SUPPLEMENTARY MATERIAL: HALTON SCHEDULER FOR MASKED GENERATIVE IMAGE TRANSFORMER

Victor Besnier[1] **Mickael Chen**[2,⋆] **David Hurych**[1]

Eduardo Valle[2] **Matthieu Cord**[2,3]

[1]**Valeo.ai, Prague** [2]**Valeo.ai, Paris** [3]**Sorbonne Université, Paris** ⋆now at **H company, Paris**

{firstname}.{lastname}@valeo.com

1 TRAINING DETAILS

Table 1 provides all the hyperparameters used to train our models across both modalities. In Table 2, we detail the architecture of our models.

For class-to-image generation, we employed an architecture similar to DiT-XL (Peebles & Xie, 2023), utilizing a patch size of 2 to reduce the number of tokens from 32×32 to 16×16. Due to GPU memory constraints, we opted not to use Exponential Moving Averages (EMA). Additionally, we used the 'tie_word_embedding' technique, where the input and output layers share weights, reducing the number of trainable parameters.

We used the T5-XL encoder for text-to-image synthesis, which processes 120 text tokens per input, resulting in a text embedding of size [120, 2048] for each sentence. To integrate text conditioning, we employed a transformer architecture similar to DiT-L (Peebles & Xie, 2023), the largest model we could fit on our GPU with EMA. The condition is incorporated using classical cross-attention.

Condition	**text-to-image**	**class-to-image**
Training steps	5×10^5	2×10^6
Batch size	2048	256
Learning rate	5×10^{-5}	1×10^{-4}
Weight decay	0.05	5×10^{-5}
Optimizer	AdamW	AdamW
Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.96$
Lr scheduler	Cosine	Cosine
Warmup steps	2500	2500
Gradient clip norm	0.25	1
EMA	0.999	–
CFG dropout	0.1	0.1
Data aug.	No	Horizontal Flip
Precision	bf16	bf16

Table 1: Hyper-parameters used in the training of text-to-img and class-to-img models.

2 EUCLIDEAN DISTANCE IN THE TOKEN SPACE

One assumption of our analysis is that parts of the image that are closer together tend to be more similar in appearance. This relationship is well understood in pixel space (Huang & Mumford, 1999). In Figure 1, we show that the principle also holds in the token space. We measure the appearance dissimilarity of tokens using the Euclidean distance of their corresponding latent representation on the LlamaGen tokenizer on the ImageNet dataset. As shown in the figure, tokens that are spatially

Condition	text-to-image	class-to-image
Parameters	479.8M	705.0M
Input size	32×32	32×32
Hidden dim	1024	1152
Codebook size	16384	16384
Depth	24	28
Heads	16	16
Mlp dim	4096	4608
Patchify (p=)	2	2
Dropout	0.0	0.0
Conditioning	Cross-attention	AdaLN

Table 2: Architecture design of the text-to-img and class-to-img models.

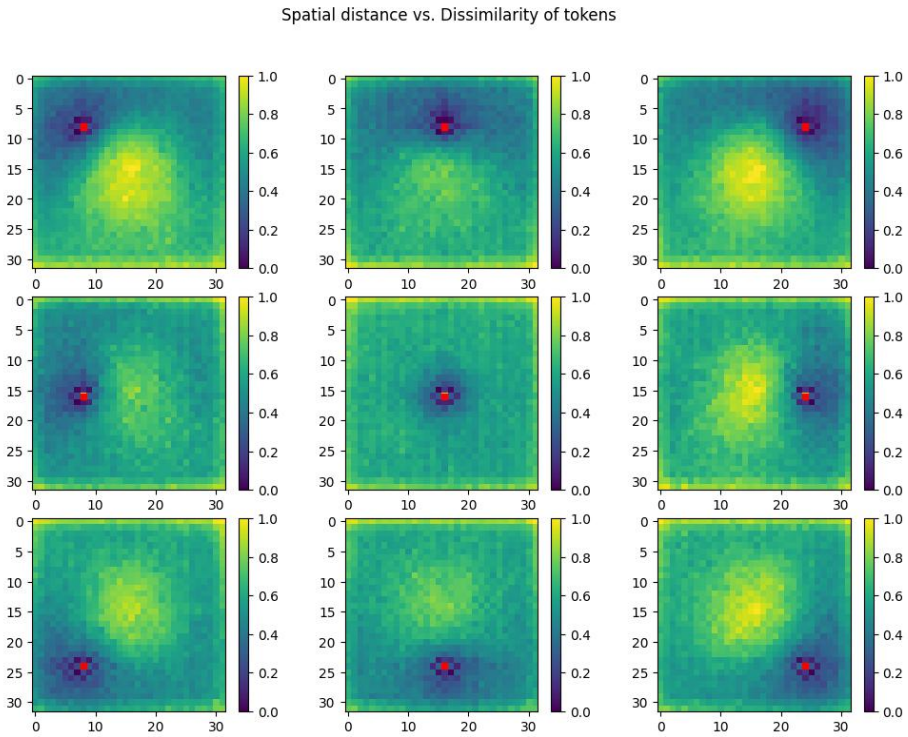


Figure 1: **Spatial distance vs. appearance dissimilarity of tokens.** The color map indicates the normalized appearance dissimilarity between each token and the reference token (shown in red). The closest tokens to the reference token are the most similar (dark blue). Tokens further away tend to be dissimilar (bright yellow).

close to the reference token (in red) also have the closest representations in feature space (dark blue). Tokens that are spatially further apart tend to have dissimilar representations (green to yellow).

3 Generative Method Comparison

We show here a complete evaluation of our methods with the Halton scheduler again recent methods from the literature in Table 3. The analysis of Figure 2 shows that we are narrowing the gap between diffusion, auto-regressive, and masked image modeling, bringing the latter to the competitive forefront

Type	Model	#Para.	FID↓	IS↑	Precision↑	Recall↑
GAN	BigGAN (Brock et al., 2018)	112M	6.95	224.5	**0.89**	0.38
	GigaGAN (Kang et al., 2023)	569M	3.45	225.5	0.84	**0.61**
	StyleGan-XL (Sauer et al., 2022)	166M	**2.30**	**265.1**	0.78	0.53
Diffusion	ADM (Dhariwal & Nichol, 2021)	554M	10.94	101.0	0.69	**0.63**
	CDM (Ho et al., 2020)	–	4.88	158.7	–	–
	LDM-4 (Rombach et al., 2022)	400M	3.60	247.7	–	–
	DiT-XL/2 (Peebles & Xie, 2023)	675M	**2.27**	**278.2**	**0.83**	0.57
AR	VQGAN (Esser et al., 2020)	1.4B	15.78	74.3	–	–
	ViT-VQGAN (Yu et al., 2021)	1.7B	4.17	175.1	–	–
	RQTran.-re(Lee et al., 2022)	3.8B	3.80	**323.7**	–	–
	LlamaGen-3B (Sun et al., 2024)	3.1B	**2.18**	263.3	**0.81**	**0.58**
MIM	MaskGIT (Chang et al., 2022)	227M	6.18	182.1	0.80	0.51
	Token-Critics (Lezama et al., 2022)	–	4.69	174.5	0.76	0.53
	AutoNAT-L (Ni et al., 2024)	194M	**2.68**	278.8	–	–
	FSQ (Mentzer et al., 2023)	–	4.53	–	**0.86**	0.45
	MAGE (Li et al., 2023)	439M	7.04	123.5	–	–
	Ours + Halton	705M	3.74	**279.5**	0.81	**0.60**

Table 3: **Model comparison on class-conditional ImageNet 256×256 benchmark.** The proposed Halton scheduler outperforms the original MaskGIT considerably, demonstrating competitive performance among Masked Image Modeling (MIM) approaches.

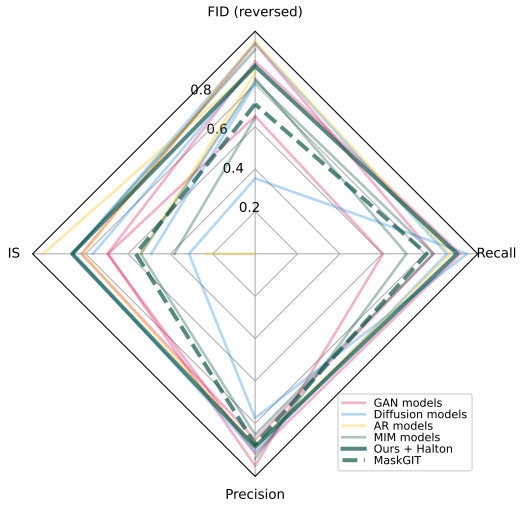


Figure 2: **Analysis of the SOTA results of Table 3.** In this radar plot, each model is represented as a line. Each metric is normalized to 1 for its best model. The FID is reversed, so higher is always better. The improvement brought by the Halton scheduler over the vanilla MaskGIT with Confidence scheduler is immediately noticeable. Our scheduler brings fast masked generative transformers to the competitive vanguard of existing methods.

of generative methods while maintaining their speed advantage (dozens of inference steps for MIM *vs.* hundreds for auto-regressive and for diffusion).

4 Intermediate generation

An interesting property of our approach is shown in Figure 3 depicting the intermediate construction of the macaw (`088`). It highlights that most images are already fixed after a few steps. First, the bird's

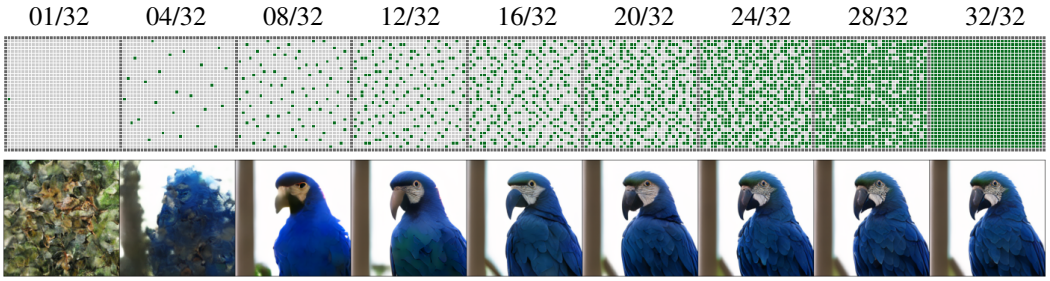


Figure 3: **Evolution of the sampling using Halton scheduler.** The macaw's (**088**) color, texture, and shape, as well as the background, are set after only 12 steps, with only ∼16% tokens predicted. That showcases the ability of the Halton scheduler to extract information from the tokens by reducing their correlation.

Steps	Percentage of tokens	FID ↓	IS ↑
4/32	2%	-	-
8/32	8%	146.2	7.31
12/32	16%	76.7	24.1
16/32	28%	24.9	79.6
20/32	43%	8.85	142.2
24/32	61%	6.25	174.3
28/32	82%	6.12	182.2
32/32	100%	6.11	184.0

Table 4: **Evaluation of intermediate generated samples on ImageNet 512×512.** Most of the gains are on early steps, which are crucial to achieving good FID and IS. Later steps keep improving but may be skipped as a compromise between quality and compute.

blue color and the white background are completely set after only four steps with $25/1024 \approx 2\%$ unmasked tokens. The shape is fixed at the 8th step (8% tokens unmasked), and the texture starts to appear at 12 steps (16% tokens unmasked). This means that the rest of the token will only influence the high-frequency details of the generated image. We push the analysis further by computing the FID and IS for these intermediate samples (see Table 4), where we evaluate the results given the generated intermediate images. While the first 16 steps significantly increase both the FID and the IS, the last 12 steps only decrease the FID score by 0.14 points.

5 Pseudo-code for Halton Sequence

In algorithm 1, we detail the generation of the Halton sequence, producing a sequence of size n' with a base b. In practice, we generate two sequences with $b = 2$ and $b = 3$, respectively, representing 2D coordinates of the points to select. We then discretize the space in a 32×32 grid. Duplicate points are discarded, ensuring complete grid coverage by setting n' appropriately. The coordinates of the remaining points determine the order of token unmasking during sampling.

6 Text Prompts

Prompts used for our text-to-image model, corresponding to Figure 7 in the main paper, from top-left to bottom-right:

1. A robot chef expertly crafts a gourmet meal in a high-tech futuristic kitchen, intricate details.
2. An old-world galleon navigating through turbulent ocean waves under a stormy sky lit by flashes of lightning.
3. A cozy wooden cabin perched on a snowy mountain peak, glowing warmly in the night, styled like a classic Disney movie, featured on ArtStation.

Algorithm 1: Compute the Halton sequence

Parameters:
    b: the base of the Halton sequence,
    n': the number of points in the sequence to compute
Results:
    S: the first n' points of the Halton sequence in base b

$n \leftarrow 0$
$d \leftarrow 1$
$S \leftarrow []$

for $i \leftarrow 0$ **to** n' **do**
    $x \leftarrow d - n$
    if $x = 1$ **then**
        $n \leftarrow 1$
        $d \leftarrow d \times b$
    end
    else
        $y \leftarrow d \div b$
        while $y \geq x$ **do**
            $y \leftarrow y \div b$
        end
        $n \leftarrow ((b + 1) \times y) - x$
    end
    S.append($n \div d$)
end

return S

4. A blue sports car is parked. The sky above is partly cloudy, suggesting a pleasant day. The trees have a mix of green and brown foliage. There are no people visible in the image.
5. An oil painting of rain in a traditional Chinese town.
6. Volumetric lighting, spectacular ambient lights, light pollution, cinematic atmosphere, Art Nouveau style illustration art, artwork by SenseiJaye, intricate detail.
7. A mystical fox in an enchanted forest, glowing flora, and soft mist, rendered in Unreal Engine.
8. Photo of a young woman with long, wavy brown hair tied in a bun and glasses. She has a fair complexion and is wearing subtle makeup, emphasizing her eyes and lips. She is dressed in a black top. The background appears to be an urban setting with a building facade, and the sunlight casts a warm glow on her face.
9. Photo of a young man in a black suit, white shirt, and black tie. He has a neatly styled haircut and is looking directly at the camera with a neutral expression. The background consists of a textured wall with horizontal lines. The photograph is in black and white, emphasizing contrasts and shadows. The man appears to be in his late twenties or early thirties, with fair skin and short, dark hair.
10. Selfie photo of a wizard with a long beard and purple robes, he is apparently in the middle of Tokyo. Probably taken from a phone.
11. An image of Pikachu enjoying an elegant five-star meal with a breathtaking view of the Eiffel Tower during a golden sunset.
12. A sleek airplane soaring above the clouds during a vibrant sunset, with a stunning view of the horizon.
13. A towering mecha robot overlooking a vibrant favela, painted in bold, abstract expressionist style.
14. Anime art of a steampunk inventor in their workshop, surrounded by gears, gadgets, and steam. He is holding a blue potion and a red potion, one in each hand
15. Pirate ship trapped in a cosmic maelstrom nebula rendered in cosmic beach whirlpool engine.

16. A futuristic solarpunk utopia integrated into the lush Amazon rainforest, glowing with advanced technology and harmonious nature.
17. A teddy bear wearing a blue ribbon taking a selfie in a small boat in the center of a lake.
18. Digital art, portrait of an anthropomorphic roaring Tiger warrior with full armor, close up in the middle of a battle.

7 Random Samples from our class conditioned model

In Figure 4, we show that our model can generate diverse images and more intricate details compared to the confidence scheduler. Furthermore, a comparison with the Confidence sampler reveals that the latter produces overly simplistic and smooth images, often with poorly defined backgrounds. In contrast, our approach consistently produces greater diversity, particularly in rendering background elements.

8 Failure Cases

Multiple Objects. The initial tokens sampled by the Halton scheduler tend to spread across the image, leading to instances where multiple objects or entities appear within a single image. For example, this can result in multiple occurrences of a specific object, such as multiple goldfish, or even multiple parts of the same entity, such as a bird with two heads, see Figure 5a for class conditioning and Figure 5c for text-to-image.

Inability to Self-Correct. Unlike diffusion models, MIM-based methods cannot iteratively correct earlier predictions. Diffusion models generate predictions over the entire image at each step, allowing for refinement and correction of previous errors. In contrast, once a token is predicted in MIM, it remains fixed, even if incorrect, as there is no mechanism for subsequent correction during the generation process.

Challenges in Complex Class/Prompt. The model exhibits difficulties in generating certain complex classes and adhering closely to prompts. As demonstrated in Figure 5b, the model struggles to accurately generate human faces or bodies in ImageNet.

Similarly, in text-to-image conditioned tasks, it can fail to produce coherent scene compositions or faithfully render text, especially when dealing with intricate or abstract descriptions. Indeed, the model fails to render the word "HALTON" in Figure 5c and the bicycle below the elephant for the last image.

References

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, June 2022.

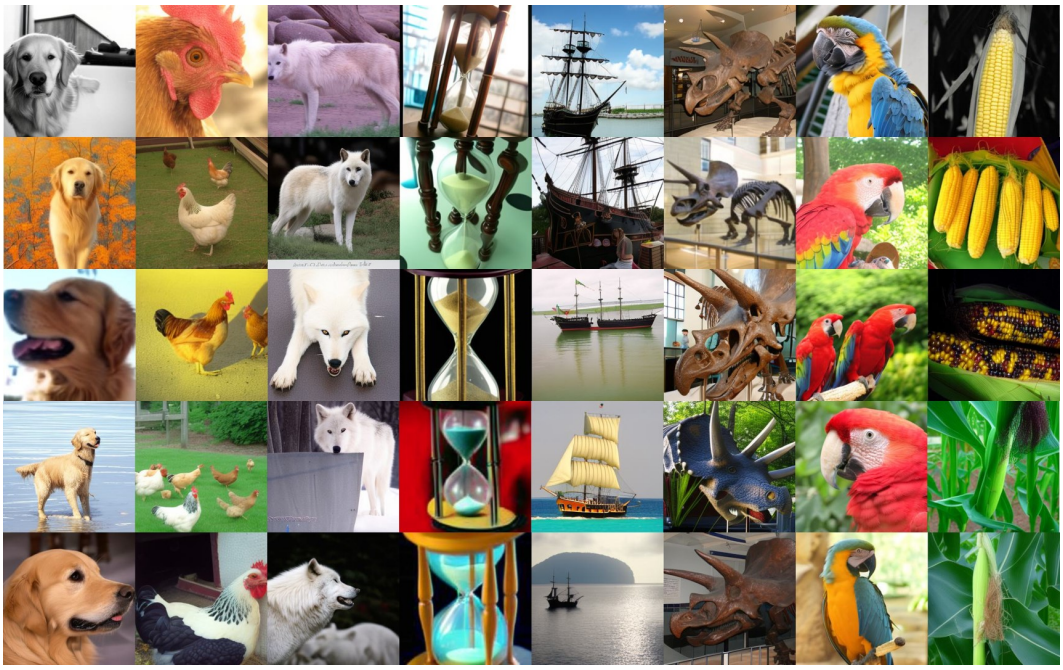
Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2020.

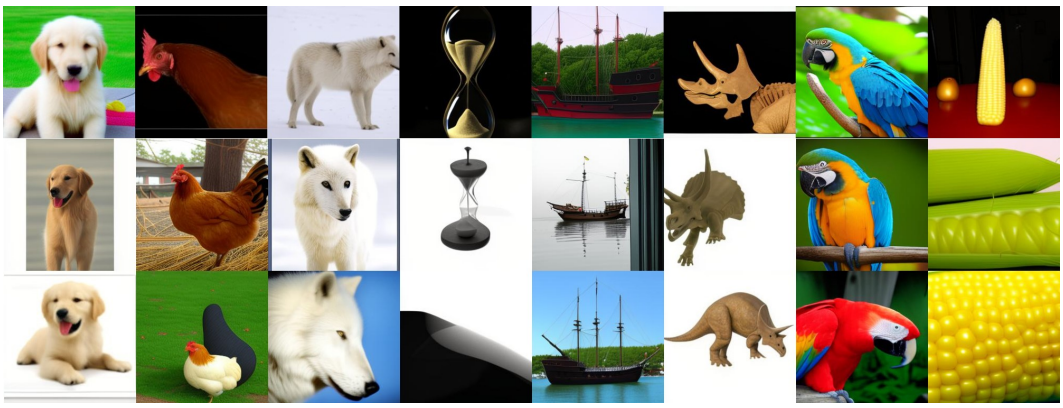
Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jinggang Huang and David Mumford. Statistics of natural images and models. In *CVPR*, volume 1, pp. 541–547. IEEE, 1999.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, pp. 10124–10134, 2023.



(a) **MaskGIT using our Halton scheduler.**



(b) **MaskGIT using the Confidence scheduler.**

Figure 4: **Scheduler comparison on random samples generated by a class-to-image model.** The Halton scheduler demonstrates a higher level of detail, capturing finer features than the Confidence scheduler, which lacks details, especially in the background.

(a) **Multiple Object Generation**

(b) **Human attributes**

(c) **Prompt adherence:**
- A female character with long, flowing hair that appears to be made of ethereal, swirling patterns resembling the NL...
- A vibrant street wall covered in colorful graffiti, the centerpiece spells 'HALTON'.
- An elephant is riding a bicycle in an empty street.

Figure 5: **Failure cases.** The Halton scheduler solves some, but not all the challenges of sampling tokens in parallel. Long-range correlations still pose a challenge for MaskGIT with the Halton Scheduler.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, pp. 11523–11532, 2022.

José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *ECCV*, 2022.

Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. MAGE: masked generative encoder to unify representation learning and image synthesis. In *CVPR*, 2023.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *CVPR*, pp. 7007–7016, June 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, October 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH*, pp. 1–10, 2022.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.