

A More Experimental Details

In this section, we provide additional experimental details, including the configurations of LoRA and other hyperparameters. For different tasks, we employ distinct settings: Section A.1 describes the spatially-aligned image generation tasks, Section A.2 covers the subject-driven image generation task, and Section A.3 presents the experimental details for style transfer.

DRA-Ctrl employs LoRA [14] to fine-tune the base model with a rank of 16. Since our method needs to simultaneously process noiseless condition image token sequences and noisy generated image token sequences, we set the LoRA scale to 0 when handling the generated image token sequences to distinguish between them. Additionally, we set δ to 12 in the Frame Skip Position Embedding (FSPE). This configuration enables 4 frames in the latent space to effectively emulate 37 frames, corresponding to $1 + 36 \times 4 = 145$ frames in pixel space — approximately equivalent to a 5-second short video at 30 frames per second (fps), which sufficiently achieves the shot transition effect.

A.1 Spatially-aligned Image Generation

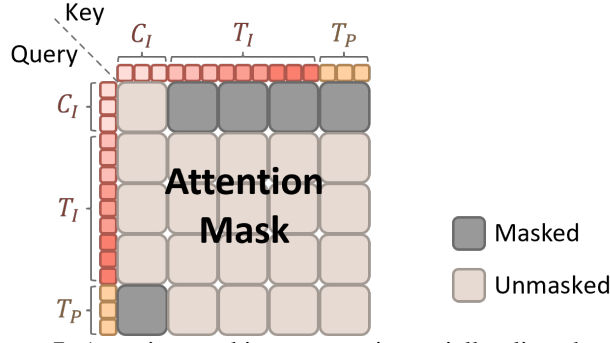


Figure 7: Attention masking strategy in spatially-aligned tasks.

In spatially-aligned image generation tasks, the condition image is directly extracted from the ground-truth image without a corresponding prompt. Therefore, we do not employ the condition image prompt C_P in our experiments, but we still utilize the attention masking strategy, with the corresponding attention mask illustrated in Figure 7. Besides, we train the model for 6,000 steps. In depth-to-image and depth prediction tasks, the depth image is extracted from the ground-truth image using Depth Anything [52]. For the depth prediction task, we prepend “[depth]” to the prompt to guide the model to generate depth maps rather than regular images. In the deblurring task, we apply Gaussian blur to the images with a randomly selected integer blur radius between 1 and 10 during training. For the in/out-painting task, we randomly select a rectangular region in the image during training, then mask either the selected region (with 0.5 probability) or the area outside it (with 0.5 probability) to create the condition image. In the super-resolution task, the condition image is obtained by downsampling the original image by a factor of 4.

A.2 Subject-driven Image Generation

For the subject-driven image generation task, we train the model for 9,000 steps. During inference, while employing attention masking, the simultaneous presence of both target image prompts T_P and condition image prompts C_P may still cause information blending. To address this, we strengthen the interaction between target image tokens T_I and T_P while suppressing C_P ’s influence on the generated output. Specifically, within the $(T_I \times T_P)$ attention mask region, we augment the attention weights by adding $0.6 \times \mu$ (where μ denotes the mean absolute value of the original weights). The modified attention computation for this region is formulated as:

$$\text{Attention}(Z) = \text{softmax} \left(\frac{Q_Z K_Z^\top}{\sqrt{d}} + 0.6 \times \text{mean} \left(\left| \frac{Q_Z K_Z^\top}{\sqrt{d}} \right| \right) \right) V_Z. \quad (4)$$

A.3 Style Transfer



Figure 8: Bitmoji-style example images in our dataset.

[USER PROMPT]:

将上传的图像分别转换为 bitmoji 风格，尺寸大小为{}:{}，输出清晰的图像。

Figure 9: The prompt format used for generating Bitmoji-style images with GPT-4o.

We collected 100 diverse images containing subjects such as humans, animals and buildings from the web. Using carefully designed prompts, we guided ChatGPT-4o to generate corresponding Bitmoji-style images, which formed our training set. The subject-driven image generation model is fine-tuned for 2,600 steps with a batch size of 8 on an NVIDIA H800 GPU to obtain the final model. Example images from our dataset are shown in Figure 8, and the prompt format we employed is shown in Figure 9, where the image dimensions are determined by their original resolutions.

B More Details about the VL Score

Current evaluation metrics for subject-driven image generation primarily employ DINO and CLIP-I to assess subject consistency, and CLIP-T for prompt adherence. However, two critical limitations exist: first, there lacks a comprehensive metric to directly evaluate subject-driven generation quality; second, these existing metrics exhibit notable shortcomings — both DINO and CLIP-I are significantly influenced by background interference, while CLIP-T struggles with fine-grained semantic alignment.

To address these issues, we propose leveraging an advanced Vision-Language (VL) model, such as QWen2.5-VL [1], as an evaluator to produce a holistic metric. Our approach consists of three steps: First, we provide the VL model with a prompt instructing it to score (prompt, reference image, generated image) triplets based on multiple fine-grained criteria for both subject consistency and prompt adherence. Next, we have the model summarize its task to confirm proper understanding. Finally, we input each triplet and collect the model’s scores. Since both metrics are discrete scores ranging from 0 to 4, we average them to derive a comprehensive metric termed the VL Score. An example input-output demonstration of the VL model is shown in Figure 10.

C More Visualization

This section presents additional qualitative experimental results across all tasks, including transition frames generated by our model. The spatially-aligned image generation results are detailed in Section C.1, while the subject-driven image generation outcomes are presented in Section C.2, and the style transfer performance is analyzed in Section C.3. Unless otherwise specified, all image generation in this paper uses 50 sampling steps by default, including both qualitative results and quantitative evaluations, and generated images maintain a consistent resolution of 512×512 pixels.

```

[USER PROMPT]:
[Task Instruction] As a professional evaluator, your role is to assess the semantic consistency between a
↪ reference image and an generated image based on a given text prompt. Focus on two aspects:
1. Subject Consistency: Compare key visual attributes (Shape: main body outline, structure, and
↪ proportions. Color: saturation, hue, brightness, and distribution. Texture: realism, clarity, and
↪ aesthetic appeal. Facial: facial features, only apply when subject is a person or animal) between the
↪ reference and generated image.
2. Prompt Adherence: Verify if the generated image accurately reflects all elements described in the text
↪ prompt (Relevance: whether core elements in text are shown. Accuracy: objects, settings, actions, or
↪ characteristics Completeness: whether the generated image omits significant details. Context:
↪ portrayal of the appropriate environment, interactions, and background elements).
You need to provide your scores Subject Consistency and Prompt Adherence based on the following criteria.
[Subject Consistency Scoring Criteria] Very Poor (0): No resemblance. The generated image's subject has
↪ no relation to the reference.
Poor (1): Minimal resemblance. The subject falls within the same broad category but differs
↪ significantly.
Fair (2): Moderate resemblance. The subject shows likeness to the reference with notable variances.
Good (3): Strong resemblance. The subject closely matches the reference with only minor discrepancies.
Excellent (4): Near-identical. The subject of the generated image is virtually indistinguishable from the
↪ reference.
[Prompt Adherence Scoring Criteria] Very Poor (0): No resemblance. The generated image's subject has no
↪ relation to the reference.
Poor (1): Minimal resemblance. The subject falls within the same broad category but differs
↪ significantly.
Fair (2): Moderate resemblance. The subject shows likeness to the reference with notable variances.
Good (3): Strong resemblance. The subject closely matches the reference with only minor discrepancies.
Excellent (4): Near-identical. The subject of the generated image is virtually indistinguishable from the
↪ reference.
[Input Format] Every time you will receive two images and a prompt, the first image is a reference image,
↪ and the second image is the generated image, and the text is the prompt describing the generated
↪ image.
[Output Format] return the scores (0-4) in JSON format, strictly following this example:
↪ {"subject_consistency": 3, "prompt_adherence": 4}.

[VL MODEL OUTPUT]:
Sure, I understand the task. Please provide me with the reference image, the generated image, and the
↪ text prompt so that I can evaluate them according to the provided criteria.

[USER PROMPT]:
Did you understand the task above? Please summarize the tasks you need to do and show how you will
↪ execute the detailed plan for the task.

[VL MODEL OUTPUT]:
Yes, I understood the task clearly. too long, omitted

[USER PROMPT]:
{the first image} {the second image} {prompt}

[VL MODEL OUTPUT]:
```json\n{"subject_consistency": 3, "prompt_adherence": 2}\n```

More...

```

Figure 10: An example of VL Score evaluation process.

## C.1 Spatially-aligned Image Generation Results

Our method performs image-to-video generation conditioned on input images, where the state of these condition images significantly impacts the output quality. We found that directly using canny edges, depth maps with black representing maximum depth, or black masks in in/out-painting tasks often resulted in unnaturally dark generated images. To address this, we implemented a color normalization scheme that remaps the darkest values (0, 0, 0) to medium-gray (128, 128, 128) while linearly scaling all other color values proportionally, preventing extreme darkening.



### C.1.1 Canny-to-image



Figure 11: More canny-to-image generation results.

### C.1.2 Colorization

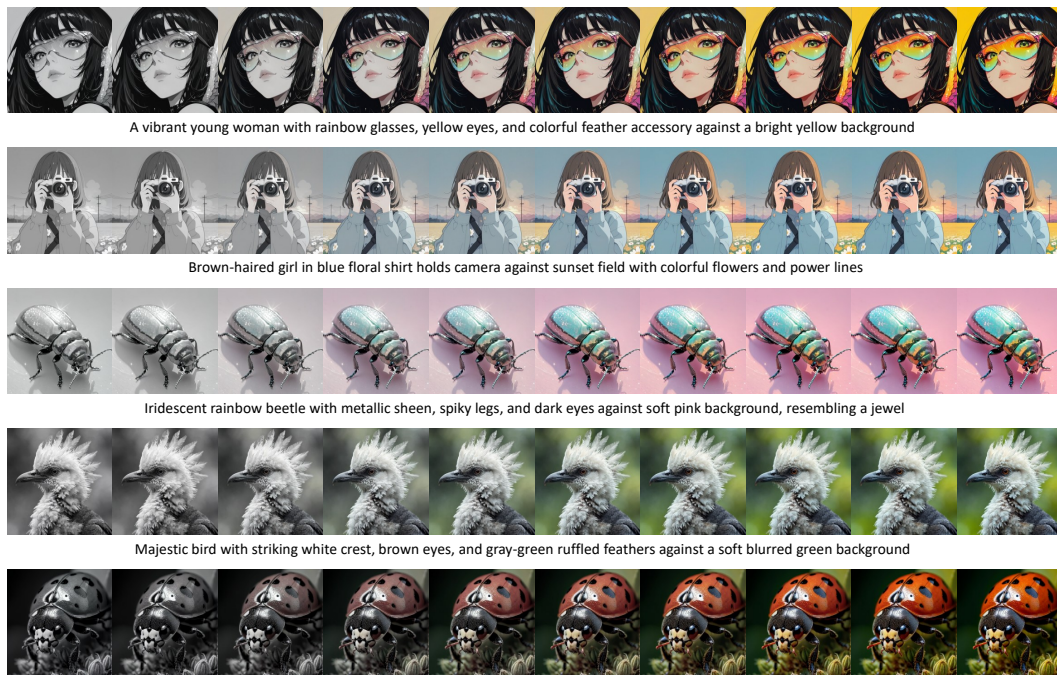


Figure 12: More colorization generation results.



### C.1.3 Deblurring

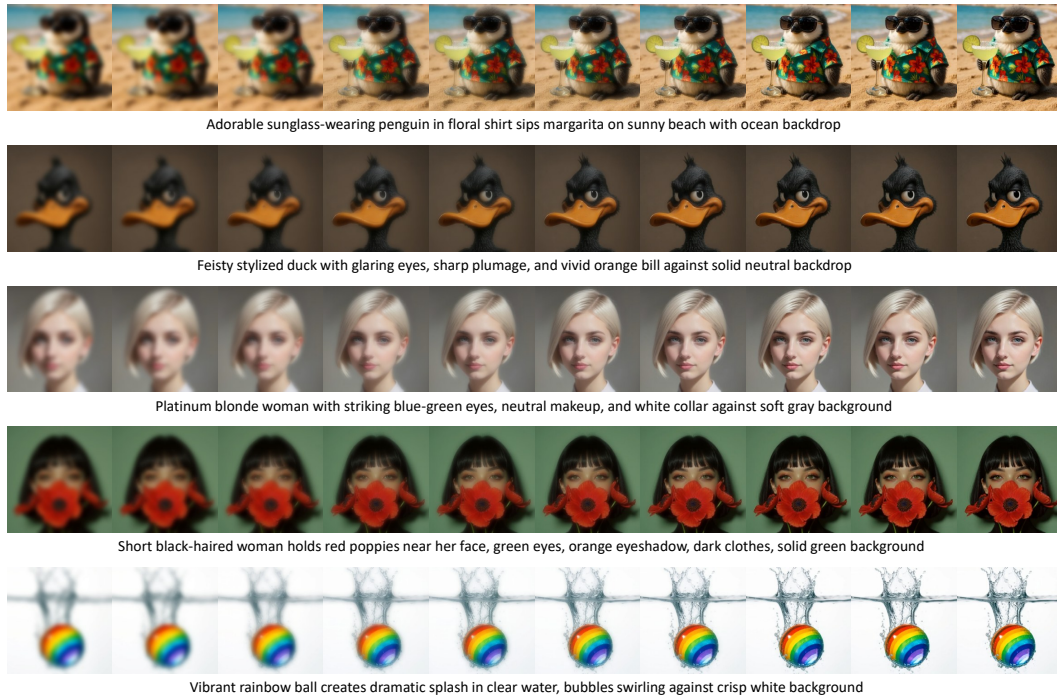


Figure 13: More deblurring generation results.

### C.1.4 Depth-to-image

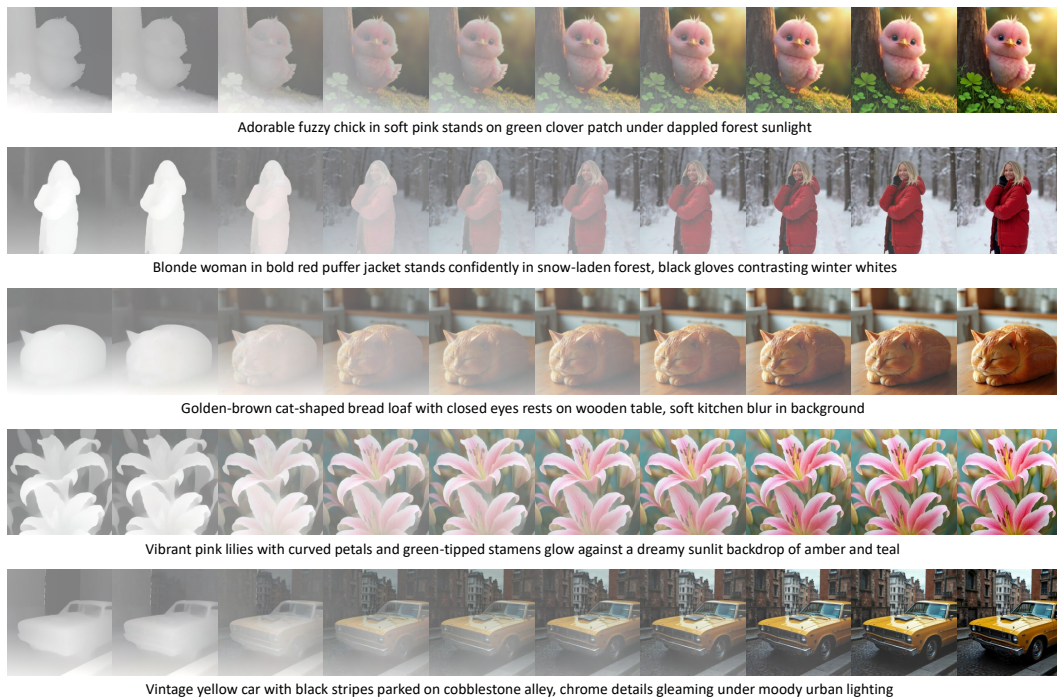


Figure 14: More depth-to-image generation results.

### C.1.5 Depth Prediction

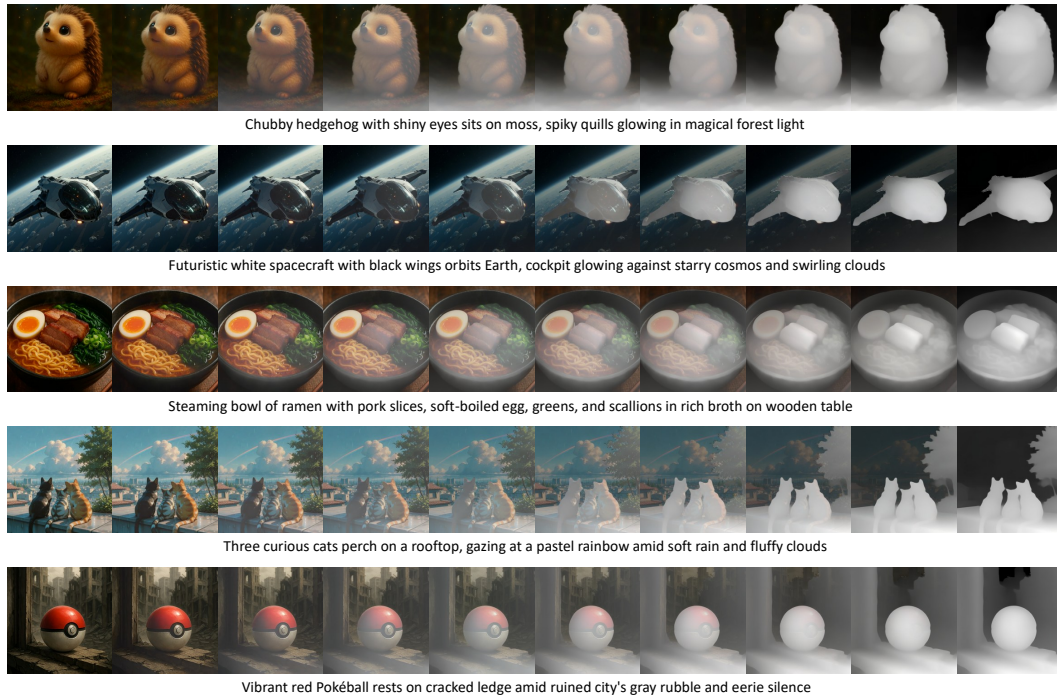


Figure 15: More image-to-depth generation results.

### C.1.6 In/out-painting

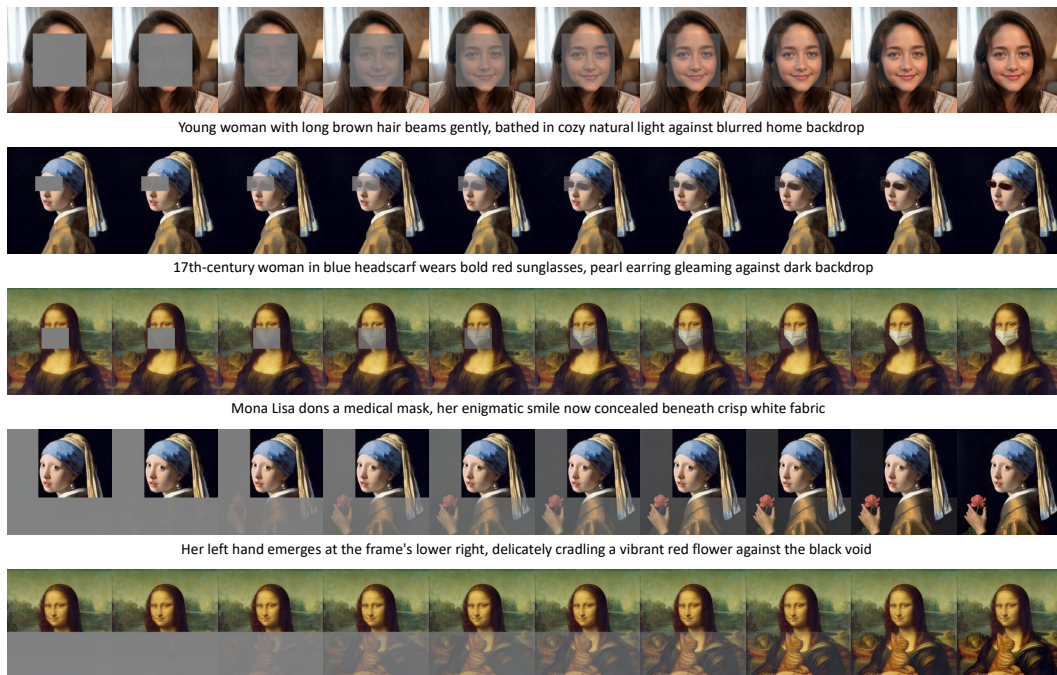


Figure 16: More in/out-painting generation results.



### C.1.7 Super-resolution

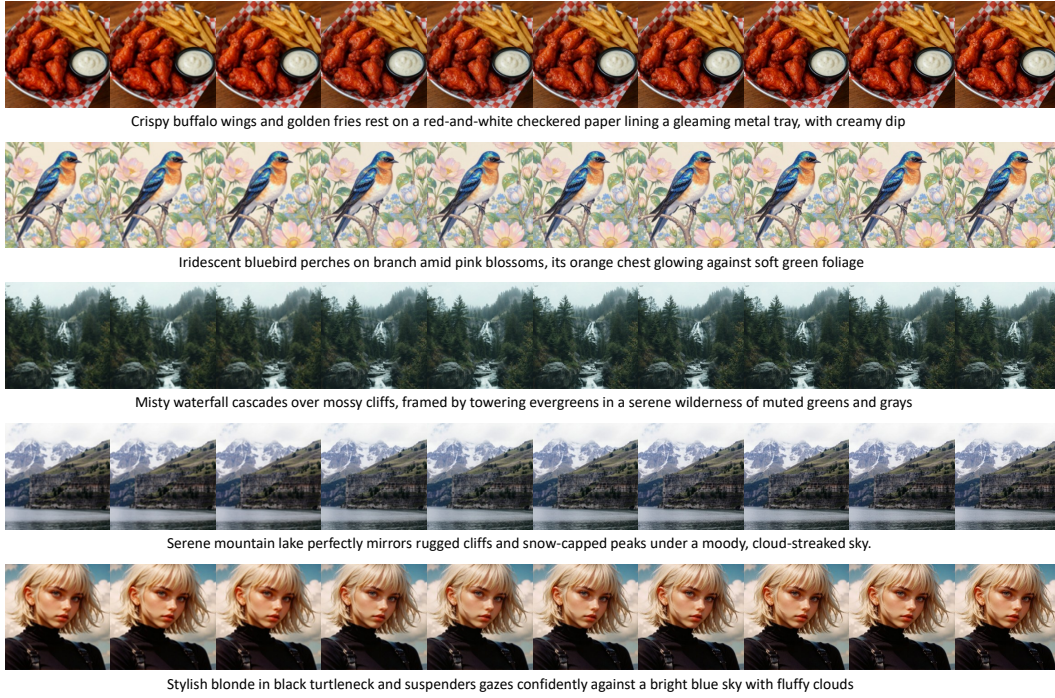


Figure 17: More super-resolution generation results.

### C.2 Subject-driven Image Generation Results

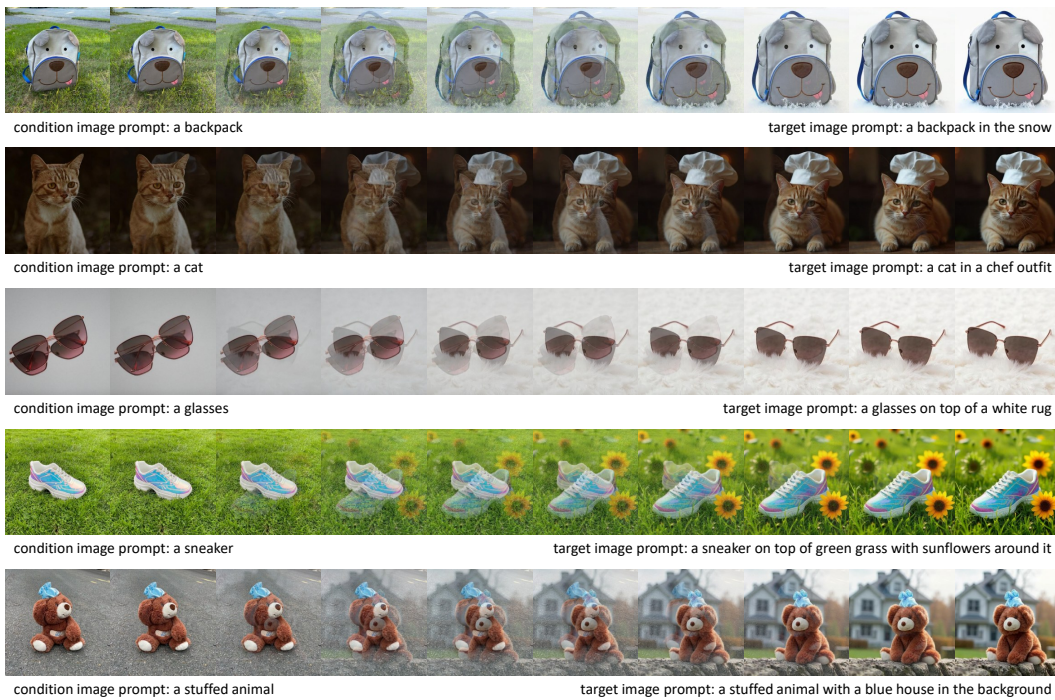


Figure 18: More subject-driven generation results on DreamBench.





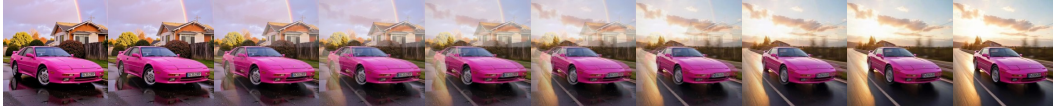
condition image prompt: A wooden violin rests on the ground beside flowers and a clock  
target image prompt: A wooden violin lies on sandy beach by the ocean



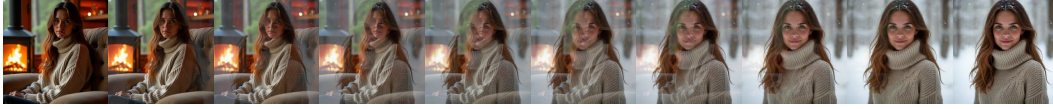
condition image prompt: Chair with white leather cushions and smooth wood grain, angled legs on minimalist gray backdrop  
target image prompt: Chair before floor-to-ceiling windows, skyscrapers glowing through glass as sunlight traces its polished frame



condition image prompt: Tiger sits politely on wooden chair beside stacked pancakes and cream container, gazing upward indoors  
target image prompt: Cool tiger with sunglasses sprawls in sunny grass, beside stacked pancakes



condition image prompt: Pink sports car parked on wet road, rainbow arching over suburban house with autumn trees and glistening raindrops  
target image prompt: Pink sports car streaks down sunlit highway, silver rims flashing, silhouette slicing through golden summer air



condition image prompt: Woman in cream knit sweater sits calmly by a crackling fireplace, surrounded by warm candlelight and rustic wooden shelves  
target image prompt: The woman stands in a snowy forest, captured in a half-portrait

Figure 19: More subject-driven generation results.

Interestingly, we discover that during subject-driven image generation, DRA-Ctrl can occasionally control two subjects in the condition image simultaneously. As shown in the third row of Figure 19, our method successfully makes the tiger wear sunglasses while placing the stacked pancakes on the grass.

### C.3 Style Transfer

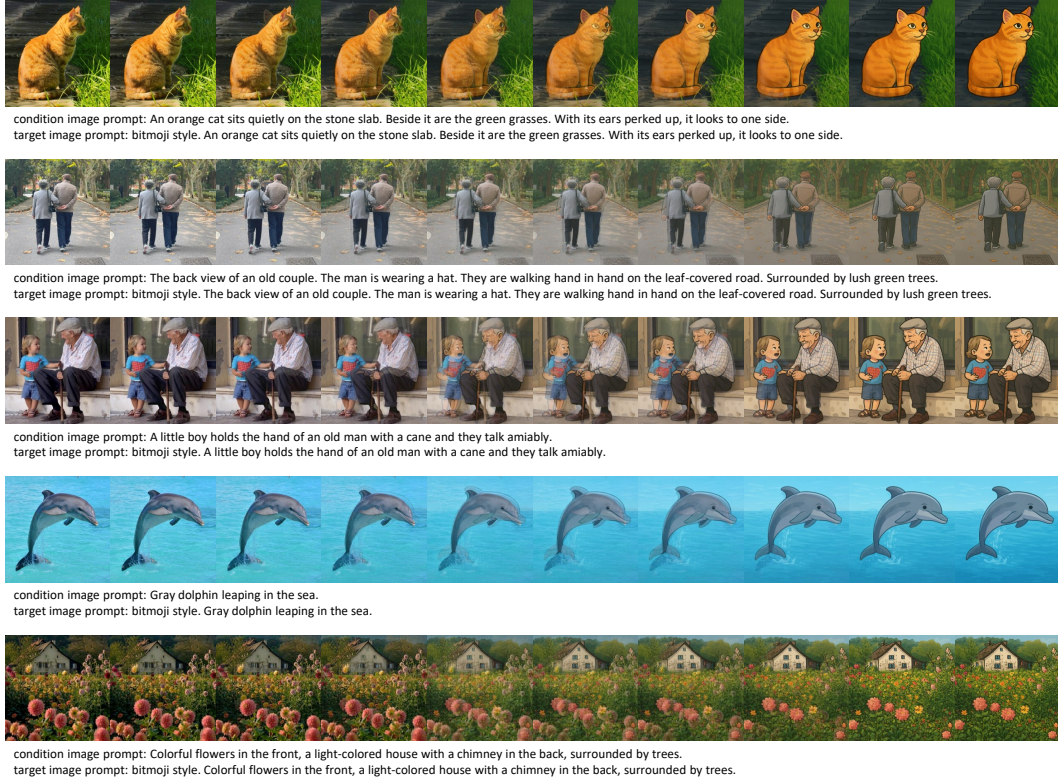


Figure 20: More style transfer generation results.

### D Failure Cases

While DRA-Ctrl successfully achieves controllable image generation in most cases, it may occasionally fail in the image-to-depth task, primarily manifesting as the presence of colored regions in the generated depth images. We attribute this limitation to the inherent nature of video models, which predominantly generate color data. A failure case is presented in Figure 21.

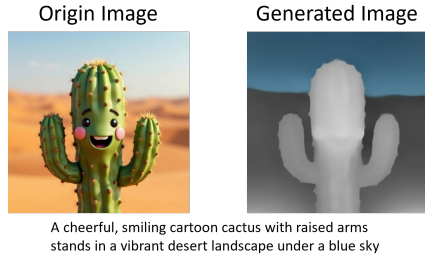


Figure 21: A failure case of DRA-Ctrl.

### E Societal Impact

Our work advances controllable image generation with significant societal implications, offering both opportunities for innovation and risks requiring proactive mitigation. Below, we outline the potential positive and negative impacts, alongside measures to address the latter.

On the positive side, our high-quality, controllable generation method empowers creative and practical applications. Artists and designers can leverage it to produce imaginative content efficiently, while

educators benefit from dynamically generated visual aids for teaching. The fine-grained control also enables ethical uses in journalism and advertising, enhancing productivity and accessibility across domains.

However, negative impacts must be acknowledged. Malicious actors could exploit the technology to create convincing fake images for disinformation, fraud, or impersonation; to mitigate this, we adopt a gated release of models to restrict access. Bias in training data might lead to stereotypical or discriminatory outputs, disproportionately harming marginalized groups — addressed through rigorous bias testing during development. Further, misuse for non-consensual imagery (e.g., deepfakes) necessitates monitoring mechanisms and legal safeguards to protect privacy.

In summary, while our technology unlocks creative and educational potential, its risks—particularly around misinformation, bias, and privacy—demand deliberate countermeasures. By combining technical safeguards with policy-oriented solutions, we aim to foster responsible use and maximize societal benefit.

## **F Safeguards**

To mitigate potential misuse risks associated with our controllable image generation technology, we will implement a gated release strategy when making the models publicly available. This will include: comprehensive usage guidelines explicitly prohibiting malicious applications such as disinformation campaigns and non-consensual imagery generation; an access control mechanism requiring users to agree to ethical use terms before obtaining the model. While we recognize no safeguards can eliminate all risks, these measures represent our proactive commitment to responsible AI development and deployment.