EasyControl: Adding Efficient and Flexible Control for Diffusion Transformer

Supplementary Material

A. Preliminary: Diffusion Transformer

Currently, The state-of-the-art text-to-image Diffusion model architectures are based on Diffusion Transformers (DiT)[50], including models such as SD3[10], FLUX[31]. These models integrate diffusion processes with Transformer architectures to improve text-to-image generation, yielding high-quality and accurate text-to-image synthesis.

Our approach is based on the FLUX.1 pre-trained model, which consists of three key components: T5 as the text encoder, a VAE for image compression, and a Transformer-based denoising network. Specifically, The denoising network divides the latent noise into several patches and treats each patch as a noise token, denoted as $X \in \mathbb{R}^{N\times d}$, where N is the number of noise tokens and d is the dimensionality of each token. To effectively introduce spatial position information in different noise patches, FLUX.1 employs rotary position encoding (RoPE)[75] to encode spatial information within each noise patch. Meanwhile, the text prompts are encoded into text tokens $C_T \in \mathbb{R}^{M\times d}$ through the T5 text encoder, where M represents the number of text tokens. These image and text tokens are then fused by concatenation to form a joint representation. After feature fusion, the image-text token sequence is fed parallel into a transformer-based denoising model. The model iteratively denoises the image, progressively restoring a clear image and ultimately generating a high-quality image that aligns with the textual description.

B. Position Encoding Offset

The PE offset strategy was proposed by method[76], which applies a fixed displacement to position encodings and was proved to lead to faster convergence. This offset is uniform across all encodings within the subject condition image. In our experiments, we set this offset to 64 in the height dimension. Mathematically, for each position encoding $PE(i,j)$ in the subject condition image, the adjustment is:

$$PE(i,j) \leftarrow PE(i,j) + \Delta_h \cdot \mathbf{e}_h \qquad (16)$$

where $\mathbf{e}_h$ is the unit vector along the height dimension, and $\Delta_h = 64$ ensures distinct separation between spatial and subject conditions.

C. Trianing Data

For spatial control tasks such as depth, canny, and OpenPose, we employ the MultiGen-20M dataset[98] as our primary training resource. Regarding subject control, our training is conducted using the Subject200K dataset[76]. For face control, we utilize a curated subset of the LAION-Face dataset[99], supplemented by a collected private multi-view human dataset (See Fig 6), where all human images are preprocessed through InsightFace[9] for precise cropping and alignment to ensure consistency and accuracy in our training inputs.

Figure 6. Visualization of samples in private Multi-view Human Dataset.

D. Details of KV Cache

More details of KV Cache for Efficient Conditional Image Generation is shown in the algorithm 1.

E. Single Condition Quantitative Comparison

Settings. In this section, we compare our method with *Controlnet*[93], *OmniControl*[76], and *Uni-Control*[98] using two types of conditioning: Depth and Canny. For the subject condition, We compare our method with *OmniControl*[76], *IP-Adapter(IPA)*[89], and *Uni-ControlNet*[98]. (To ensure a fair and consistent comparison, all methods are implemented using FLUX.1 dev as the base model, with configurations and parameters sourced from publicly available official[32, 76, 93] and community resources[27, 85] with recommended parameters, while Uni-ControlNet employs its official implementation[98] based on the SD1-5 architecture.)

Data. For the Depth map and Canny edge conditions, comprehensive evaluations were conducted on the COCO

Condition	Method	Controllability	Generative Quality		Text Consistency
		F1 ↑ /MSE ↓	FID ↓	MAN-IQA ↑	CLIP-Score ↑
Canny	ControlNet	0.232	20.325	0.420	0.271
	OminiControl	**0.314**	17.237	0.471	0.283
	Uni-ControlNet	0.201	17.375	0.402	0.279
	Ours	0.311	**16.074**	**0.503**	**0.286**
Depth	ControlNet	1781	23.968	0.319	0.265
	OminiControl	1103	**18.536**	0.431	0.285
	Uni-ControlNet	1685	21.788	0.423	0.279
	Ours	**1092**	20.394	**0.469**	**0.289**

Table 2. Quantitative comparison with baseline methods on single condition tasks.

Condition	Method	Identity Preservation		Generative Quality		Text Consistency
		CLIP-I ↑	DINO-I ↑	FID ↓	MAN-IQA ↑	CLIP-Score ↑
Subject	IP-Adapter	**0.700**	0.429	79.277	0.511	0.266
	OminiControl	0.663	**0.445**	72.298	0.579	0.276
	Uni-ControoNet	0.641	0.417	86.369	0.439	0.204
	Ours	0.667	0.443	**71.910**	**0.595**	**0.283**

Table 3. Quantitative comparison with baseline methods on single condition tasks.

Condition	Method	ID Preservation	Controllability	Generative Quality		Text Consistency
		Face Sim. ↑	MJPE ↓	FID ↓	MAN-IQA ↑	CLIP-Score ↑
Openpose+Face	ControlNet+IPA	0.049	166.7	227.06	0.229	0.156
	ControlNet+Redux	0.027	141.5	200.70	0.293	0.217
	Uni-ControlNet	0.048	258.8	203.31	0.481	0.147
	ControlNet+InstantID	0.521	83.9	203.17	0.345	0.250
	ControlNet+PhotoMaker	0.343	86.3	213.83	0.420	0.281
	ControlNet+Uni-portrait	0.456	46.0	203.07	0.564	0.253
	Ours	**0.530**	**36.7**	**184.93**	**0.586**	**0.285**

Table 4. Quantitative comparison with baseline methods on multi-condition tasks.

2017[39] validation set comprising 5,000 images. All generated outputs strictly preserved the original image resolutions and aspect ratios, with textual prompts derived from the corresponding ground-truth captions of the dataset. For subject control scenarios, we adopted the Concept-101 benchmark dataset[30] to assess model performance. Each reference image was paired with its semantically aligned textual description as the conditioning prompt.

Metrics. To comprehensively evaluate the performance of each algorithm, we assess four key aspects: *1. Controllability:* We extract structural information from the generated images using the corresponding structure extractor, obtaining the structure map. The F1 Score is computed between the extracted and input edge maps in the edge-conditioned generation, and the MSE is calculated between the extracted and original condition maps for the depth task. *2. Text Consistency:* We use the CLIP-Score[18, 55] to evaluate the consistency between the generated images and the input text. *3. Generative Quality:* The diversity and quality of the generated images are assessed using FID[19], MAN-IQA[87].*4. Identity Preservation:* For the subject condition, we use CLIP-I[55] and DINO-I[4] to evaluate identity preservation. Specifically, CLIP-I computes the cosine similarity between image embeddings extracted by the CLIP image encoder for both generated and reference images. Similarly, DINO-I measures identity preservation by calculating the cosine similarity of image embeddings obtained through the DINO encoder framework.

Quantitative Analysis. As shown in Table 2, our proposed method demonstrates superior performance across multiple evaluation metrics. Under the Canny condition, our approach outperforms all comparative methods in terms of generation quality and text consistency, while achieving the second-highest score in controllability. In the depth condition, our method exhibits dominant performance in both controllability and text consistency. Regarding generation

Algorithm 1 KV Cache for Efficient Conditional Image Generation

```
Require:
  Conditional features $\{\text{cond}_i\}_{i=1}^{m}$ for $m$ conditions
  Denoising steps $T = \{t_0, t_1, ..., t_T\}$
Ensure: Generated image $x_0$ with reduced latency
  Initialize KV cache dictionary $\mathcal{D} \leftarrow \emptyset$
  Generate initial noisy image $x_N \sim \mathcal{N}(0, I)$
  for timestep $t \in \{t_T, t_{T-1}, ..., t_0\}$ do
      if $t = t_T$ (First step) then
          for each condition branch $i \in \{1, ..., m\}$ do
              Compute keys/values: $K_{C_i}, V_{C_i} = f_\theta(\text{cond}_i)$
              Cache KV pairs: $\mathcal{D}[i] \leftarrow (K_{C_i}, V_{C_i})$
          end for
      end if
      Retrieve cached KV pairs: $\{(K_{C_i}, V_{C_i})\}_{i=1}^{m} \leftarrow \mathcal{D}$
      Compute self-attention using current noise and text features: $Q_{\text{denoising}}, K_{\text{denoising}}, V_{\text{denoising}} = f_\theta(\mathrm{x}_t, t)$
      Fuse conditions via cached K/V pairs:
          $$Q = Q_{\text{denoising}},$$
          $$K = \text{Concat}\left(K_{\text{denoising}}, K_{C_1}, \ldots, K_{C_m}\right),$$
          $$V = \text{Concat}\left(V_{\text{denoising}}, V_{C_1}, \ldots, V_{C_m}\right)$$
          $$\text{Output} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$
      Update latent: $x_{t-1} = \text{Denoise}(x_t, t, \text{Output})$
  end for
  return Final image $x_0$
```

quality, while our method ranks second position in the FID metric, it achieves first according to the MAN-IQA metric. These comprehensive results substantiate the superiority of our approach across most evaluation criteria, particularly highlighting its exceptional performance in controllability, generation quality, and text consistency. As shown in Table 3, Under the Subject condition, our approach outperforms all comparative methods in terms of generation quality and text consistency and achieves competitive results on identity preservation.

Qualitative Analysis. We show some results about spatial control in figure 9. Under identical conditional input configurations, both ControlNet and OminiControl demonstrate significant blurring artifacts in the synthesized images. In contrast, our framework consistently preserves superior visual fidelity across all evaluated scenarios. This qualitative advantage is particularly pronounced in the preservation of fine-grained details and structural integrity, thereby substantiating the enhanced performance of our Position-Aware Training Paradigm. We have also visualized several subject control results in Figure 10 to demonstrate the effectiveness of our method in terms of identity preservation, generative quality, and text consistency. (The prompts utilized in the generated images include: *in the forest, in the library, on a snow-covered mountain, in the city, in a room, in front of a castle, floating on water, on the beach, on a mountain, in the desert, and on a snowy day.*)

F. Multi-Condition Quantitative Comparison

Settings. In this section, we conduct comparisons using face + OpenPose as multi-condition configurations, against several plug-and-play baseline methods including: *Controlnet+IP-Adapter*[89], *Controlnet+Redux[32]*, and *Uni-Controlnet*[98] and several SOTA play-and-plug identity customization methods, including *Controlnet+InstantID[78]*, *Controlnet+PhotoMaker[37]*, and *ControlNet+Uni-portrait[17]*. (For several plug-and-play baseline methods, we utilize official FLUX-based implementations such as OminiControl and community-driven implementations including ControlNet, IPA, and Redux. For identity customization methods, we adopt official implementations based on SD1-5 (e.g., Uni-portrait) and SDXL (e.g., PhotoMaker, InstantID) as base models, along with their corresponding ControlNet modules.)

Data. For evaluation, we constructed a comprehensive dataset comprising three components: (1) 1,000 randomly sampled face images from the FFHQ dataset for face control inputs; (2) 1,000 full-body or half-body human images crawled from Laion face dataset[99], from which OpenPose information was extracted for spatial control inputs; and (3) 1,000 text prompts generated by GPT, each describing a person with specific characteristic and locations. Each algorithm generated 1,000 images based on these inputs for evaluation. This diverse dataset ensures a thorough assessment of the models' capabilities in handling various control conditions.

Metrics. To comprehensively evaluate the performance of each algorithm, we assess several key aspects: *1. Controllability:* The Mean Joint Position Error (MJPE) metrics computed between the extracted and input openpose maps in the pose generation. *2. Text Consistency:* We use the CLIP-Score to evaluate the consistency between the generated images and the input text. *3. Generative Quality:* The diversity and quality of the generated images are assessed using FID[19], MAN-IQA[87]. *4. Identity Preservation:* For the face condition, we use face similarity[9] to evaluate identity preservation. For the OpenPose condition, our controllability metric is quantitatively assessed through the Mean Joint Position Error (MJPE) metrics, which measure the spatial consistency between the generated image and the input OpenPose map. The evaluation procedure involves three sequential steps: initially, key point information is extracted from both the generated image and the input condition using OpenPose. Subsequently, The Euclidean distance for each joint is computed and averaged to obtain the MJPE for a single image. This process is repeated for all images in the test set, and the average MJPE serves as the model's controllability metric. By quantifying joint position deviations, MJPE effectively evaluates the consistency between generated images and input conditions. It is noteworthy that a lower MJPE indicates superior spatial alignment and better pose consistency between the generated image and the input condition, thus reflecting higher controllability in the pose generation process.

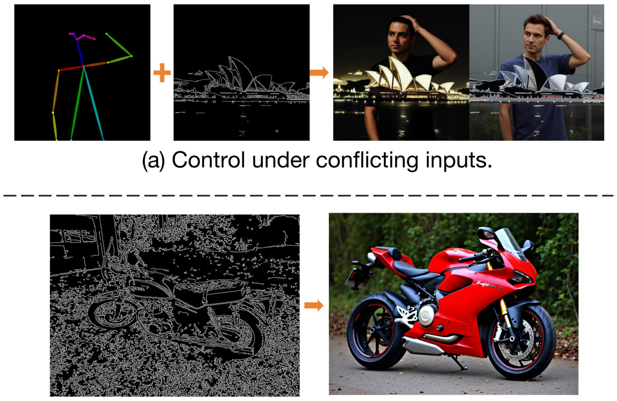
F.1. Quantitative Comparison

The quantitative results are presented in Table 4. Our method achieves state-of-the-art performance across all metrics. Specifically, it obtains the best Face Similarity, demonstrating superior ID preservation. For controllability, our approach achieve the lowest MJPE score and significantly outperforms others. In terms of generative quality, our method achieves the lowest FID and highest MAN-IQA, indicating better image quality and diversity. Additionally, it maintains strong text consistency with the highest CLIP score. These results collectively demonstrate the effectiveness of our framework in balancing control precision, identity preservation, and generation quality under a multi-condition combination.

It is noteworthy that in Table 4, certain algorithms exhibit significantly inferior performance in terms of Face Similarity (Face Sim) and Mean Joint Position Error (MJPE) metrics compared to other methods. This is primarily attributed to the fact that many competing methods fail to effectively transfer facial or pose features from the control images, often resulting in generated images that are blurry, distorted, or lack detectable facial or pose features. Consequently, these methods are unable to accurately compute the metrics required for face similarity or pose alignment. In contrast, our approach ensures robust feature transfer and precise alignment, enabling the generation of high-quality images with clearly detectable facial and pose attributes, which contributes to the superior performance reflected in the metrics.

F.2. Visual Comparison

As illustrated in the figure 8, we present a visual comparison with ID customization methods. Our method demonstrates superior performance in facial similarity, controllability, and image quality compared to other approaches. This indicates that our framework, despite being trained on single conditions, exhibits strong plug-and-play adaptability, effectively integrating multiple conditions without conflicts. In contrast, other methods often suffer from incompatibility between different modules, leading to degraded facial similarity, controllability, and poor generation quality.


(a) Control under conflicting inputs.

(b) Control under very high resolution(2560x3520).

Figure 7. Visualization of results (1) under conflicting condition inputs (2) under very high-resolution generation.

The visual results further validate the robustness and versatility of our approach in handling complex multi-condition generation tasks.

G. Visual Comparison of Resolution Adaptability

As shown in the figure 11, we compare the controllability of our method with DiT-based controllable baseline methods, including ControlNet and OmniControl, across different resolutions. Clearly, our approach consistently demonstrates strong controllability, high text consistency, and superior image quality across resolutions ranging from low to high. However, at lower resolutions, ControlNet exhibits image distortion, while at higher resolutions, OmniControl also suffers from image degradation. This demonstrates that our method exhibits strong adaptability across different resolutions.

H. Limitations

While the proposed framework demonstrates significant improvements in flexibility and computational efficiency compared to existing DiT-based approaches, certain technical limitations remain and warrant further investigation. For example, in multi-conditional scenarios involving conflicting inputs, the model may generate artifacts characterized by overlapping layers, as illustrated in Figure 7. Additionally, our method cannot indefinitely upscale generated resolutions. When the resolution becomes extremely high, there is a decrease in the ability to control the output.

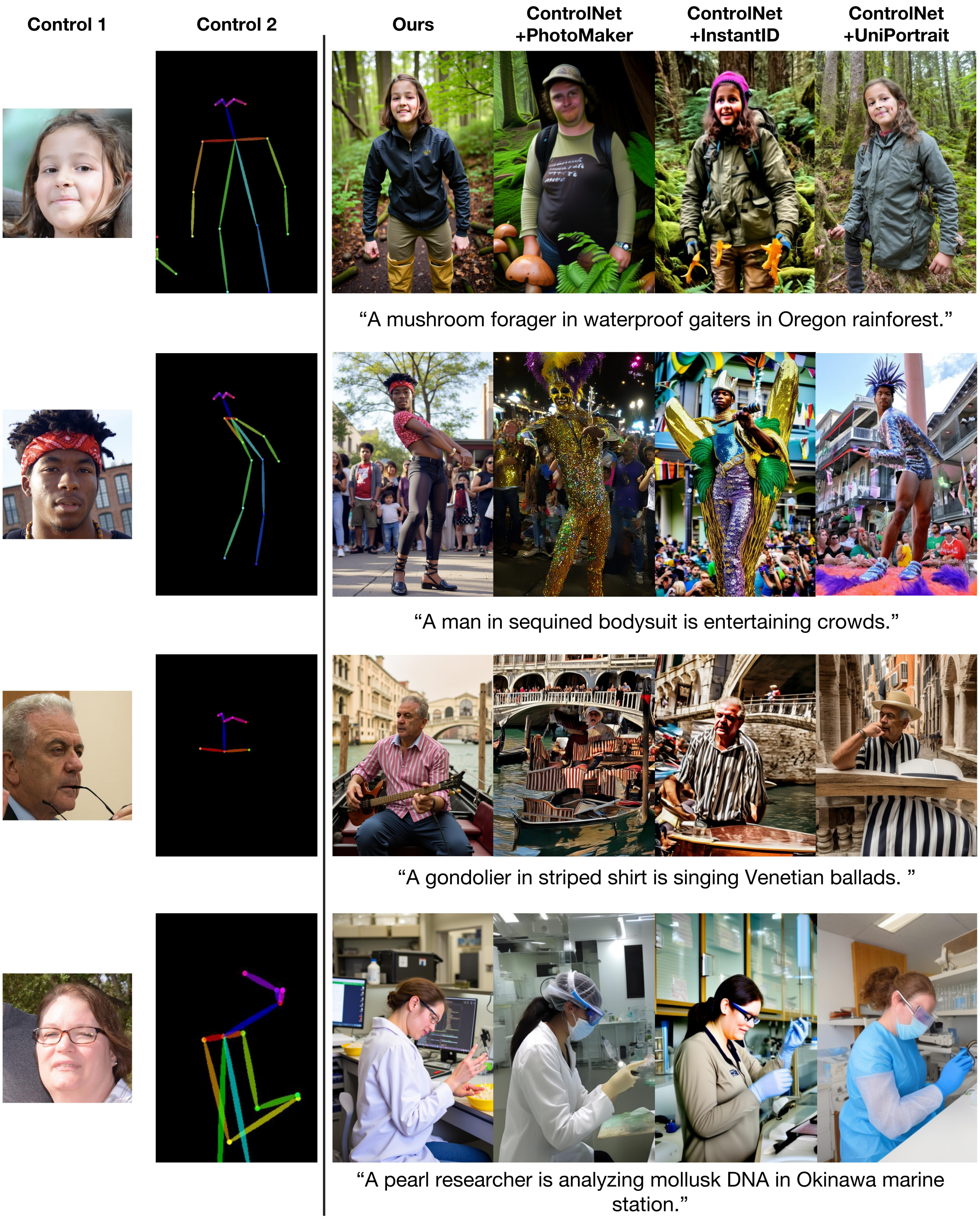

Control 1
Control 2
Ours
ControlNet +PhotoMaker
ControlNet +InstantID
ControlNet +UniPortrait
“A mushroom forager in waterproof gaiters in Oregon rainforest.”
“A man in sequined bodysuit is entertaining crowds.”
“A gondolier in striped shirt is singing Venetian ballads. ”
“A pearl researcher is analyzing mollusk DNA in Okinawa marine station.”

Figure 8. Visual comparison with Identity customization methods under multi-condition generation setting.

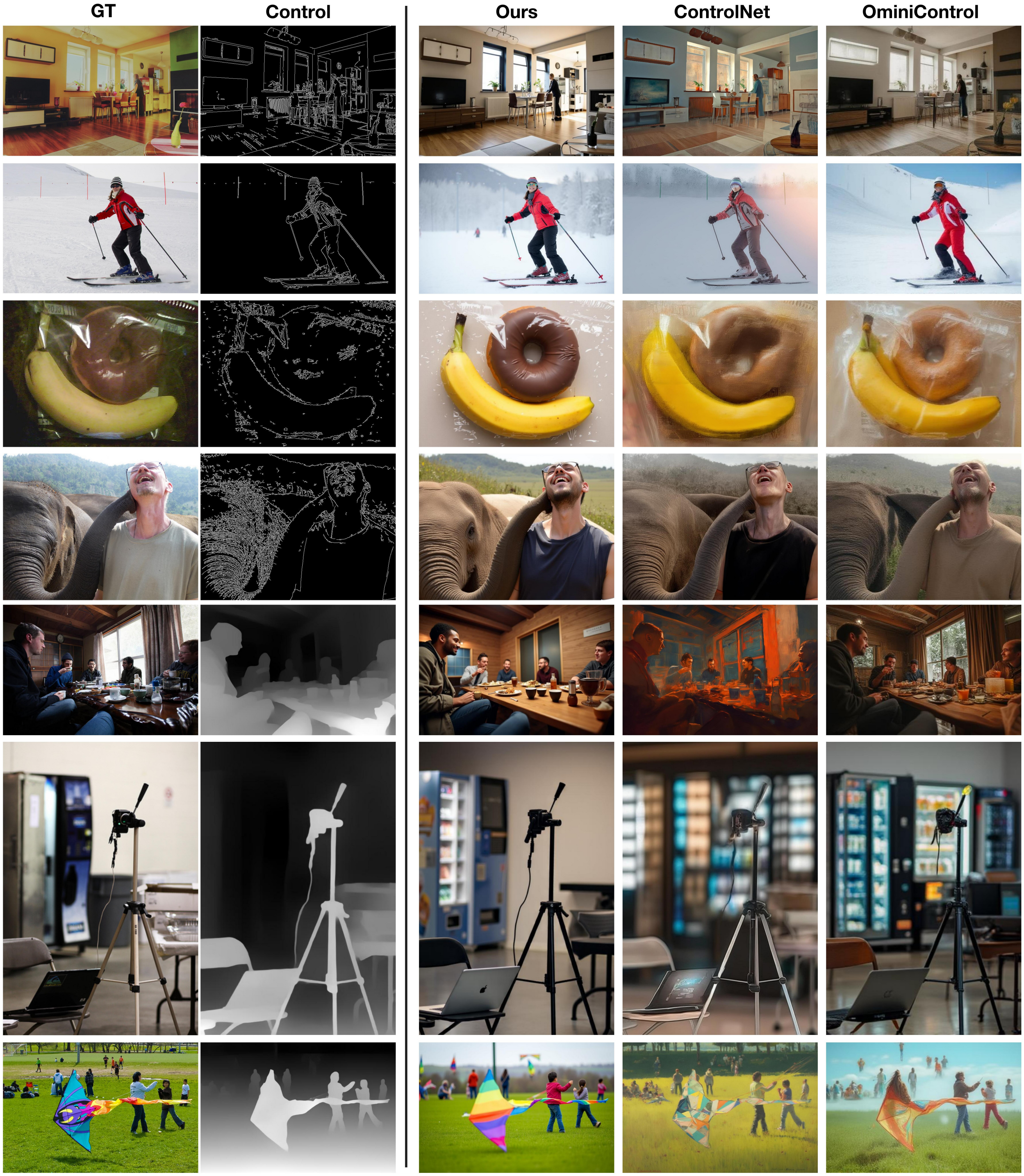


Figure 9. Visualization of spatial control generation.

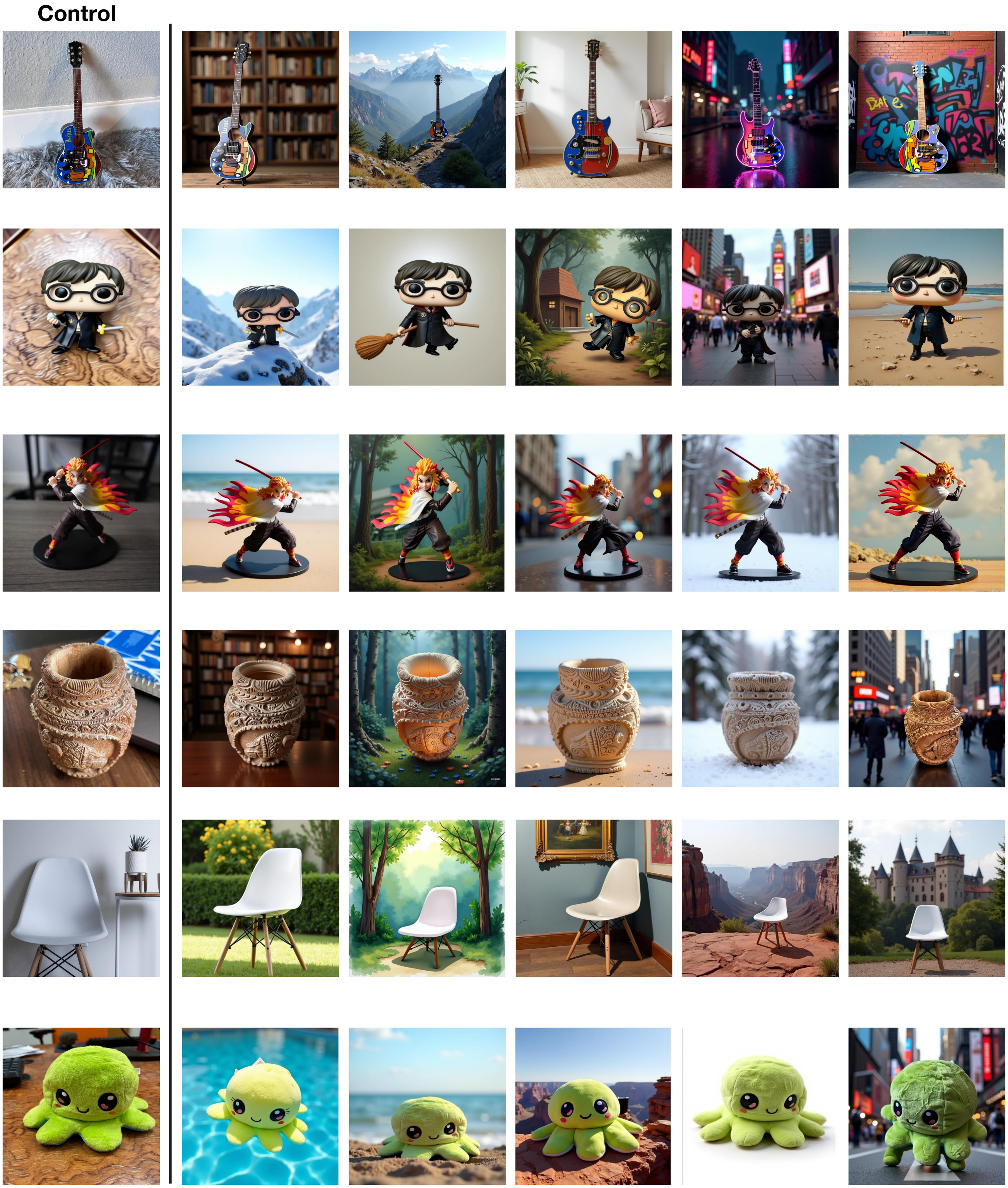


Figure 10. Visualization of subject control generation.

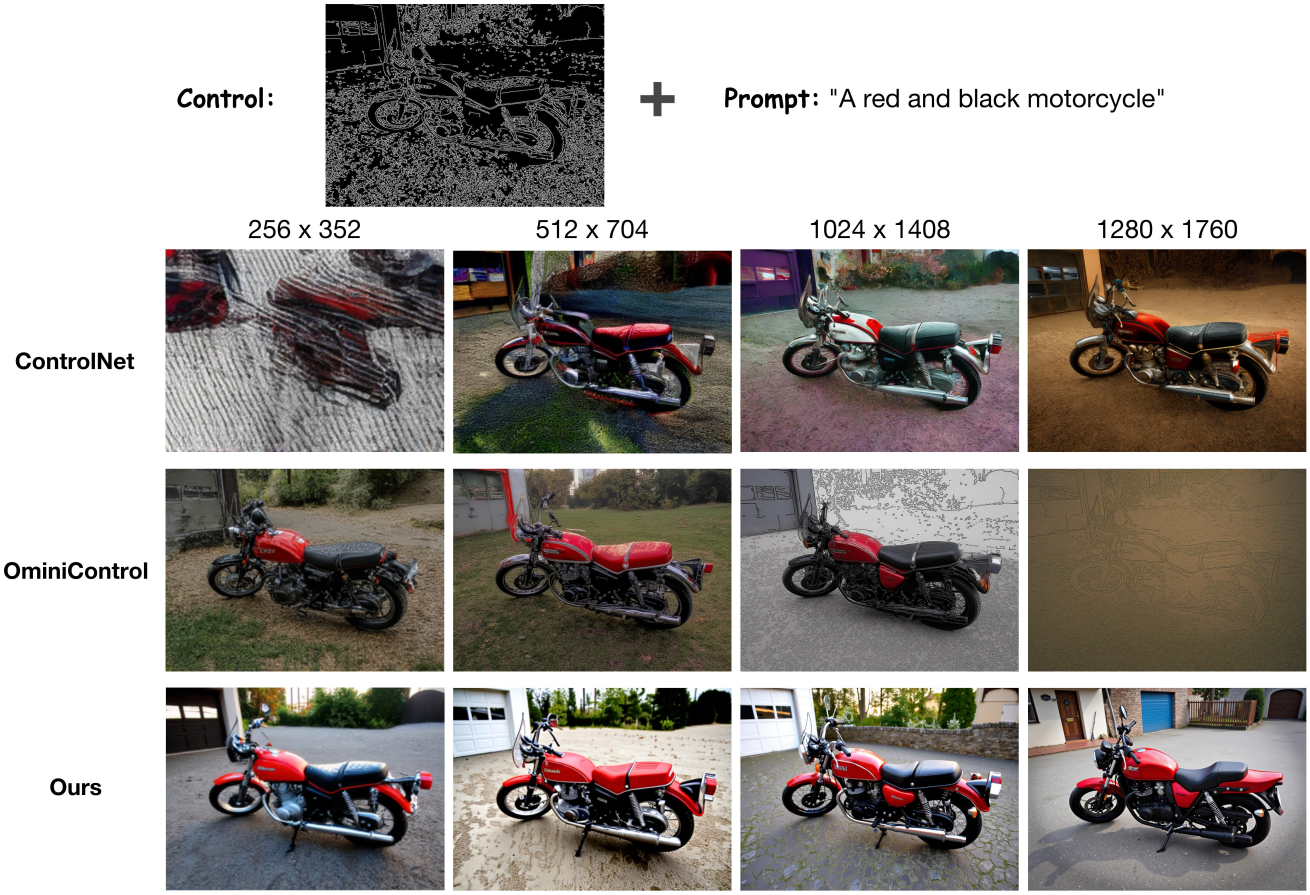


Figure 11. Visual comparison with baseline methods under different resolution generation settings.(zoom in for a better view)