

---

# Supplementary Material — Understand Before You Generate: Self-Guided Training for Autoregressive Image Generation

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A1 Supplementary Experimental Results

2 **Applying ST-AR on vanilla Transformer.** To verify the generalizability of ST-AR, we conduct  
3 experiments using a vanilla autoregressive Transformer. We follow the training setup of LlamaGen  
4 and disable designs such as RoPE and QK-Norm. The experimental results, shown in Table A1,  
5 indicate that with ST-AR, the vanilla Transformer achieves better generation performance.

Table A1: Results of vanilla Transformers on *ImageNet*-256 × 256 Benchmark.

|         | Model          | #Params | Epochs | FID↓         | sFID↓       | IS↑           | Prec.↑      | Rec.↑       |
|---------|----------------|---------|--------|--------------|-------------|---------------|-------------|-------------|
| w/o CFG | Transformer    | 111M    | 50     | 35.30        | 6.66        | 31.90         | 0.56        | <b>0.59</b> |
|         | <b>+ ST-AR</b> | 111M    | 50     | <b>29.37</b> | <b>6.08</b> | <b>38.88</b>  | <b>0.60</b> | <b>0.59</b> |
| w/ CFG  | Transformer    | 111M    | 50     | 9.67         | 6.94        | 129.50        | <b>0.84</b> | 0.37        |
|         | <b>+ ST-AR</b> | 111M    | 50     | <b>6.86</b>  | <b>6.40</b> | <b>159.13</b> | 0.83        | <b>0.43</b> |

6 **Text-to-Image Generation Experiment.** We also validate the effectiveness of ST-AR on text-  
7 conditional generation. We train LlamaGen-XL on a 2M subset of the SAM-11M dataset with an  
8 image resolution of 256 × 256. Pre-trained FLAN-T5 XL is used to extract text embeddings, with a  
9 maximum length of 120 for the embeddings. FID and CLIP Score are employed to evaluate image  
10 quality and text-image alignment. As shown in Table A2, ST-AR also demonstrates a significant  
11 improvement, highlighting the importance of extracting high-quality image representations in image  
12 generation.

Table A2: Results of LlamaGen-XL on the text-conditional COCO-2014 benchmark.

| Model          | Epochs | FID↓         | CLIP↑       |
|----------------|--------|--------------|-------------|
| LlamaGen-XL    | 50     | 17.08        | 0.25        |
| <b>+ ST-AR</b> | 50     | <b>13.52</b> | <b>0.29</b> |

## 13 A2 Additional Visualization Results

14 **Individual effects of the three losses.** In Figure A1, we visualize the attention maps of models  
15 trained with each of the three losses individually to support our claims in the main submission.  
16 The MIM loss significantly expands the attention range but fails to capture semantic information,

17 consistent with the results in Table 3. Both the inter-step contrastive loss and inter-view contrastive  
 18 loss slightly expand the attention range but significantly highlight semantically relevant areas.

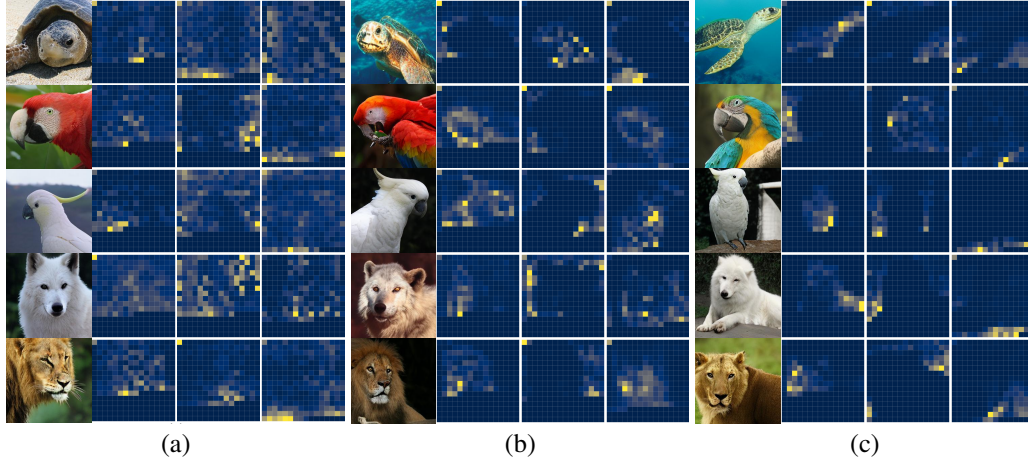


Figure A1: Attention maps of LlamaGen-B trained with (a) MIM loss, (b) inter-step contrastive loss, and (c) inter-view contrastive loss individually.

19 **Qualitative Comparison.** In Figure A2, we generate images using the same random seed for  
 20 LlamaGen-B and LlamaGen-B + ST-AR. It can be observed that, due to the lack of global and  
 21 semantic information, images generated by LlamaGen exhibit distortions and object discontinuities,  
 22 while images generated with ST-AR appear more natural. This also highlights the importance of  
 23 incorporating high-level semantic information.



Figure A2: Qualitative comparison between (a) LlamaGen-B and (b) LlamaGen-B + ST-AR.