PLACID: Identity-Preserving Multi-Object Compositing via Video Diffusion with Synthetic Trajectories

Supplementary Material

This Supplementary Material provides complementary information and visuals on different aspects presented in the main paper. It includes additional visualizations of emerging capabilities (Sec. S.1) and ablation studies (Sec. S.2). Further experiments on video length and first frame initialization are presented in Sec. S.3. Expanded information on training data and the evaluation set are detailed in Sec. S.4 and Sec. S.5, respectively. Comparisons with state of the art works are shown in Sec. S.6, while model limitations are discussed in Sec. S.7.

S.1. Emerging Capabilities

As explained in Sec. 4.3, our model's capabilities extend beyond object compositing. By leveraging its ability to maintain background integrity, align with text instructions, and faithfully reproduce colors, it can perform text-based and color-based image editing, as demonstrated in Fig. S.1. When no background image is provided, the model can also synthesize text-aligned backgrounds, as shown in Fig. S.2, and generate extra objects or props that interact with the main subjects, as in Fig. S.3. Additionally, PLACID produces spatiotemporally consistent videos that can be used in their entirety, as seen in Fig. S.4 or in the project page.

S.2. Ablation Studies

(i) Training Strategy. We visualize the effects of each additive change in our training pipeline in Fig. S.5. The base model struggles with our task due to the significant difference between evaluation and training data. Finetuning on our entire training set improves results but still yields blurry, low-quality images with poor identity preservation. Conditioning on object and background images enhances object identity preservation and consistency, though text instruction adherence remains challenging (Fig. S.5 rows 1,3,4). Adding special text tokens for better text-visual input correlation improves text understanding and composite coherence. However, using noisy intermediate frames from Subject-200k [3] data leads to missing (Fig. S.5 row 2) and distorted objects (Fig. S.5 row 3). By using only the final frame from Subject-200k for training supervision in our final model, we reduce this noise, resulting in improved text-image alignment, more natural and complete scenes, and better fine-grained preservation of objects and backgrounds.

(ii) Training Data. Fig. S.6 compares models trained on different data sources. In-the-Wild images from Unsplash [6] offer good photorealistic background preservation but tend towards copy-pasting, especially on plain backgrounds, and struggle with textual instructions and reposing. Manually designed compositions provide appealing layouts on plain backgrounds but struggle with photorealistic backgrounds, object interactions, and text-image alignment. Subject-200k data greatly improves text alignment (Fig. S.6, row 2) but performs poorly in background preservation and retaining specific object details. Side-by-side compositions provide a stronger bias towards object relighting and relative scaling, but struggle with drastic reposing and integrating objects in photorealistic scenes (Fig. S.6, rows 3,4). Our final model effectively combines these complementary data sources, offering natural-looking scenes with coherently transformed and relighted objects, good text alignment, preserved object identities, and faithful background scenes, including fine-grained colors.

S.3. Additional Experiments

(i) Study on Number of Frames. We train our model on 9-frame videos, balancing the need for sufficient length to learn object animation along synthetic trajectories with efficient use of training resources. At inference time, the model can generate videos of varying lengths, with the last frame always serving as the target image. Tab. S.1 illustrates how different quantitative metrics change with video length, from 9 to 81 frames (the default in our base model [15]). Longer videos allow for greater transformations, improving text-image alignment but potentially compromising identity, background, and color fidelity. Conversely, shorter videos (9 or 17 frames) may not fully transition objects from initial random locations to the desired composition, resulting in lower text-image alignment. For our compositing experiments in the main paper and SupMat, we use 33 frames to achieve an optimal balance between identity preservation and text-image alignment. Notably, users can choose to adjust this trade-off by simply generating shorter or longer videos, providing flexibility in the model's application.

(ii) Study on First Frame Initialization. As detailed in Sec. S.3 (i), our model's use of fewer frames at inference time limits the extent to which objects can move from their initial to final positions. While this might seem restrictive, it actually provides an additional layer of control. Users can take advantage of this characteristic by manually choosing how to arrange objects on the background to create the first frame, rather than relying solely on text descriptions to specify the desired layout. Due to the short video length,

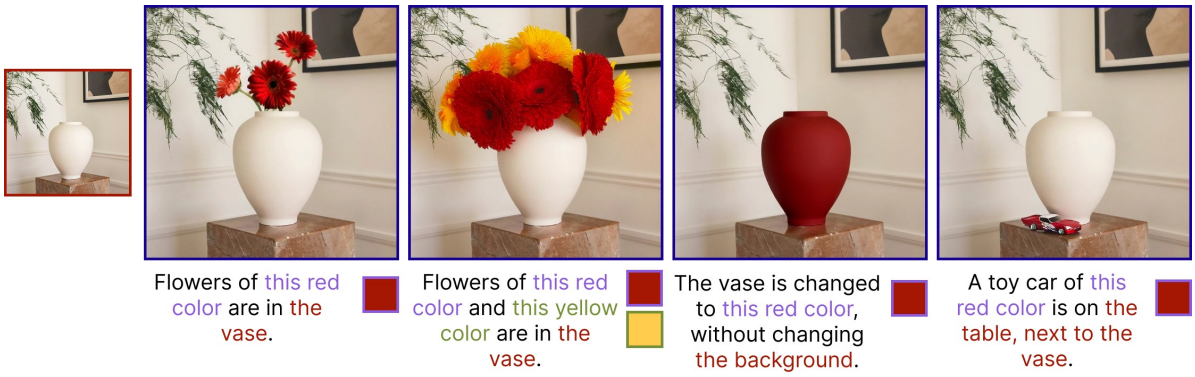

Figure S.1. Text- and Color-Guided Image Editing. Four edits are applied to a single background image using text and color instructions.

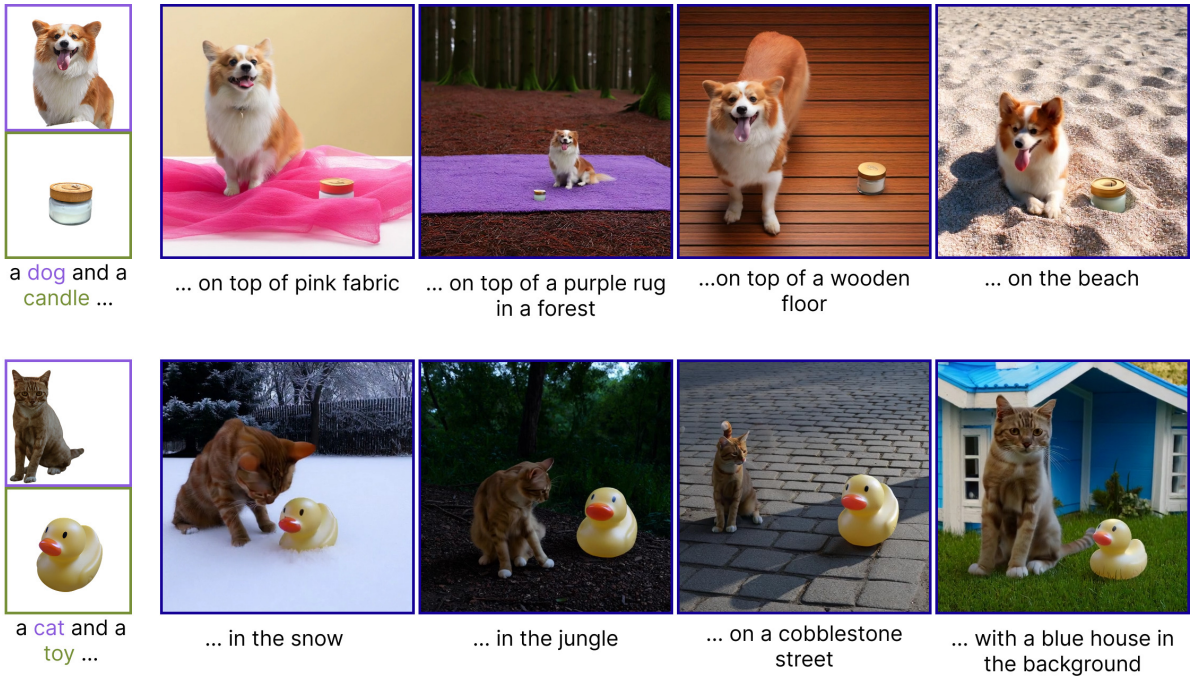

Figure S.2. Subject-Guided Image Generation. PLACID can generate images where objects are contextualized based on textual prompts.

# Frames	CLIP-I↑	DINO↑	CLIP-T↑	MSE-BG↓	Chamfer ↓	Missing ↓
9	0.694	**0.450**	0.330	0.027	3.228	0.045
17	0.688	0.444	0.330	0.027	5.590	0.048
33	**0.705**	0.440	0.336	**0.019**	4.641	**0.044**
61	0.699	0.436	0.334	0.032	**3.061**	0.054
81	0.693	0.422	**0.337**	0.043	5.500	0.057

Table S.1. Impact of Video Length on Image Quality Metrics. This study examines the effect of varying the number of generated video frames on the quality of the final output image. We quantitatively evaluate: identity preservation (CLIP-I, DINO), text alignment (CLIP-T), background preservation (MSE-BG), color fidelity (Chamfer), and object omission rate (Missing).

objects tend to remain close to their initial positions and scale, as shown in Fig. S.7. This method is only effective when the text caption aligns with the initial object arrangement or when no caption is provided. However, it's important to note that if the caption specifies a different layout, these textual instructions will override the first frame initialization. This feature offers users a visual means of controlling object placement while still preserving the model's ability to respond to text prompts.

S.4. Training Data Generation

PLACID is trained on videos where objects follow synthetic trajectories from random initial positions to form coherent, visually appealing layouts in the final frame. Text descriptions are used during training for additional guidance but are optional at test time. During training, objects in the ini-

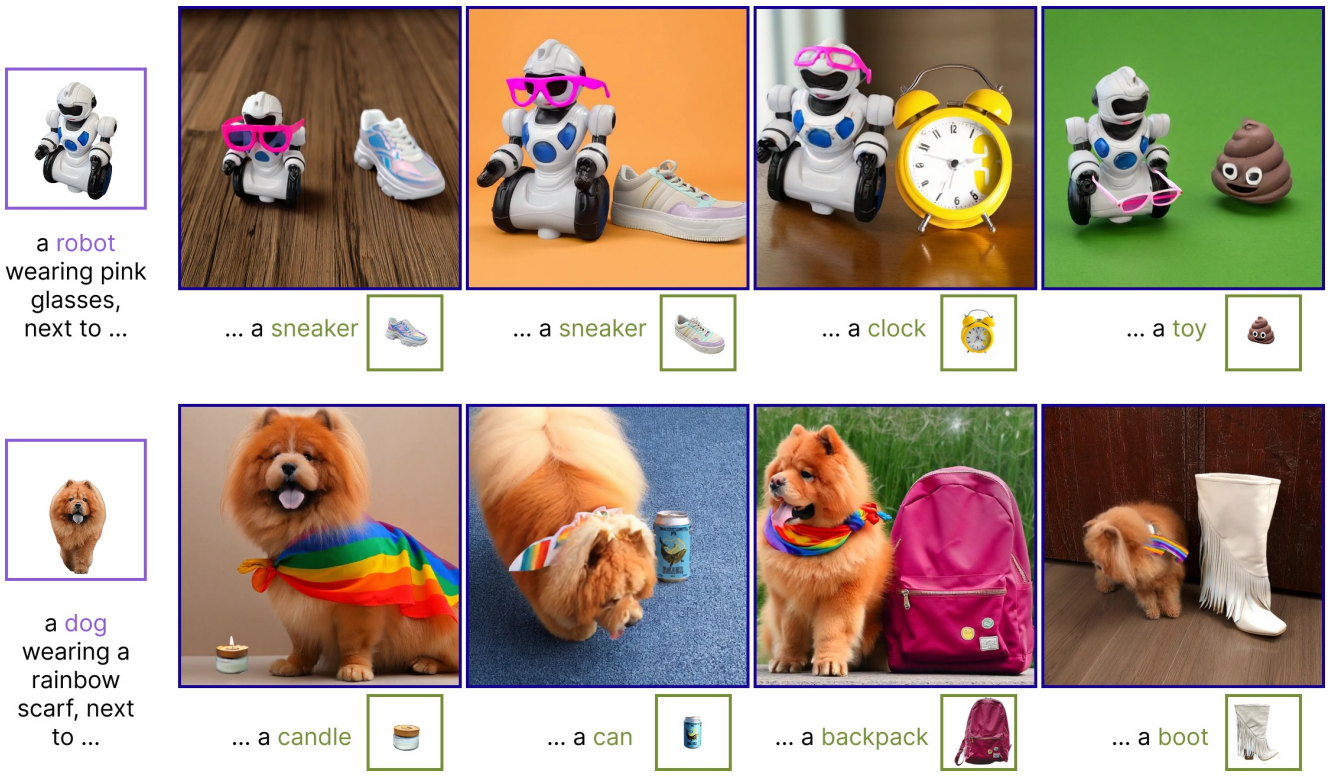

Figure S.3. Subject-Guided Generation with Text-Guided Interacting Objects. PLACID can generate background scenes and additional interacting objects (*e.g.*, glasses, scarves) based on textual prompts, incorporating them with the main subjects.

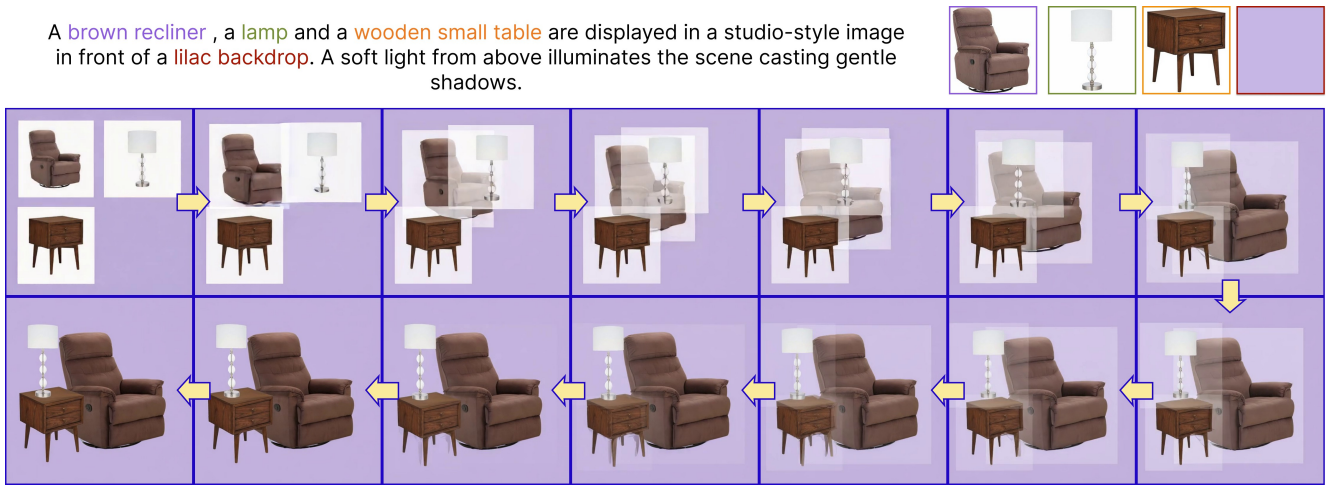

Figure S.4. Frame-by-Frame Generation. PLACID outputs a spatiotemporally consistent video, with the desired image as the final frame.

tial frame are enclosed with white bounding boxes to improve robustness and usability, enabling the direct use of most studio images without preprocessing and increasing tolerance to imperfect segmentations, while still supporting pre-segmented objects at test time.

As detailed in Sec. 3.3, the training data is generated from three distinct image sources: (i) Professional Multi-Object Images, (ii) a Subject-Driven Generation Paired Dataset, and (iii) Synthetic Side-by-side Compositions.

This section provides additional information and visualizations for each of these sources, along with more details on the augmentations applied to the training data.

Figure S.5. Training Strategy Ablation Studies. We visually compare generation on (i) base model [15], (ii) finetuning on our data, (iii) conditioning on object images and background scene, (iv) adding special grounding tokens, and (v) adapting loss into our final model.

S.4.1. Professional Multi-Object Images

(a) In-the-Wild Images: All 13,103 images labeled 'Product Photography', or 'Flat Lay' on Unsplash [6] are parsed together with their captions (*e.g.*, yellow box contents in Fig. S.8). For each image we first identify the relevant foreground objects by using the grounding pipeline of Grounding-DINO [9] together with SAM [8], as in [13]. The raw detections are then cleaned by (i) merging those with the same label, (ii) discarding duplicates with different labels, (iii) eliminating overlap by separating small objects largely interacting with larger ones (*e.g.*, pen and notebook in Fig. S.8), and (iv) keeping only objects with mask-coverage between 0.5% and 80% of image area. Images that end up with no valid detections are dropped, and the remaining backgrounds are restored with the inpainting method of [20] on the union of all foreground objects, using a dilation of 50. Each retained object is extracted, randomly reordered, rescaled, and optionally warped with a mild perspective transform; it is then pasted onto a white box to emulate the unsegmented object image I_i. To build F_1, the first frame of the training video, we scatter these boxed objects across an empty canvas. The canvas background can be (i) plain white, (ii) a randomly chosen background, or (iii) the inpainted original background. Finally, a short video is rendered as in Fig. S.8 (bottom): every object travels in a straight line at constant speed from its random start location to its target position in the professional

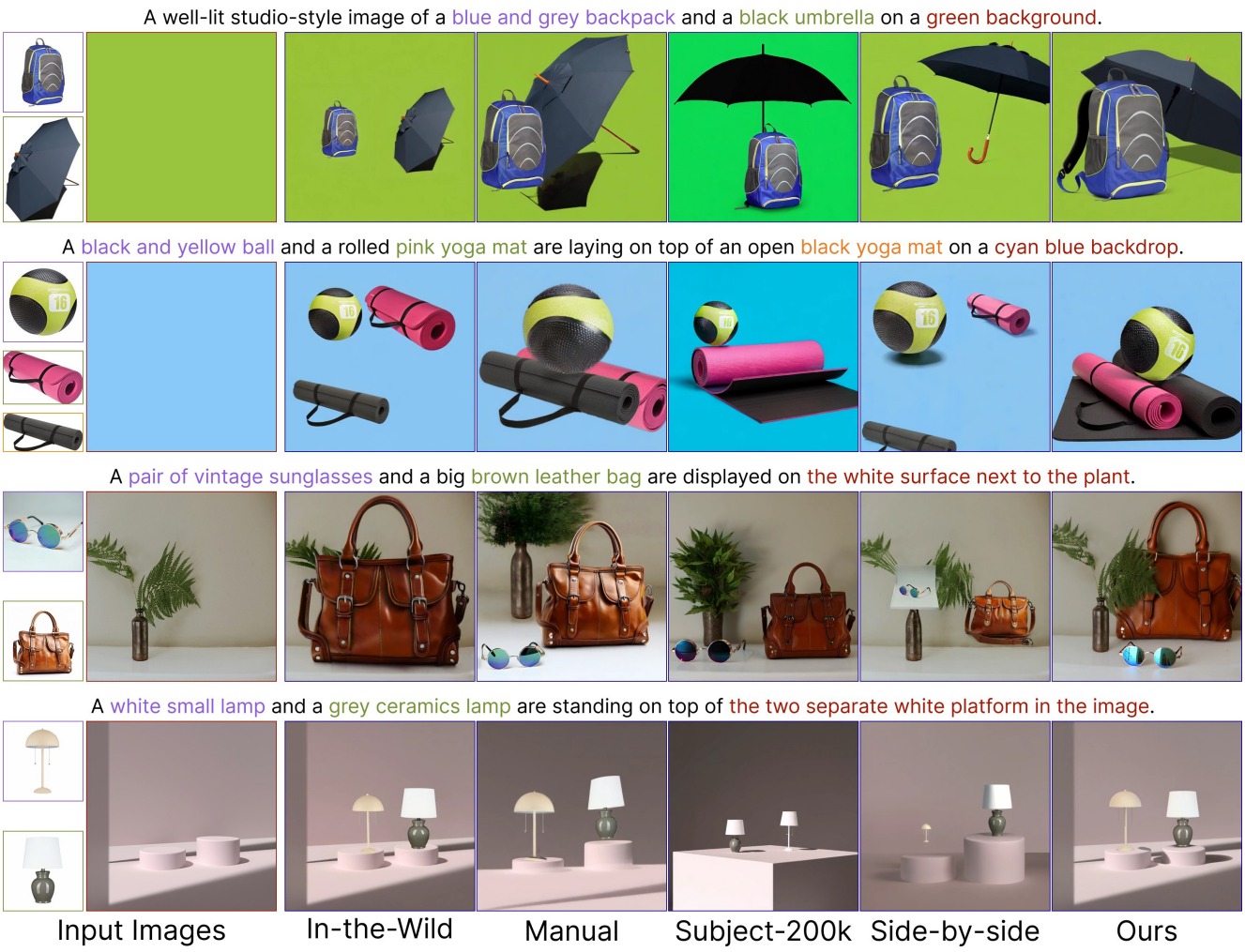

Figure S.6. Training Data Ablation Studies. We visually compare training with (i) In-the-Wild Professional Multi-Object Images from Unsplash [6], (ii) Professional Manual Designs, (iii) Pairs from Subject-200k [3], (iv) Synthetic Side-by-side Compositions, and (v) Our model, simultaneously trained with all data sources.

composition, while simultaneously the white boxes fade out and, if a plain-white background was used initially, the true background fades in. The last frame of the video, F_K, is thus the desired multi-object composite image, and the intermediate frames provide a temporally and spatially coherent progression used for training the model.

(b) Manual Designs: In-the-wild multi-object images often contain severe occlusions or cut-off objects, making it impossible to recover their full true appearance (*e.g.*, notebook in Fig. S.8). Since faithful detail preservation is essential in our task, we request professional designers to curate a small set of ~ 400 manually designed compositions. For each composition, we have access to a separate image of each object, showing its entire appearance. As visualized in Fig. S.9, given a background image and a set of high-resolution images for 2-5 objects, the designers provide a visually appealing image F_K, including all objects on the provided background. We synthesize the short video $F_{1..K}$ by randomly initializing the first frame, placing all unsegmented objects on the background image, and progressively rotate, translate and transform the objects to their final arrangement in F_K. The caption paired with each video is created by following one of a few pre-computed template prompts that describe the desired transformation and scene, completed with a short description of each individual object and background, individually generated using [1].

S.4.2. Subject-Driven Generation Paired Dataset

To preserve the text-image alignment and reposing abilities of the pre-trained I2V base model [15], while encouraging plausible object placements (*e.g.*, a laptop on a table, shoes on the floor), we use a filtered subset of the

Figure S.7. Effect of First Frame Initialization on Final Composition. When generating short videos, the initial object placement in the first frame can be used as rough guidance for the layout in the last frame, when combined with a complementary caption. **(Top):** First frame initialization; **(Bottom):** Last frame of generated 9-frame video, guided by "The vase and the lamp are on the table".

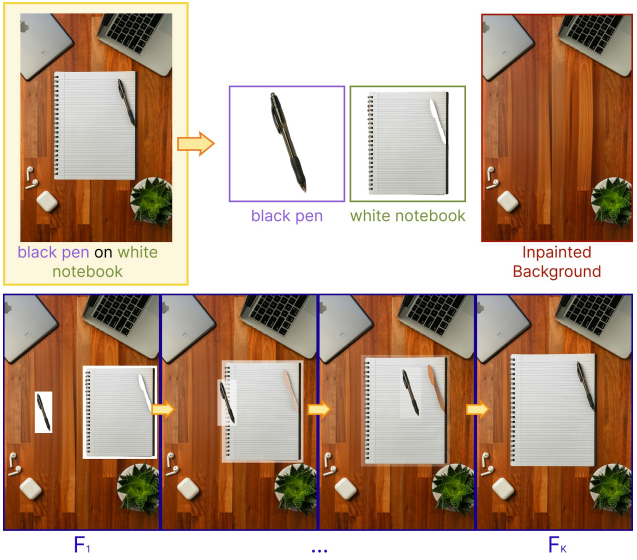

Figure S.8. Training Data Generation Pipeline from In-the-Wild Images. Given an image-caption pair from Unsplash [6], the process extracts relevant objects and an inpainted background. These elements are used to create the $F_{1..K}$ video frames for training.

Subject-200k dataset [14]. Each pair in this subset contains (a) a white-background image of an object and (b) an image of the same object in a different context, together with a descriptive caption. We filter out pairs with object identity changes, often due to AI-generated images, using Grounding-DINO [9]. We retain pairs where object descriptions yield bounding boxes with confidence > 0.55 in both images. For each pair, we use the white-background image as F_1 and the contextual image as F_K for generating short videos as in Fig. S.10. For objects undergoing significant

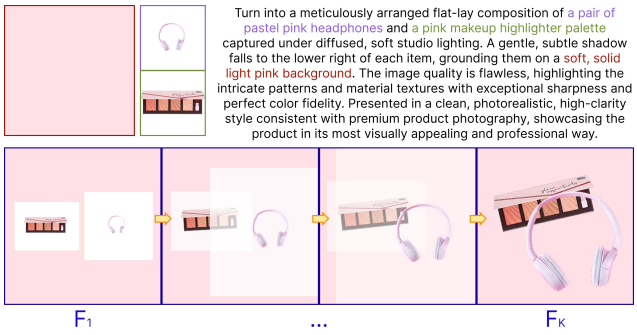

Figure S.9. Training Data Generation Pipeline from Manual Designs. This process transforms a random arrangement of objects on a background image F_1 into a professionally designed layout F_K over a short video sequence. Objects follow synthetic trajectories, while a template-based caption describes the animation.

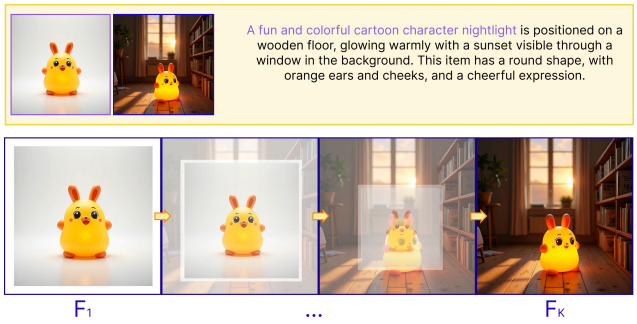

Figure S.10. Training Data Generation Pipeline from a Subject-Driven Generation Paired Dataset. A short animation is built for transforming a white-background object image into an in-context scene described by the accompanying caption. Components in the yellow box are extracted from a Subject-200k subset [3].

reposing between the initial and final frames, we employ frame interpolation to create a smooth transition.

S.4.3. Synthetic Side-by-side Compositions

To enhance our model's ability to realistically rescale and relight objects, we augment our dataset with animations placing objects side-by-side on a shared ground in the final frame. We use 3D-rendered object images similar to those in [4], clustering them into three size groups using K-Means [10]. For each video, we randomly sample two objects from one size group, extract their RGBA images with consistent lighting, and place them with shadows side-by-side on a random background for the final frame F_K. Importantly, when creating F_K, we use the known real-life sizes of the objects to scale them relative to each other, ensuring accurate size relationships. The initial frame F_1 is created by assembling images of the same objects with random lighting conditions. As illustrated in Fig. S.11, a short animation progressively relights, rescales, and repositions the objects from F_1 to F_K. The accompanying caption combines a tem-

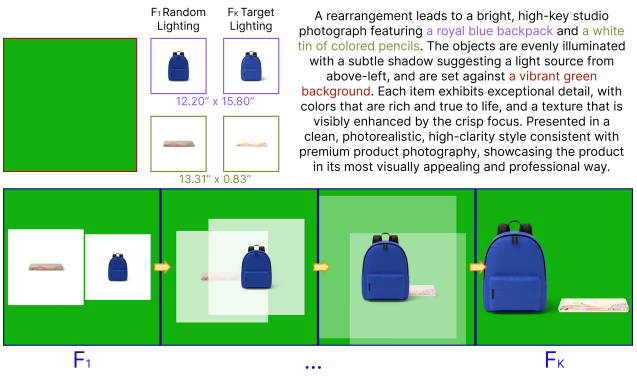

Figure S.11. Training Data Generation Pipeline from Synthetic Side-by-side Compositions. The animation transforms a random arrangement of unsegmented objects with arbitrary scales and lighting on a plain background (F_1) into a coherent natural-looking scene (F_K). The final frame shows objects with consistent lighting and scales based on their real-life sizes. A template-based caption describes this transformation.

plate with object descriptions generated using [1].

S.4.4. Details on Data Augmentation

Object Augmentation. To discourage the model from simply copying the objects from the supplied images $I_{1..N}$, during training, each object is randomly (i) rescaled, (ii) rotated, and (iii) applied a small perspective warping. The resulting object images are scattered around the unchanged background in randomized order to minimize any rearrangement biases, while ensuring all objects are fully present in the image and there is no overlap between them.

Background Augmentation. To make the model resilient to a wide variety of background textures, each training sample using an aleatory background, draws it at random from one of three pools: (i) a collection of high-quality photorealistic or textured background photos that match the aesthetic typically used by professional designers, (ii) plain backgrounds with a randomly sampled RGB color, and (iii) backgrounds created on the fly by combining simple primitives such as linear gradients, block textures, or radial gradients with harmonious color palettes. By mixing these three sources during training, the network learns to operate equally well on simple solid colors, patterned textures, and realistic photographic backdrops, thereby improving its robustness to any background it may encounter at test time.

Scene Completion. During training we randomly apply a "scene-completion" augmentation with a 20% probability. In such cases the selected objects are treated as part of the background: their final positions are incorporated into the background description of the caption, and the background image (if one is provided) is edited to contain those objects throughout the entire K-frame video. The remaining objects are still supplied as explicit conditioning inputs and are animated from their initial random locations to their target positions; in which they may even be partially occluded by the added background objects. This augmentation encourages the model to handle both added objects and items that naturally belong to the scene, improving robustness to varied compositional scenarios and diversity of generation.

Design Elements. With probability 10% we treat a conditioning object as a design element rather than as an explicit visual cue. In that training sample, the object's image I_i is omitted from the visual conditioning set and does not appear in the initial frame of the video. Its description is inserted only in the caption c (outside any $< OBJ > \ldots < /OBJ >$ block). Along the video duration, the object "flies in" from outside the image to reach its target pose in the final frame. This forces the model to rely on the textual description alone for synthesizing that item, encouraging robust text-driven generation and improving editing skills.

Object Replacement. To enable the model to substitute objects in existing scenes, we introduce a replacement augmentation. When none of the objects in a training sample are designated as background elements or design-only elements, with probability 7%, we select one object and substitute it with a new item in the generated animation. Simultaneously, the caption is updated to reflect the new edit.

S.5. Evaluation Dataset

As detailed in Sec. 4, our evaluation set combines object images from the Amazon Berkeley Object Dataset (ABO) [4] and DreamBench++ [12], with backgrounds from Unsplash [6] or plain-color canvases. The set comprises 122 combinations of objects, background images, and descriptive captions. Of these, 44 feature plain-colored backgrounds, while 78 use photorealistic images from Unsplash [6]. The object distribution varies: 17 sets contain a single object, 51 have two objects, 32 include three objects, 15 feature four objects, 6 contain five objects, and 1 set has seven objects. Descriptive captions were initially hand-written to ensure realistic and appealing compositions, then augmented via an LLM [11] to incorporate diverse writing styles, varied compositions, and additional elements (*e.g.*, the flower in Fig. 5 row 3). All figures in the main paper, except for Fig. 4 and Fig. 7, as well as Figs. S.5, S.6 and S.15 to S.17 in this SupMat, showcase inputs from this evaluation set.

S.6. Comparison to State of the Art

We show additional comparisons to state of the art models UNO [18], DSD [2], OmniGen [19], MS-Diffusion [16], VACE [7], NanoBanana [5], Qwen Image Edit [17], and Wan 2.1 [15] in Fig. S.15 (one object compositing), Fig. S.16 (two object compositing), and Fig. S.17 (three object compositing). Additionally, we assess the consistency

Figure S.12. Comparison to Base Model Wan 2.1. Our training adapts Wan 2.1 to multi-object compositing, improving text, color, and background alignment while preserving object identities and enabling new capabilities.

of quantitative metrics by evaluating the full test set over 5 runs and reporting the standard deviation for each metric: CLIP-I (0.0063), DINO (0.0078), CLIP-T (0.0022), BG-MSE (0.0086), Chamfer (0.8370), Missing (0.0089).

As shown in the main paper (Tab. 1), some models such as OmniGen [19] achieve higher identity preservation metrics than our model. However, this can be misleading, as illustrated in Fig. S.15 column 4. In this example, OmniGen's output obtains substantially higher identity preservation metrics (CLIP-I: 0.940, DINO: 0.869) compared to our model (CLIP-I: 0.855, DINO: 0.692). Paradoxically, our model better preserves fine-grained details, such as the duck's beak. This discrepancy arises because our model relights, harmonizes, and slightly reposes the object to integrate it realistically into the scene, while in this case OmniGen simply copy-pastes the object without any significant transformation or background integration. Consequently, while OmniGen maintains higher metric scores by not adapting the object to its new context, our model prioritizes realistic scene integration, which can slightly alter the object's appearance. The effectiveness of our approach is corroborated by the user studies presented in the main paper (Fig. 6), where participants showed a preference for our model in terms of both identity preservation and overall image quality, despite the slightly lower metric scores.

Comparison to Base Model. As reported in Tab. 1, Wan2.1 [15] achieves text alignment comparable to PLACID but struggles with color fidelity and background preservation. Its high identity scores largely reflect copy-paste behavior, as illustrated in Fig. S.12 (1st column).

Figure S.13. Comparison to Image-to-Image (I2I) Models. Qwen Image Edit [17] duplicates objects, while PLACID produces a consistent multi-object composition.

Trained on natural images and designed for a single input image, Wan2.1 is out of distribution for collage-style inputs with multiple objects on a background image, often leading to unsatisfactory results (Fig. S.12, top). Our training strategy and pipeline adaptations (Sec. 4.2, Sec. S.2) provide richer background context and higher-fidelity object representations via concatenated inputs, while special tokens reduce reference ambiguity. This adapts the model to the task of multi-object compositing and extends its generation and editing capabilities, as shown in Fig. S.12, bottom.

Comparison to I2I models. PLACID is based on an image-to-video model (I2V). As part of our comparison to state of the art methods, we consider off-the-shelf image-

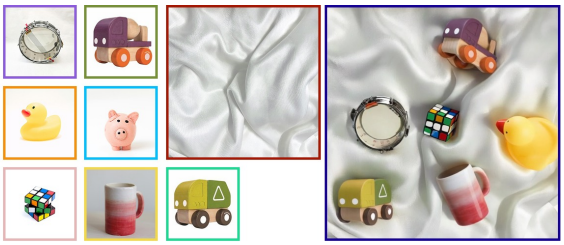

Figure S.14. Model Limitation. Visualization of a failure case where the model struggles to introduce seven new objects into a scene, resulting in one object being omitted.

to-image (I2I) models such as NanoBanana [5] and Qwen Image Edit [17], which are trained on millions of curated I2I samples. Despite this, even these strong I2I models often produce duplicated (Fig. S.13), merged (Fig. 5, row 2), or missing objects (Fig. 5, rows 1,4). In contrast, our video-based approach enforces spatiotemporal consistency when rearranging objects, reducing duplication, merging, or omission while preserving object identities.

S.7. Limitations

We provide additional limitations of our model beyond those highlighted in Sec. 4.4. First, since we use a video model to solve an image-to-image problem, the time and computational requirements at inference are slightly higher than if we used a similar architecture to generate a single image. However, we consider the benefits in terms of model versatility, image quality, completeness, identity and background preservation to outweigh this issue. If necessary, shorter videos could be used to train the model, reducing the gap between I2V and I2I models. Additionally, even though our model can successfully handle compositing a large number of objects in a scene, as shown in Fig. 1 (top, middle), the more inputs and constraints provided by the user, the more challenging the task becomes, resulting in an increased number of failure cases. We illustrate this in Fig. S.14 with an example where the model struggles to adhere to all constraints when attempting to composite seven objects into a natural, cohesive, and appealing display.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 7

[2] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *CVPR*, 2025. 7

[3] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *NeurIPS*, 2023. 1, 5, 6

[4] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, 2022. 6, 7

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7, 9

[6] https://unsplash.com/. 1, 4, 5, 6, 7

[7] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 7

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 4

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 4, 6

[10] James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, 1967. 6

[11] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. 7

[12] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *ICLR*, 2025. 7

[13] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4

[14] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *ICCV*, 2025. 6

[15] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *CoRR*, 2025. 1, 4, 5, 7, 8

[16] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 7

[17] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 7, 8, 9

[18] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 7

[19] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, 2025. 7, 8

[20] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 4

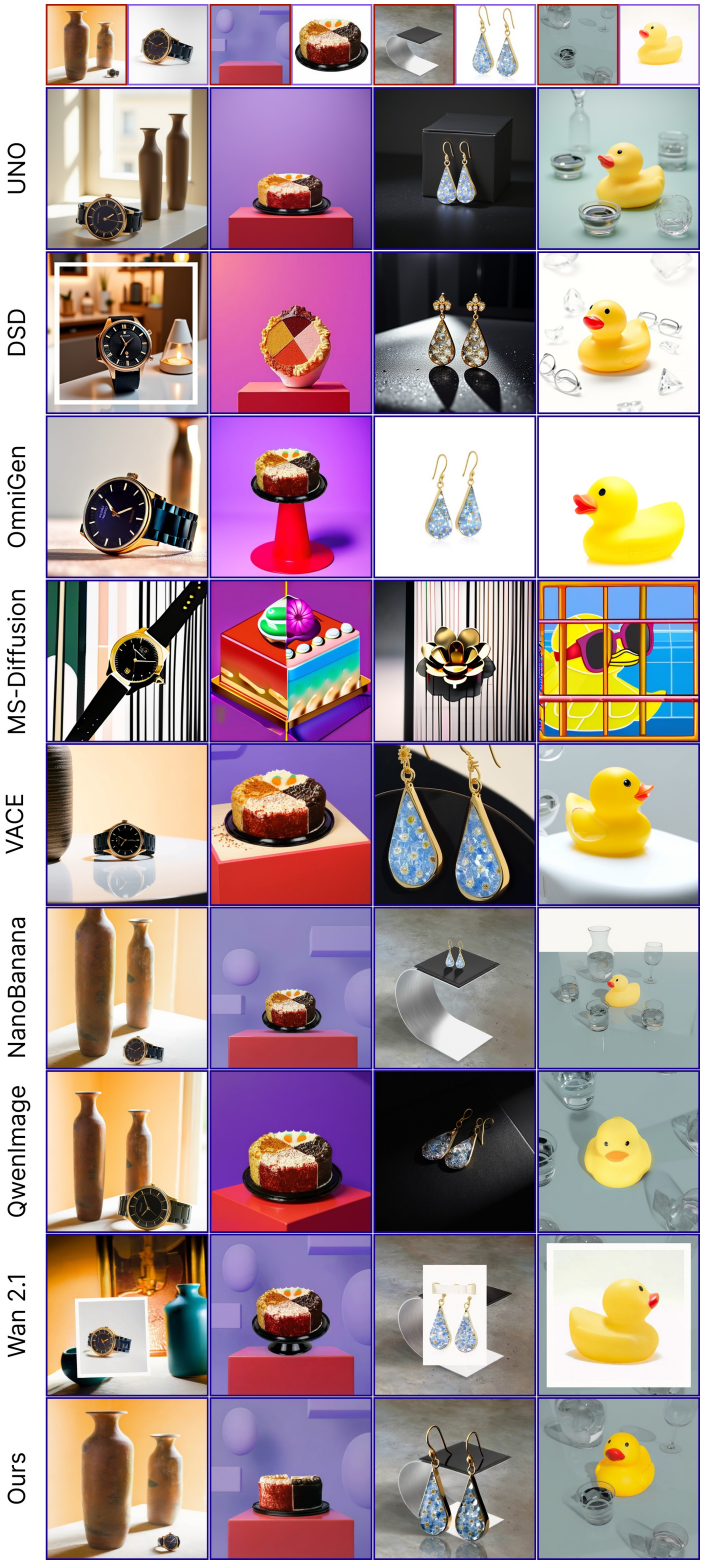

Figure S.15. Comparison to State of the Art. Captions used to guide generation are, from left to right: (i) *"A shiny black and gold watch is clearly displayed on the table next to the front vase and in front of the other."*, (ii) *"A four flavours cake is placed on top of a red platform, in front of a vibrant purple background."*, (iii) *"A gorgeous pair of flowery drop golden earrings is showcased on the black surface in the image."*, (iv) *"A yellow rubber duck is carefully placed in the middle of the image and surrounded by see-through glasses."*

Figure S.16. Comparison to State of the Art. Captions used to guide generation are, from left to right: (i) *"A flat lay display of a pink and white mug and a tall burgundy vase laying down in a diagonal on a luxurious burgundy background."*, (ii) *"A ring with a heart-shaped stone is displayed next to a classy ring with a big diamond in a studio-style display with shadows using a red backdrop."*, (iii) *"A studio-style photo of a silver ring with blue accents and a pair of blue earrings on a blue backdrop. The objects are close to the camera and casting a shadow to the side, laying on the same ground."*, (iv) *"A pair of hoop earrings and a diamond ring are sitting on the table next to the statue."*

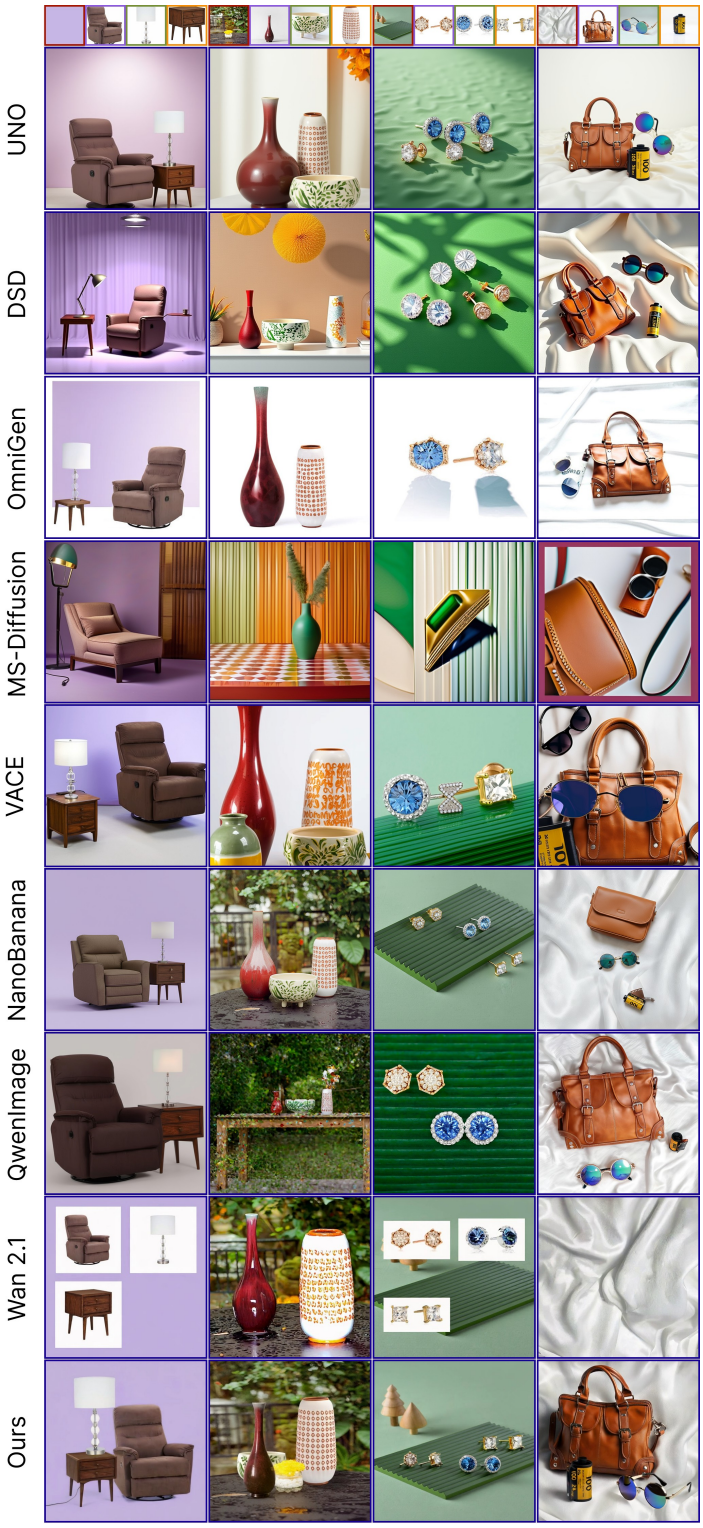

Figure S.17. Comparison to State of the Art. Captions used to guide generation are, from left to right: (i) *"A brown recliner , a lamp and a wooden small table are displayed in a studio-style image in front of a lilac backdrop. A soft light from above illuminates the scene casting gentle shadows."*, (ii) *"A tall red vase, a white and green ceramics element and a tall white vase with orange elements are clearly displayed on the table around and behind the white and yellow decoration."*, (iii) *"A golden shiny pair of earrings, a silver and blue pair of earrings and a golden squared diamond earrings are kept in a diagonal line on top of the wavy green pattern on the green backdrop. They cast a soft long shadow to the bottom right of the image."*, (iv) *"A leather bag, a pair of retro sunglasses and a roll of film are laying down on a satin white surface."*