

---

# TTS-VAR: A Test-Time Scaling Framework for Visual Auto-Regressive Generation (Supplementary Material)

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A Algorithm of TTS-VAR

2 We describe the algorithm of TTS-VAR in Alg. 1. Following the generation process of VAR [1]  
3 (Infinity [2]), TTS-VAR first predicts the residual tokens at the current scale and adds them to the  
4 accumulated feature maps. At scales that require clustering, TTS-VAR uses the extractor to gather  
5 features from  $b_i$  intermediate images decoded from the feature maps. It then clusters the samples  
6 based on these features and selects  $b_{i+1}$  ones as the next batch. At scales that require resampling,  
7 TTS-VAR employs the potential function to calculate scores for each image and samples  $b_{i+1}$  indexes  
8 from the multinomial distribution for superior intermediate states.

---

### Algorithm 1 TTS-VAR

---

**Require:** Scales  $\mathcal{S} = \{s_1, s_2, \dots, s_K\}$ , Descending batch sizes  $\mathcal{B} = \{b_1, b_2, \dots, b_K\}$ , Clustering  
scales set  $S_c$ , Resampling scales set  $S_r$ , Generative model  $\theta$ , Reward model  $r_\phi$  Extractor  $F$ ,  
Potential Score function  $P$ , Text prompt  $c$ .

```
1: Initialize accumulated feature map  $f_0$  with zeros.  
2: for  $i \in \{1, 2, \dots, K\}$  do ▷ Iterate through scales  
3:    $r_i \leftarrow \text{Generate}(\theta, b_i, s_i, f_{i-1}, c)$   
4:    $f_i \leftarrow f_{i-1} + r_i$   
5:   if  $s_i \in S_c$  then ▷ Clustering phase  
6:      $x \leftarrow \text{Decode}(f_i)$   
7:      $feat \leftarrow F(x)$   
8:      $index \leftarrow \text{KMeans++}(feat, b_{i+1})$   
9:      $f_i \leftarrow f_i[index]$   
10:  else if  $s_i \in S_r$  then ▷ Resampling phase  
11:     $x \leftarrow \text{Decode}(f_i)$   
12:     $rw \leftarrow r_\phi(x)$   
13:     $p \leftarrow P(rw)$   
14:     $index \leftarrow \text{Multinomial}(p, b_{i+1})$   
15:     $f_i \leftarrow f_i[index]$   
16:  end if  
17: end for  
    return Final generated images  $\text{Decode}(f_K)$ 
```

---

$N$	Strategy	GenEval	ImageReward	HPS	CLIP	Aesthetic
1	Raw Inference	0.6946	1.1320	0.3042	0.3366	0.5811
1	Ours	<b>0.7253</b>	<b>1.3226</b>	<b>0.3084</b>	<b>0.3395</b>	<b>0.5822</b>
2	Importance Sampling	0.7022	1.1941	0.3051	0.3374	0.5807
2	Best-of-N	0.7087	1.2545	0.3069	0.3384	0.5813
2	Ours	<b>0.7403</b>	<b>1.4136</b>	<b>0.3106</b>	<b>0.3411</b>	<b>0.5821</b>
4	Importance Sampling	0.7116	1.2883	0.3067	0.3387	0.5815
4	Best-of-N	0.7244	1.3471	0.3083	0.3397	0.5820
4	Ours	<b>0.7437</b>	<b>1.4605</b>	<b>0.3112</b>	<b>0.3414</b>	<b>0.5821</b>
8	Importance Sampling	0.7181	1.3657	0.3085	0.3395	0.5810
8	Best-of-N	0.7364	1.4144	0.3103	0.3406	<b>0.5820</b>
8	Ours	<b>0.7530</b>	<b>1.4995</b>	<b>0.3122</b>	<b>0.3420</b>	0.5810

Table 6: **Scores over Different Strategies.** This table shows results with different scaling strategies.

## 9 B Detailed Main Results

We present detailed results of variant curves in Table 6. As evident, *TTS-VAR* demonstrates clear advantages across all indicators [3–7] compared to the baselines. In Table 7, we list each item of the GenEval [3] metric. Generally, our method significantly improves performance on handling two objects and counting tasks. We attribute this to the importance of structural accuracy in multi-character scenes, particularly when two objects and multiple identical objects (counting) are involved. For instance, when provided with a prompt for three objects, there is a possibility that the model may incorrectly generate a layout with four objects. Once this error occurs, following the structure-to-detail generation process in VAR, it becomes challenging for subsequent scales to rectify. However, *TTS-VAR* facilitates structure diversity, thereby enabling the selection of a layout with the correct configuration and avoiding the irreversible wrong generation process for inferior samples.

$N$	Strategy	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.
1	Raw Inference	0.6946	0.9938	0.8351	0.5923	<b>0.9293</b>	0.2020	0.6150
1	Ours	<b>0.7253</b>	<b>0.9938</b>	<b>0.9072</b>	<b>0.6518</b>	0.9192	<b>0.2096</b>	<b>0.6700</b>
2	Importance Sampling	0.7022	<b>0.9969</b>	0.8497	0.6071	0.9268	0.1869	0.6475
2	Best-of-N	0.7087	0.9906	0.8789	0.6339	0.9242	0.1944	0.6300
2	Ours	<b>0.7403</b>	0.9936	<b>0.9278</b>	<b>0.7113</b>	<b>0.9318</b>	<b>0.1995</b>	<b>0.6775</b>
4	Importance Sampling	0.7116	0.9906	0.8840	0.6339	<b>0.9318</b>	0.1970	0.6325
4	Best-of-N	0.7244	<b>1.0000</b>	0.8969	0.6756	0.9242	0.1944	0.6550
4	Ours	<b>0.7437</b>	0.9906	<b>0.9510</b>	<b>0.6994</b>	0.9293	<b>0.2045</b>	<b>0.6875</b>
8	Importance Sampling	0.7181	0.9906	0.8969	0.6220	0.9318	0.2121	0.6550
8	Best-of-N	0.7364	0.9938	0.9201	0.6756	<b>0.9444</b>	0.2146	0.6700
8	Ours	<b>0.7530</b>	<b>0.9969</b>	<b>0.9501</b>	<b>0.7411</b>	0.9318	<b>0.2172</b>	<b>0.6800</b>

Table 7: **GenEval Details.** This table shows each item of the GenEval benchmark, "Object" is short for "Obj.", and "Attribute" is short for "Attri.".

## 20 C Performance over Computational Consumption

We here display the changing curves of GenEval, ImageReward, and HPSv2 over the increment of computation in Fig. 8, along with the increment of sample number  $N$ . As shown, our method *TTS-VAR* has higher computational efficiency and surpasses Importance Sampling and Best-of-N with less than half TFLOPs.

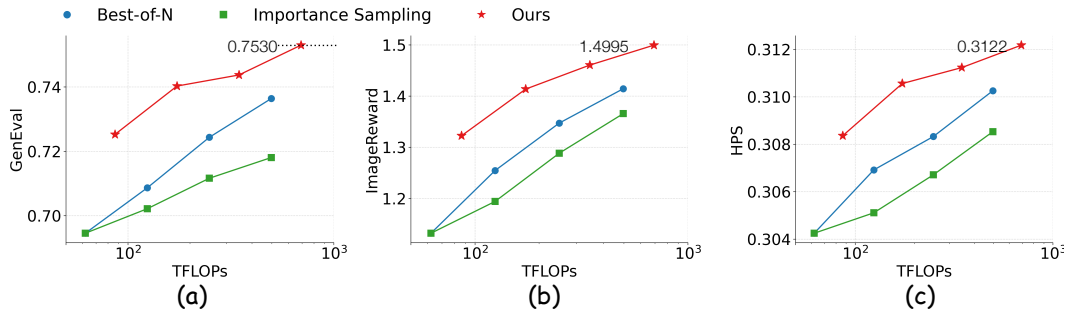


Figure 8: **Performance over Flops.** This figure shows the variant curves of different methods with computational consumption as the x-axis, demonstrating the efficiency of our method.

## D Ablation Study

### D.1 Pipeline Ablation

We present the ablation study of different design components in the overall pipeline in Table 8. As adaptive batch sampling alone (without integrated sample selection mechanisms) cannot directly enhance generation performance, these cases are denoted by "-". Excluding these baseline cases, both clustering-based diversity search and resampling-based potential selection demonstrate performance improvements, with statistically significant gains observed in reward and related evaluation indicators.

Notably, the clustering approach yields relatively moderate improvements, which can be attributed to its primary function of maintaining structural diversity rather than actively identifying superior samples for subsequent generation. The combination of diversity maintenance through clustering and quality-based selection via resampling synergistically enhances the effectiveness of the pipeline. This dual-mechanism framework ultimately achieves substantial performance gains over the baseline system, with the resampling component playing the pivotal role in selecting high-quality candidates for iterative refinement.

$N$	Method	GenEval	ImageReward	HPS	CLIP	Aesthetic
2	Infinity	0.6946	1.1320	0.3042	0.3366	0.5811
	+BoN	0.7087	1.2545	0.3069	0.3384	0.5813
	+Adaptive Batch Sampling	-	-	-	-	-
	+Clustering-Based Diversity Search	0.7220	1.2591	0.3072	0.3385	0.5816
	+Resampling-Based Potential Selection	0.7403	1.4136	0.3106	0.3411	0.5821
4	Infinity	0.6946	1.1320	0.3042	0.3366	0.5811
	+BoN	0.7244	1.3471	0.3083	0.3397	0.5820
	+Adaptive Batch Sampling	-	-	-	-	-
	+Clustering-Based Diversity Search	0.7294	1.3608	0.3095	0.3403	0.5824
	+Resampling-Based Potential Selection	0.7437	1.4605	0.3112	0.3414	0.5821

Table 8: **Pipeline Ablation.** This table shows gains from each design.

### D.2 Reward Models

We implement comparisons on using different reward models to rate the intermediate images and calculate the potential scores (VALUE), including Aesthetic [6], ImageReward [4], HPSv2 [5], and HPS+ImageReward. Owing to different value ranges of HPS and ImageReward, for HPS+ImageReward, we first calculate scores using the two models separately, then softmax the values of each model into the range  $[0, 1]$ , and finally take the average as the potential scores.

As shown in Table 9, generally, each reward model motivates an increase in the corresponding metric. For instance, with  $N = 4$ , the Aesthetic model, the ImageReward model, and the HPS model achieve the highest scores in the associated indicators, respectively. Among different models, ImageReward promotes improvements more. Especially with  $N = 2$ , ImageReward demonstrates a clear lead in GenEval and even defeats HPS in HPS score. We attribute this to the ability to clearly distinguish between superior and inferior samples, along with scoring that better aligns with human preferences.

$N$	Reward Model	GenEval	ImageReward	HPS	CLIP	Aesthetic
2	-	0.7087	1.2545	0.3069	0.3384	0.5813
2	Aesthetic	0.6966	1.123	0.3054	0.3366	<b>0.6004</b>
2	ImageReward	<b>0.7403</b>	<b>1.4136</b>	<b>0.3106</b>	<b>0.3411</b>	0.5821
2	HPS	0.7135	1.2246	0.3102	0.3391	0.583
2	HPS+ImageReward	0.7238	1.3522	0.3088	0.3402	0.5824
4	-	0.7244	1.3471	0.3083	0.3397	0.5820
4	Aesthetic	0.6842	1.1172	0.3056	0.3363	<b>0.6114</b>
4	ImageReward	<b>0.7437</b>	<b>1.4605</b>	0.3112	<b>0.3414</b>	0.5821
4	HPS	0.7255	1.2812	<b>0.3154</b>	0.3402	0.5843
4	HPS+ImageReward	0.7413	1.4128	0.3101	0.3406	0.5818

Table 9: **Reward Model Ablation.** This table shows results using different models for the potential.

### D.3 $\lambda$ Setting

In Table 10, we exhibit the results using different temperature  $\lambda$  in the resampling process with fixed clustering operations. Intuitively, higher temperature promotes the expression of intermediate states

with higher potential scores and prevents superior samples. However, excessively high temperatures can also widen the gap between intermediate states with the highest scores and those with scores that are only slightly lower. This can directly inhibit the generation of these slightly lagging intermediate states, which may ultimately become the optimal results. As shown, though there is a steady increase in ImageReward,  $\lambda = 10.0$  falls behind  $\lambda = 5.0$  with  $N = 4$  in GenEval.

$N$	<b>Lambda</b>	GenEval	ImageReward	HPS	CLIP	Aesthetic
2	-	0.7087	1.2545	0.3069	0.3384	0.5813
2	0.1	0.7065	1.2458	0.3065	0.3387	0.5811
2	0.5	0.7210	1.3167	0.3080	0.3400	0.5819
2	1.0	0.7222	1.3459	0.3087	0.3403	0.5821
2	5.0	0.7361	1.4010	0.3101	0.3410	0.5821
2	10.0	<b>0.7403</b>	<b>1.4136</b>	<b>0.3106</b>	<b>0.3411</b>	<b>0.5821</b>
4	-	0.7244	1.3471	0.3083	0.3397	0.5820
4	0.1	0.7308	1.3576	0.3089	0.3400	0.5816
4	0.5	0.7347	1.3918	0.3094	0.3406	0.5819
4	1.0	0.7418	1.4097	0.3099	0.3407	0.5820
4	5.0	<b>0.7465</b>	1.4500	0.3108	0.3412	0.5820
4	10.0	0.7437	<b>1.4605</b>	<b>0.3112</b>	<b>0.3414</b>	<b>0.5821</b>

Table 10:  $\lambda$  **Ablation**. This table shows results with different lambda values.

## E More Visualization Results

We present pairs of results on GenEval to compare the quality of Infinity, Infinity-IS, Infinity-BoN, and Infinity-*TTS-VAR*. These cases are sampled from text prompts in GenEval, which include the single object, two objects, counting, colors, position, and color attributes. As shown, our method produces higher-quality samples for the single object, such as the airplane in the left second line, effectively avoiding the generation of artifacts. In two-object settings, *TTS-VAR* successfully distinguishes references to different objects and generates accurate outputs based on prompts. For example, in the right second line, it eliminates conceptual mixtures and object disappearances. As illustrated in lines 3-10 on the right side, *TTS-VAR* also excels at determining counting numbers, positional relationships, and color attributes. Notably, in the last right line, our method generates a counterintuitive "green carrot," demonstrating its ability to separate objects from their natural attributes.

## F Societal Impact

When applied to VAR models, *TTS-VAR* enhances the alignment of generated images with textual descriptions, making the generation process more controllable and better suited to meet creative and production demands. However, we also acknowledge the potential for misuse of this method, which could lead to privacy and copyright concerns. Nonetheless, we believe that our in-depth research into the VAR generation process will help researchers gain a clearer understanding, advance studies on controllability and safety in generation, and ultimately ensure that image generation models become safe and manageable tools.

## G Limitation and Future Work

Though *TTS-VAR* shows significant improvement over the baseline and sets a new record, it still has two main limitations. First, *TTS-VAR* does not completely address the misalignment between text prompts and generated images. As indicated by the scores in Table 7, there are still some failure cases, particularly in the Position item. Second, while *TTS-VAR* is based on a general coarse-to-fine process, its potential application to other coarse-to-fine models, such as autoregressive models that use 1-D tokenizers, remains unexplored. In the future, we will investigate the generation process more thoroughly, examining the reasons for failure and designing solutions to unlock further scaling potential. Additionally, we plan to assess the compatibility of *TTS-VAR* with other autoregressive coarse-to-fine models, including those utilizing 1-D tokenizers and hybrid architectures that combine diffusion models. These efforts aim to create a more robust scaling framework for text-to-image synthesis while enhancing the methodological transferability of coarse-to-fine paradigms.



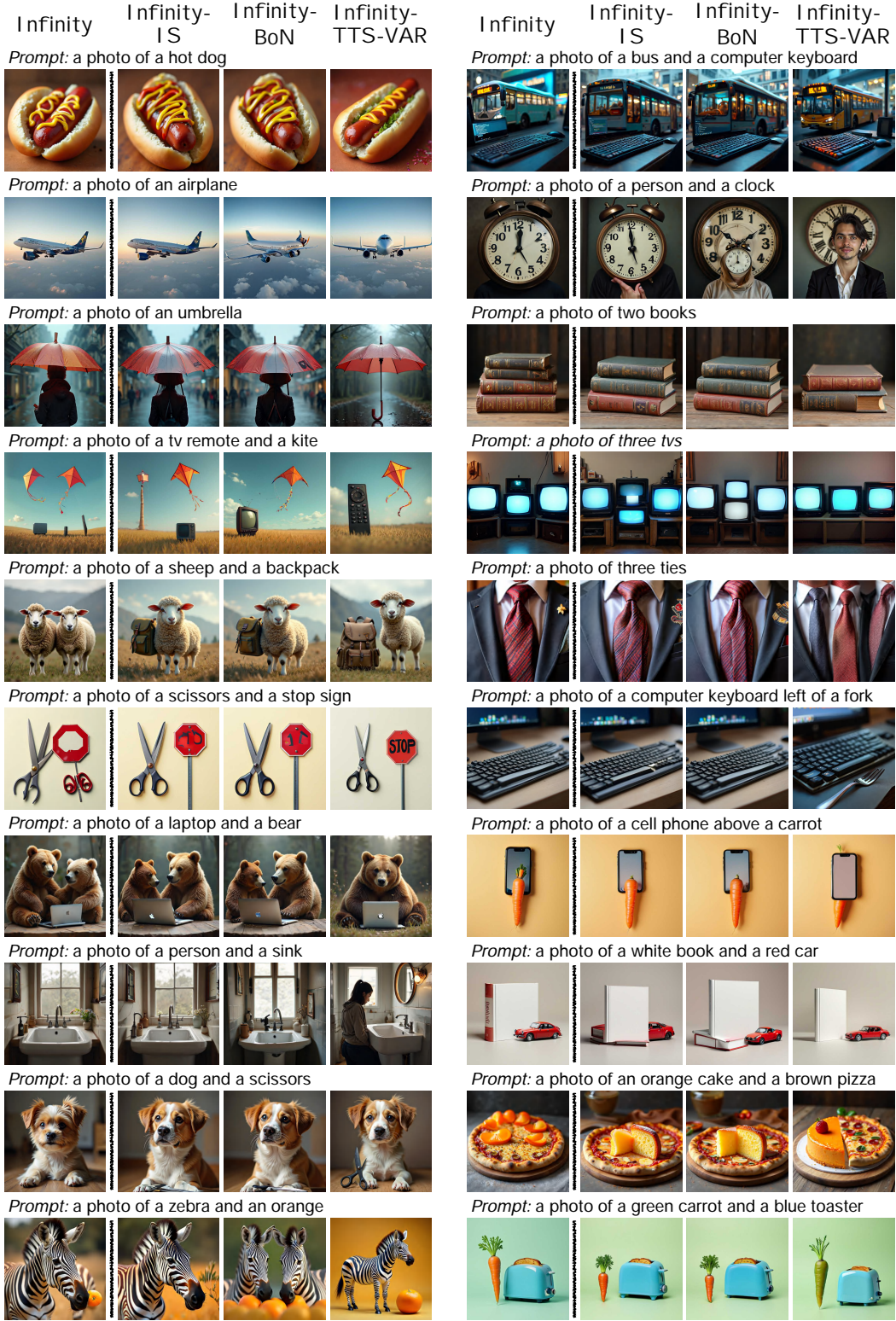


Figure 9: **More Visualization Results.** Samples are generated from GenEval prompts, with "IS" meaning Importance Sampling, "BoN" meaning Best-of-N, and our method *TTS-VAR*. We display various cases, including the single object, two objects, counting, colors, position, and color attributes.

## References

- [1] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2404.02905.
- [2] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv: 2412.04431*, 2024.
- [3] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [4] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2304.05977.
- [5] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv: 2306.09341*, 2023.
- [6] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv: 2210.08402*, 2022.
- [7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *Emnlp*, 2021.