HALTON SCHEDULER FOR MASKED GENERATIVE IMAGE TRANSFORMER SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

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 INTERMEDIATE GENERATION

An interesting property of our approach is shown in Figure 1 depicting the intermediate construction of the macaw (**088**). It highlights that most images are already fixed after a few steps. First, the bird's 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 3), 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.

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.

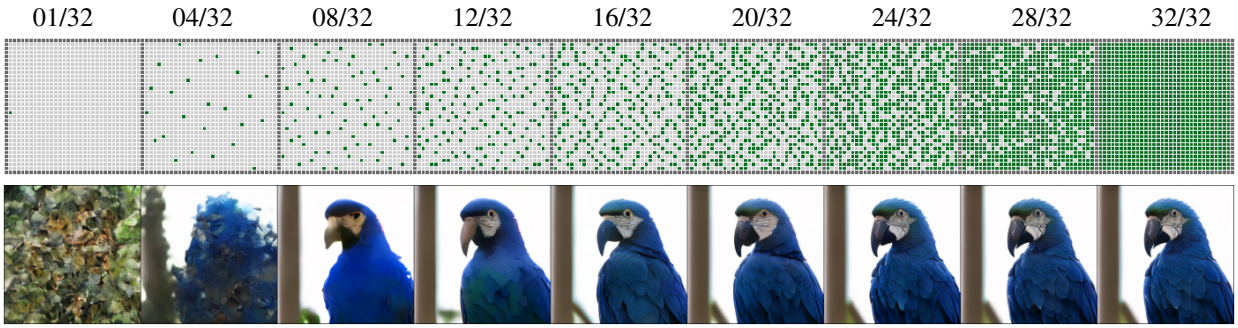


Figure 1: **Evolution of the sampling using Halton scheduler.** The macaw's (`088`) color, texture, and shape, as well as the background, are set after only 12 steps, with only ∼16% tokens predicted. That showcases the ability of the Halton scheduler to extract information from the tokens by reducing 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 3: **Evaluation of intermediate generated samples on ImageNet 512×512.** Most of the gains are on early steps, which are crucial to achieve good FID and IS. Later steps keep improving but may be skipped as a compromise between quality and compute.

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

3 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.

4 Text Prompts

Prompt use for our text-to-image model, from left-top 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.
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.

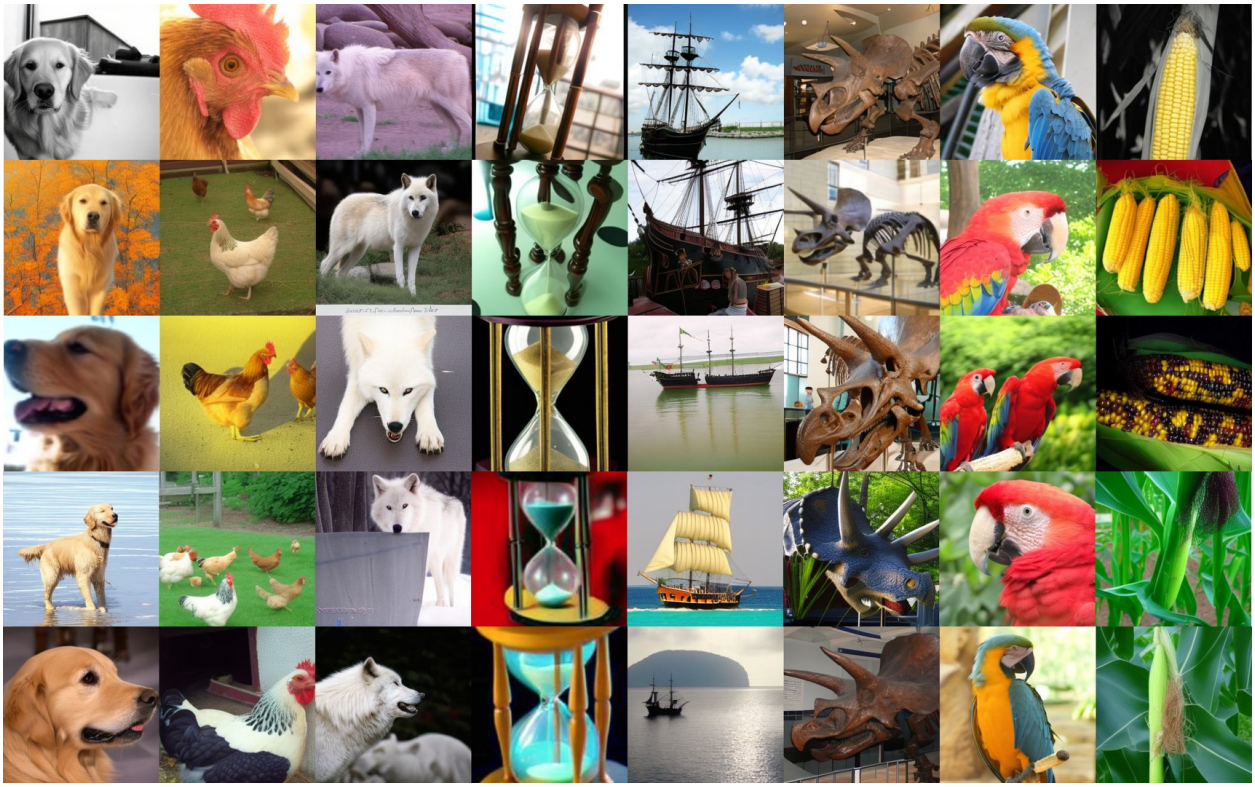
10. Selfie photo of a wizard with a long beard and purple robes, he is apparently in the middle of Tokyo. Probably taken from a phone.
11. An image of Pikachu enjoying an elegant five-star meal with a breathtaking view of the Eiffel Tower during a golden sunset.
12. A sleek airplane soaring above the 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.

5 Random Samples from our class conditioned model

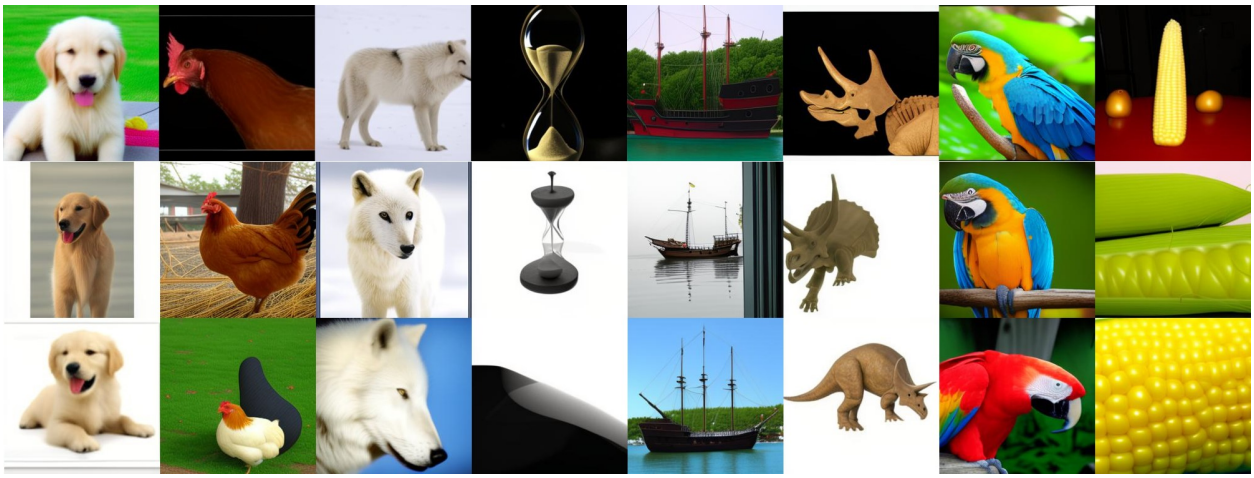
In Figure 2, we show that our model can generate diverse images and more intricate details compared to the confidence scheduler. Furthermore, a comparison with the Confidence sampler reveals that the latter produces overly simplistic and smooth images, often with poorly defined backgrounds. In contrast, our approach consistently produces greater diversity, particularly in rendering background elements.

References

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



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



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

Figure 2: **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.