UniCombine: Unified Multi-Conditional Combination with Diffusion Transformer

Supplementary Material

A1. Dataset Partitioning Scheme

In our proposed SubjectSpatial200K dataset, we utilize the ChatGPT-4o assessment scores provided by Subjects200K [37] on Subject Consistency, Composition Structure, and Image Quality to guide the dataset partitioning in our experiments.

- Subject Consistency: Ensuring the identity of the subject image is consistent with that of the ground truth image.
- Composition Structure: Verifying a reasonable composition of the subject and ground truth images.
- Image Quality: Confirming each image pair maintains high resolution and visual fidelity.

We partition the dataset into 139,403 training samples and 5,827 testing samples through Algorithm 1.

Algorithm 1: Dataset Partitioning Scheme

```
Input: example
Output: train or test
cs ← example[“Composite Structure”]
iq ← example[“Image Quality”]
sc ← example[“Subject Consistency”]
scores ← [cs, iq, sc]
if all($s == 5$ for $s$ in $scores$) then
    return train;
else if $cs \geq 3$ and $iq == 5$ and $sc == 5$ then
    return test;
```

A2. More Ablation on CMMDiT Attention

More quantitative and qualitative ablation results on the other multi-conditional generative tasks are provided here. The comprehensive ablation results in Tab. A1, Tab. A2, Tab. A3, Fig. A1, Fig. A2, and Fig. A3 demonstrate that the UniCombine performs better with our proposed CMMDiT Attention.

Method	CLIP-I ↑	DINO ↑	CLIP-T ↑	F1 ↑
Ours w/o CMMDiT	91.51	86.31	33.20	0.16
Ours w/ CMMDiT	**91.84**	**86.88**	**33.21**	**0.17**

Table A1. Quantitative ablation of CMMDiT Attention mechanism on training-free Subject-Canny task

A3. More Qualitative Results

More qualitative results are presented in Fig. A4 and Fig. A5.

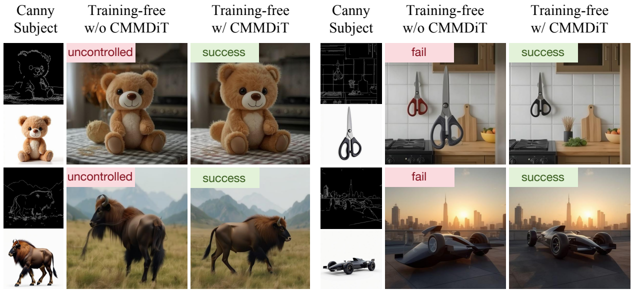

Figure A1. Qualitative ablation of CMMDiT Attention mechanism on training-free Subject-Canny task

Method	CLIP-I ↑	DINO ↑	CLIP-T ↑	MSE ↓
Ours w/o CMMDiT	90.83	85.38	33.38	547.63
Ours w/ CMMDiT	**91.15**	**85.73**	**33.41**	**507.40**

Table A2. Quantitative ablation of CMMDiT Attention mechanism on training-free Subject-Depth task

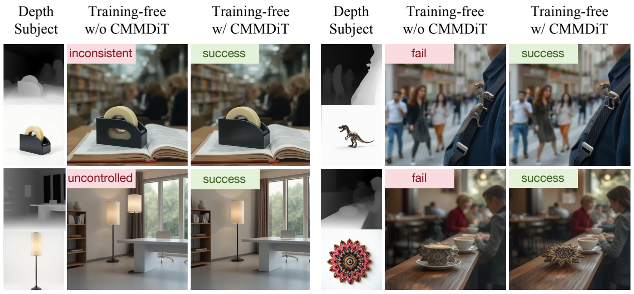

Figure A2. Qualitative ablation of CMMDiT Attention mechanism on training-free Subject-Depth task

Method	CLIP-T ↑	F1 ↑	MSE ↓
Ours w/o CMMDiT	33.70	0.17	524.04
Ours w/ CMMDiT	33.70	**0.18**	**519.53**

Table A3. Quantitative ablation of CMMDiT Attention mechanism on training-free Multi-Spatial task

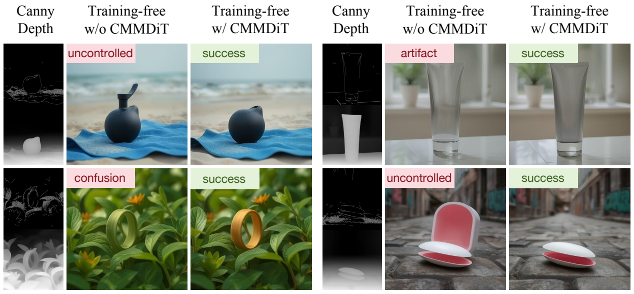

Figure A3. Qualitative ablation of CMMDiT Attention mechanism on training-free Multi-Spatial task

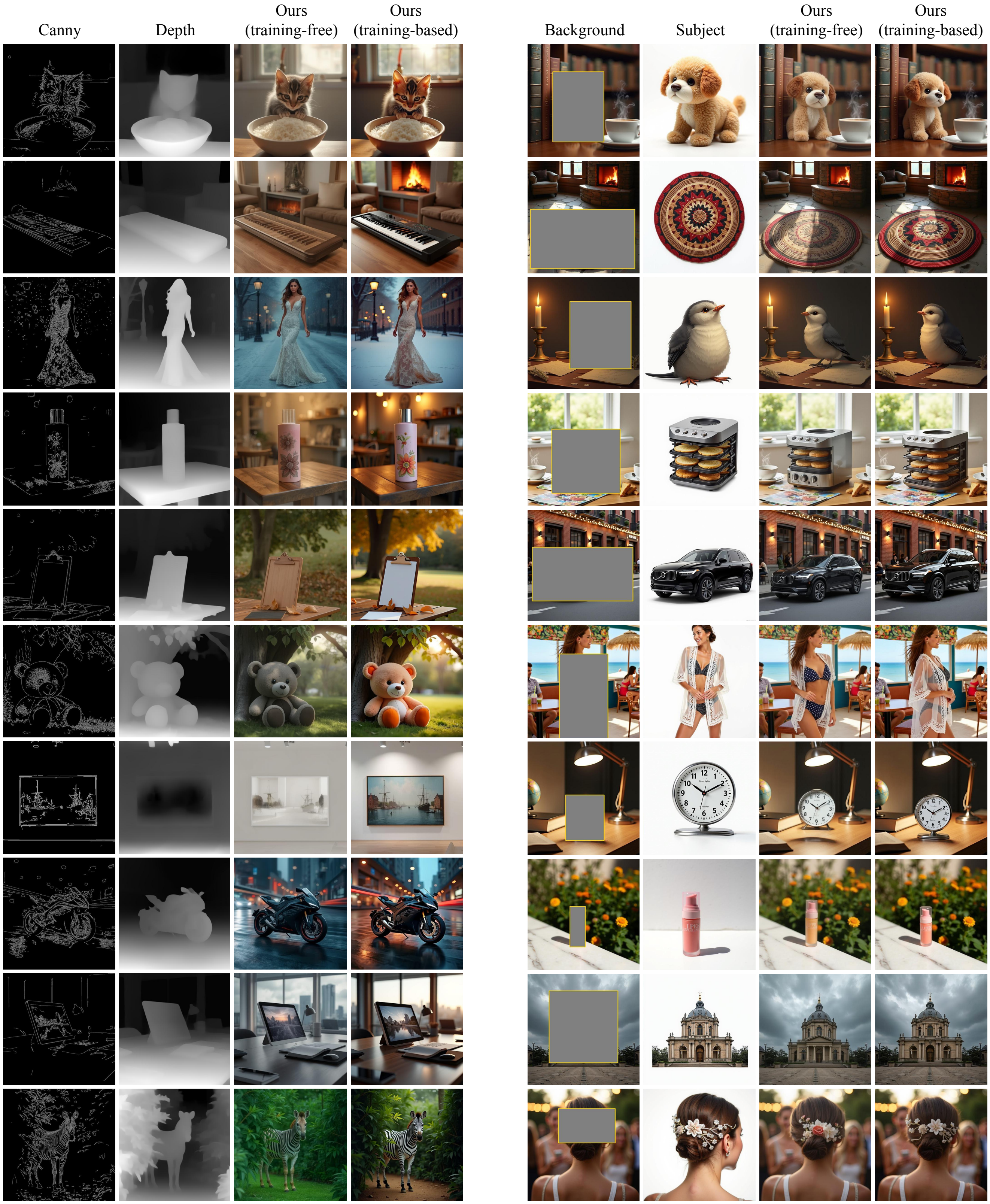

Figure A4. More qualitative results on Multi-Spatial and Subject-Insertion tasks.

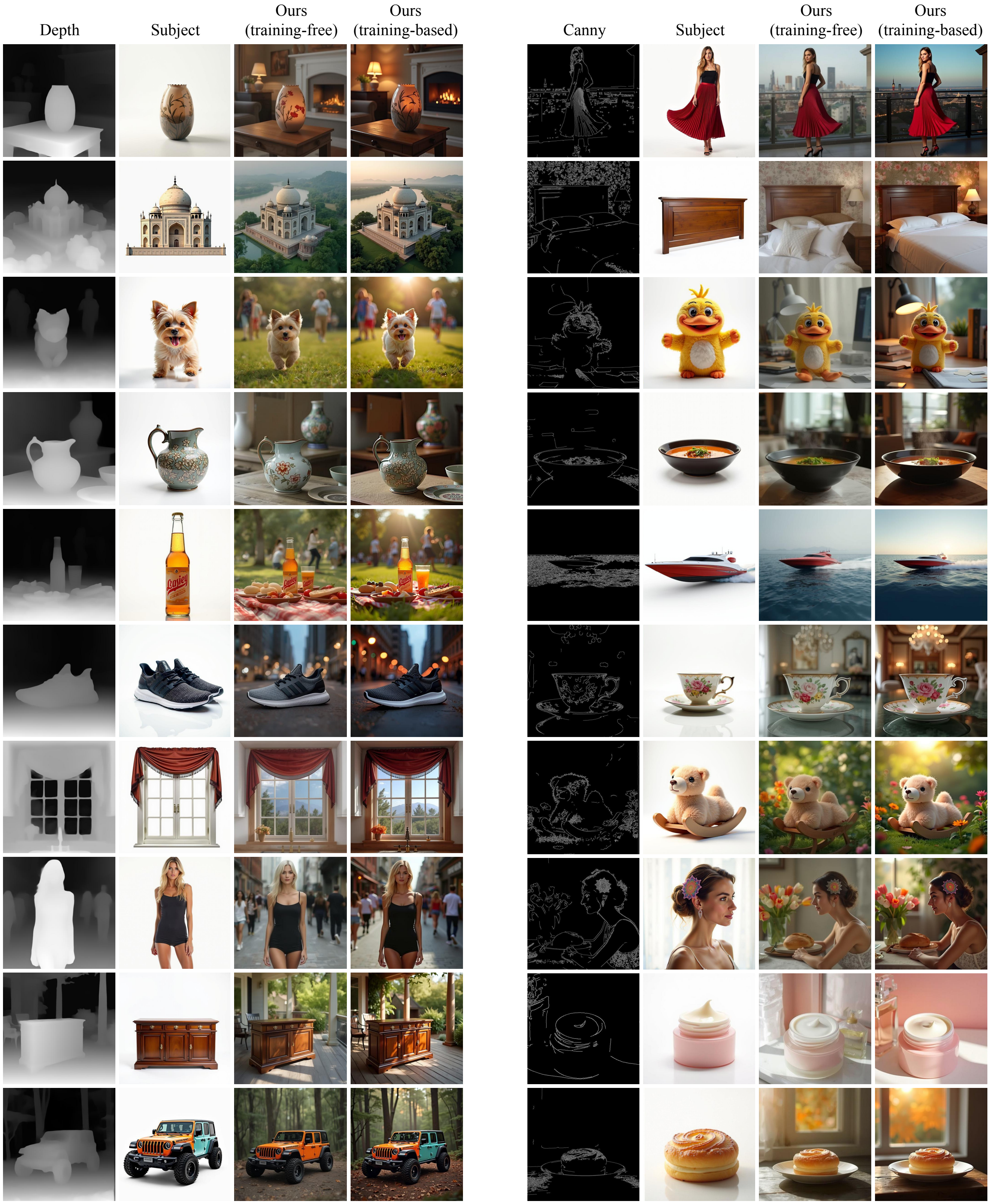

Figure A5. More qualitative results on Subject-Depth and Subject-Canny tasks.