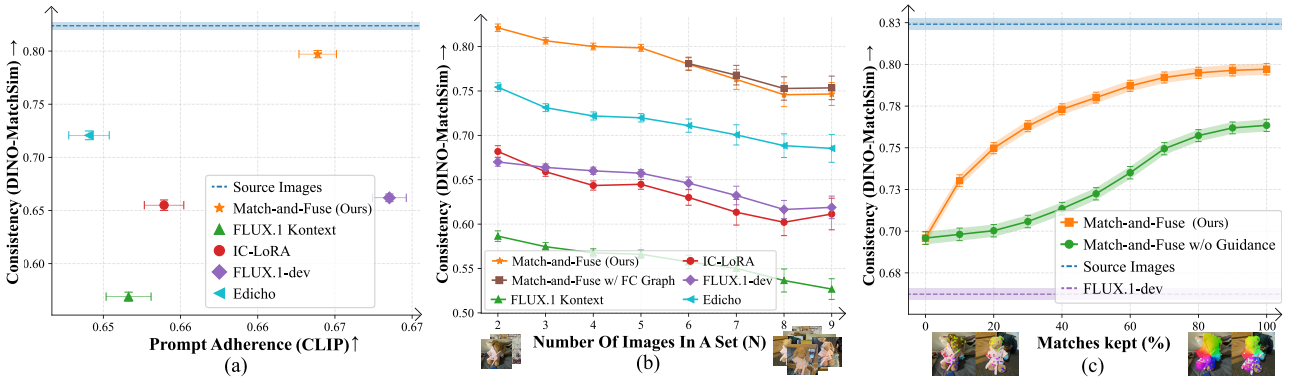


Figure 6. **Quantitative comparison and analysis.** We evaluate subject Consistency (DINO-MatchSim) of Match-and-Fuse as a function (a) of Prompt Adherence (CLIP Score), compared to baselines; (b) Number of Images, compared to baselines and method variant with a fully-connected graph; (c) % of Matches, compared to method variant w/o Feature Guidance. Error bars are SEM.

Figure 7. **Ablation.** Subtraction of each of our method's components (a-c) degrades consistency, demonstrating their necessity in the full method (d). Each setting is detailed in Sec. 4.4.

	CLIP ↑	DreamSim ↑	DINO-MatchSim ↑
FLUX Kontext	0.65 ± 0.002	0.78 ± 0.004	0.57 ± 0.004
IC-LoRA	0.65 ± 0.001	0.71 ± 0.006	0.65 ± 0.004
FLUX	**0.67 ± 0.001**	0.76 ± 0.004	0.66 ± 0.004
Edicho	0.65 ± 0.001	0.81 ± 0.004	0.72 ± 0.004
Match-and-Fuse	0.66 ± 0.001	**0.85 ± 0.004**	**0.80 ± 0.003**
w/o Guidance	0.66 ± 0.001	0.82 ± 0.004	0.76 ± 0.004
w/o MFF	0.66 ± 0.001	0.83 ± 0.004	0.78 ± 0.003
w/o Pairwise Graph	0.66 ± 0.001	0.82 ± 0.004	0.75 ± 0.004

Table 1. **Quantitative evaluation.** We compare our method to baselines and ablate method components.

4. Experiments

We extensively evaluate Match-and-Fuse. Our implementation builds on FLUX [5] with depth-conditioning [41], and uses RoMA [10] for matching. Full details are in SM.

	Users ↑	VLM ↑
Kontext	88%	82%
IC-LoRA	90%	92%
FLUX	92%	94%
Edicho	83%	78%

(a)

	Agreement ↑
DreamSim	84.3
VLM	84.9
DINO-MatchSim	**91.4**

(b)

Table 2. (a) **User Study and VLM evaluation.** 2AFC voting demonstrates the preference of humans and VLM of our method over all baselines. (b) **Agreement between metrics with human judgments.** % of samples for which the winner according to a higher metric agrees with the users' majority vote. Average across comparisons and baselines.

4.1. Evaluation Setup

Evaluating our task requires image sets that share content across diverse scenarios, along with a metric that captures fine-grained cross-image consistency. Since no existing benchmark or metric supports this setting, we introduce a setup tailored to consistent set-to-set generation.

Evaluation set. We curate a benchmark with 400 edits for 149 diverse distinct image sets, 3–15 images in each, combining all sets from [22] and [29] with frames sampled from 3D datasets [4, 16, 24], keyframes of a few publicly sourced videos, and sketched storyboards generated by ChatGPT. Data will be publicly released.

Metrics. We quantitatively evaluate our method, beginning with fine-grained cross-image consistency. Previous work relies on *global* visual-similarity metrics to assess set consistency—for example, [36] uses DreamSim [13] on object-masked regions. However, as our user study (Tab. 2b) shows, such global metrics are not ideal. To address this, we introduce *DINO-MatchSim*. For each image pair, we compute patch-level nearest-neighbor correspondences NN_{ij} between DINOv3 [34] feature maps from the source images, and measure similarity at the corresponding output locations:

$$S_{ij} = \frac{1}{|NN_{ij}|} \sum_{(p,q)\in NN_{ij}} \cos\left(\tilde{F}_i(p), \tilde{F}_j(q)\right), \qquad (6)$$

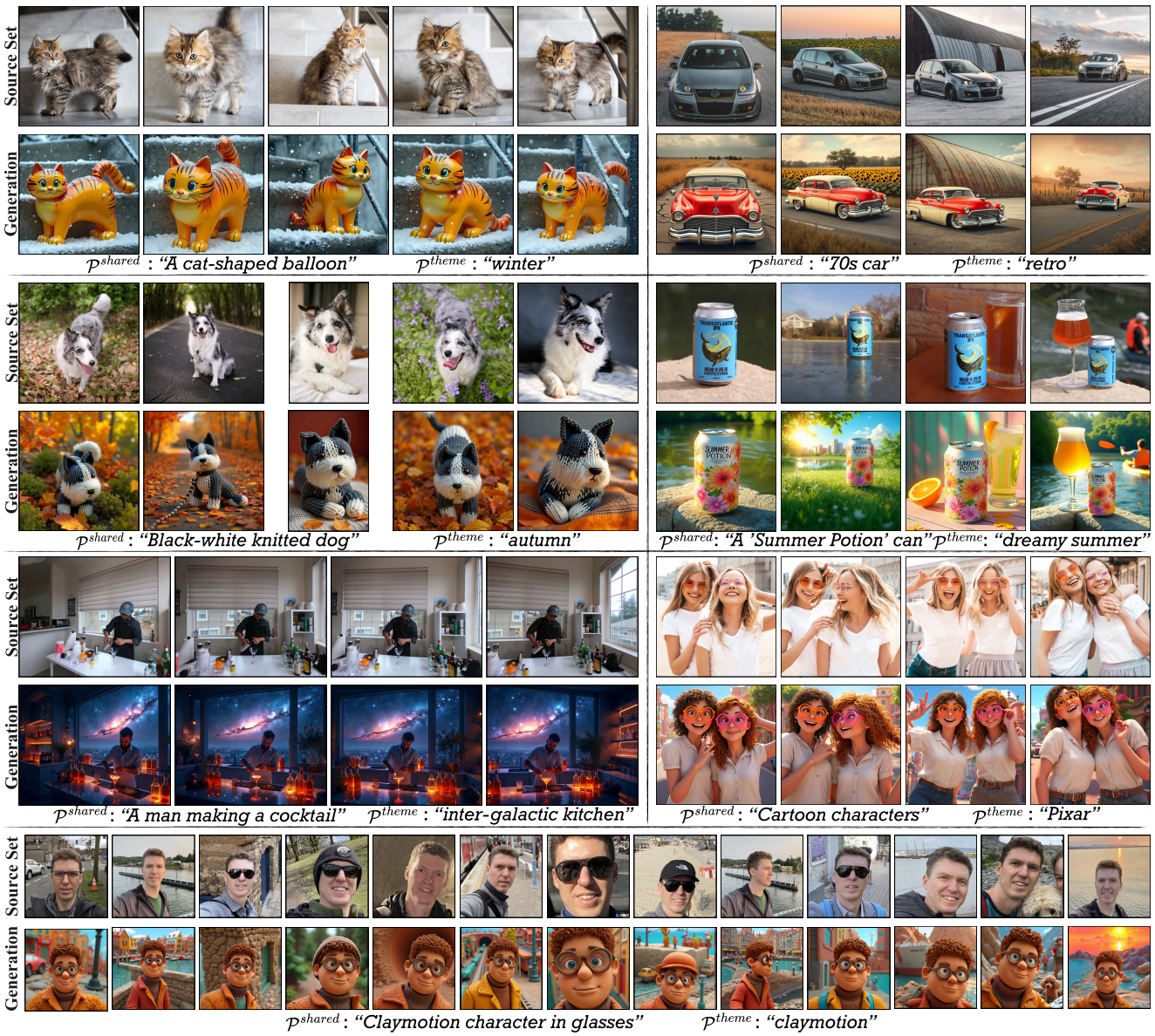


Figure 8. **Qualitative results.** Match-and-Fuse generates consistent content of rigid and non-rigid shared elements, single and multi-subject, with shared or varying background, preserving fine-grained consistency in textures, small details, and typography. Notably, it can generate consistent long sequences (last row). See SM for full sets of results.

where $\tilde{F}_i$ denotes the i^{th} output feature map. The average similarity across all image pairs is *DINO-MatchSim*. For prompt adherence, we report the average CLIP score over the set. To assess human preference, we conduct a large-scale user study on a random subset of the benchmark, comparing Match-and-Fuse to a baseline in a randomized two-alternative forced-choice (2AFC) format. Finally, we include a VLM-based evaluation using GPT-5 [27], which received the same content as the human raters.

4.2. Qualitative results

Sample qualitative results of our method are shown in Figs. 1, 8 and 9, demonstrating that it performs robustly under non-rigid pose variations, diverse viewpoints, and partial occlusions, on single or multi-object sets (*e.g.* two girls). The output sets maintain fine-grained consistency in textures (*e.g.* crochet, balloon and can print), small elements (*e.g.* kitchen items), and typography (*e.g.* the can title). Notably, Match-and-Fuse is able to handle big number of images (*e.g.* $N = 13$ in Fig. 8, last row).

4.3. Comparisons

Since no existing method is designed for our set-to-set generation task, we compare Match-and-Fuse to the closest baselines: **(1)** FLUX [5] with ControlNet [41] conditioning serves as a baseline for assessing independent genera-

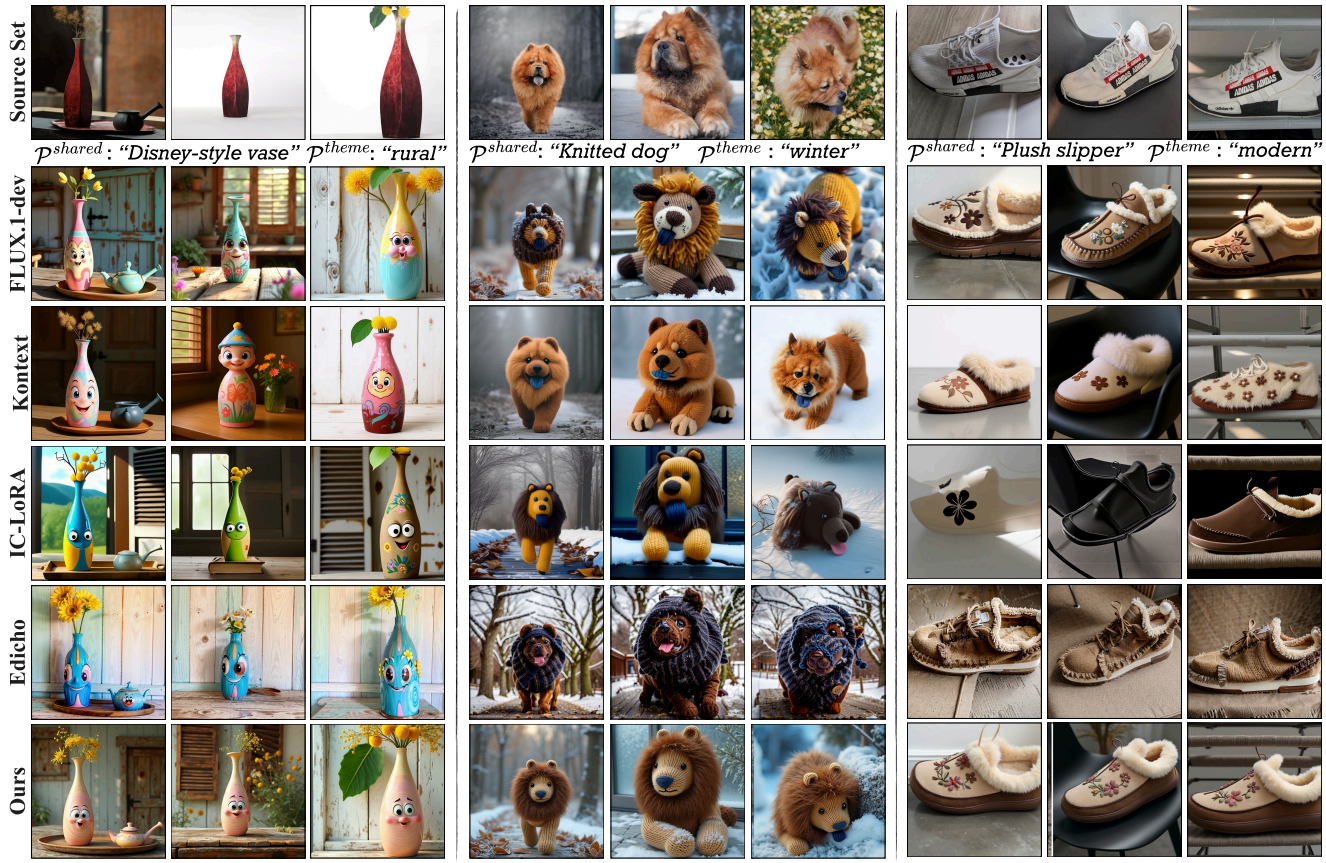


Figure 9. **Qualitative comparisons.** Match-and-Fuse (Ours) generates image sets with the highest consistency compared to our baselines FLUX [5], FLUX Kontext [23], IC-LoRA [20], and Edicho [2]. See Sec. 4.3 for comments on each result and SM for full image sets.

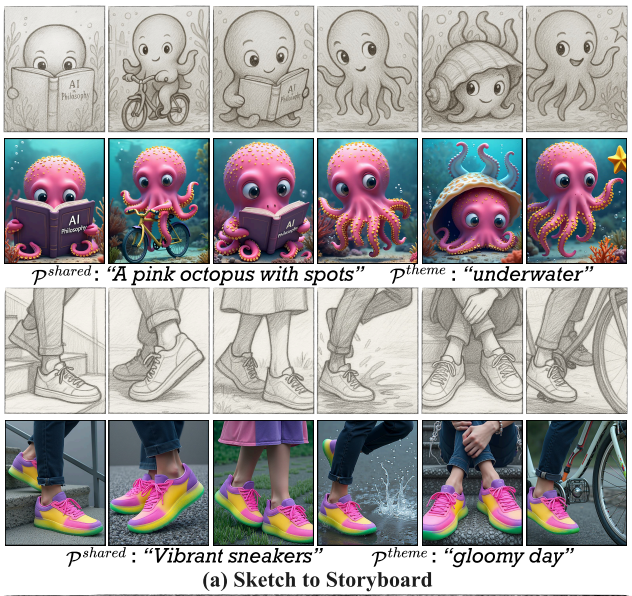
tion using the same base model as our method. **(2)** FLUX Kontext [23], an image-to-image editing model. While it provides consistent edits over multiple turns on a single image, it lacks an explicit mechanism for achieving consistency across a single-turn image set editing. **(3)** IC-LoRA [20], text-driven generation of image sets with shared elements by fine-tuning LoRA modules for specific context types. We use their most relevant public checkpoint and add spatial ControlNet conditioning for comparability. **(4)** Edicho [2], the closest baseline to our method, combines depth-conditioned generation with explicit correspondences, but is limited to an image pair editing through a single reference. Additional implementation details are in SM.

Figure 9 shows representative qualitative results. FLUX [5] produces inconsistencies in both coarse and fine details. FLUX Kontext [23] has low prompt adherence (*e.g.*, a toy remains animal-like), and often distorts object structure (*e.g.* vase, dog). IC-LoRA [20] achieves partial consistency, with coherence often restricted to subsets (vase). Its realism and fidelity are notably lower than FLUX. Edicho [2] performs best among baselines but still shows noticeable inconsistencies (dog), as its one-to-all warping enforces consistency only with the first image, making the choice of an anchor ambiguous and degrading coherence across views. In contrast, Match-and-Fuse (bottom row) maintains high image quality while substantially improving structural and fine-grained consistency. We additionally compare to the closed-source Nano Banana API [8] and find that it does not solve our task (SM).

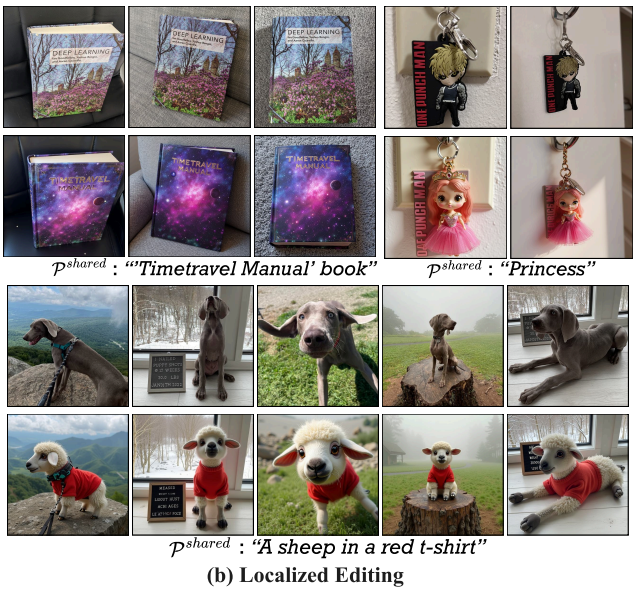
We quantitatively evaluate all methods using the metrics in Sec. 4.1. Figure 6a plots Consistency (DINO-MatchSim) versus Prompt adherence (CLIP score). As an upper bound, we report DINO-MatchSim on the source sets, which are consistent by definition. Our method achieves the highest consistency among baselines, approaching the source score, while maintaining a comparable CLIP score. Since DINO-MatchSim evaluates consistency relative to the source structure, methods that distort shape, such as FLUX Kontext, score lowest. Tab. 1 further shows that Match-and-Fuse attains the highest DreamSim score. We complement these results with a user study and VLM-based evaluation. As shown in Tab. 2a, participants and VLM preferred our method over all baselines.

4.4. Analysis & Ablations

Ablation. We ablate each component in Fig. 7 and Tab. 1. *W/o Pairwise Consistency Graph*, pairwise steps become single-image predictions, with MFF and Guidance applied



(a) Sketch to Storyboard

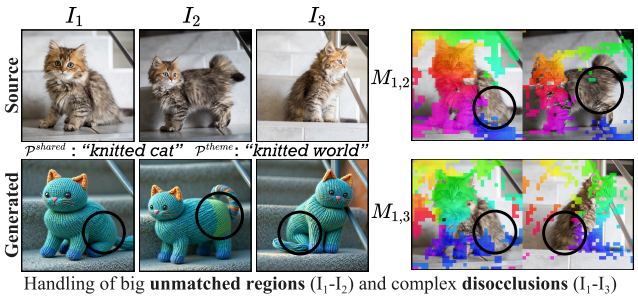


(b) Localized Editing

Figure 10. **Extended Applications.** Our method generalizes across various settings, enabling (a) consistent story generation from sketches and (b) localized editing omitting $\mathcal{P}^{theme}$ via FlowEdit [21] integration. See full sets of results in SM.

per image. Correspondences alone cannot align appearances, leading to identity drift. In *w/o Multiview Feature Fusion*, pairwise updates use only the grid prior; latent versions aggregated across edges diverge more easily, reducing consistency. *Omitting Feature Guidance* at each step leads to misaligned fine-grained details.

Number of images. A fully connected graph has $O(N^2)$ edges. To maintain scalability, we limit each node degree to 4 random neighbors, yielding full connectivity for $N \leq 5$ and increasing sparsity thereafter. This keeps runtime linear while preserving cross-image communication. We use this setup in all our experiments. Figure 6b shows consistency as a function of N. For each N-image set, we evaluate subsets of the first $N' \in [2, N]$ images and report averages over all N'. Our method is slightly below its fully connected variant but remains similar in performance. Although Match-and-Fuse degrades with N it remains more consistent for 9 images than baselines do for 2.



Handling of big **unmatched regions** (I_1-I_2) and complex **disocclusions** (I_1-I_3)

Figure 11. **Limitations.** The method may produce inconsistencies in largely unmatched regions (I_1, I_2) or under complex disocclusions (symmetry in I_2, I_3).

Match sparsity. We evaluate robustness to reduced correspondences by randomly subsampling x% of matches. As shown in Fig. 6c, DINO-MatchSim remains high even under strong sparsification, whereas the variant without Feature Guidance drops more rapidly—highlighting its role in maintaining consistency with sparse matches.

4.5. Extended Applications

As shown in Fig. 10a, our method generalizes to sketched source sets, enabling consistent visualization of *storyboards*. In Fig. 10b, we further demonstrate *localized editing* by combining our method with FlowEdit [21]. We provide default integration parameters in the SM; however, as is a known limitation of FlowEdit, achieving the desired balance between structure preservation and appearance change often requires mild per-edit hyperparameter tuning. Automating this selection is left for future work.

5. Discussion and Conclusions

We introduced Match-and-Fuse, the first method for controlled generation from unstructured image collections. Extensive experiments demonstrate that it outperforms strong baselines and enables diverse creative applications, advancing a largely underexplored yet fundamental visual modality. Several limitations remain for future work. Our performance depends on the quality and density of pixel correspondences, which may result in inconsistencies in largely unmatched or ambiguous regions (e.g., disocclusions or symmetries; see Fig. 11). It also relies on the ability of the base model to preserve the source conditioning depth maps, which may occasionally deviate from the source layout, depending on the target prompt and generative prior of the T2I model (*e.g.* body and hand poses of the girls in Fig. 8). We believe these insights lay a path toward future extensions, such as video collections and foundation models for set-to-set generation.

Acknowledgments

We thank Danah Yatim for her valuable comments. We thank Vladimir Kulikov for insightful discussions on FlowEdit. We thank Qingyan Bai and the other authors of Edicho for conducting qualitative comparisons with their method.

References

[1] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2

[2] Qingyan Bai, Hao Ouyang, Yinghao Xu, Qiuyu Wang, Ceyuan Yang, Ka Leong Cheng, Yujun Shen, and Qifeng Chen. Edicho: Consistent image editing in the wild. In *arXiv preprint arXiv:2412.21079*, 2024. 2, 7

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 3

[4] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 5

[5] Black-Forest. Flux: Diffusion models for layered image generation. https://github.com/black-forest-labs/flux, 2024. Accessed: 2024-09-24. 5, 6, 7

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2

[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 4

[8] Google DeepMind. Gemini 2.5 flash image ("nano banana") model/api, 2025. Accessible via Google Gemini API. 2, 7

[9] Jiahua Dong and Yu-Xiong Wang. Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. In *NeurIPS*, 2023. 2

[10] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching, 2023. 5

[11] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. 2023. 4

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 2

[13] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, pages 50742–50768, 2023. 5

[14] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space, 2025. 2

[15] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 2

[16] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 5

[17] Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion, 2024. 2

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 2

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[20] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. 2024. 2, 7

[21] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024. 8

[22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023. 5

[23] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 7

[24] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video, 2022. 5

[25] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6190–6200, 2024. 2

[26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[27] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025. Technical report. 6

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[31] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 4

[32] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions, 2020. 2

[33] Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. 2024. 2

[34] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 5

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 4

[36] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 2, 5

[37] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 2

[38] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024. 2

[39] Yuxuan Wang, Xuanyu Yi, Zike Wu, Na Zhao, Long Chen, and Hanwang Zhang. View-consistent 3d editing with gaussian splatting. *arXiv preprint arXiv:2403.11868*, 2024. 2

[40] Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Prisacariu. GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing. *ECCV*, 2024. 2

[41] XLabs-AI. Flux-controlnet collections. https://huggingface.co/XLabs-AI/flux-controlnet-collections, 2024. Accessed: 2024-11-25. 2, 5, 6

[42] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 2

[43] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6463–6474, 2024. 2

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[45] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *NeurIPS 2024*, 2024. 2

[46] Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation. *arXiv preprint arXiv:2502.17363*, 2025. 2