

Appendix Overview

Due to page limitation of the main body, as indicated by our submission, the supplementary material offers further discussion on the motivation of mid-frequency band and more visual results with higher resolution, which are summarized below:

- More *intuition* on hybrid frequency aware diffusion Models to text-guided image inpainting, as mentioned in Sec.2.2 of the main body (Sec.A)
- Additional quantitative results and qualitative analysis with *higher resolution* for the comparison with state-of-the-arts, as mentioned in Sec.3.2 of the main body (Sec.B).
- More ablation studies on the impact of adaptively extracting the low-and-mid frequency bands, as mentioned in Sec.2.3.2 and 2.3.3 of the main body (Sec.C).
- More ablation studies on the impact of text and null-text prompts on low-and-mid frequency bands during denoising process, as mentioned in Sec.3.3.1 of the main body (Sec.D).
- Additional visual results for the ablation study about hyperparameter sensitivity analysis of λ for the length of the early and late stage, as mentioned in Sec.3.3.2 of the main body (Sec.E).
- We list the limitations and broader impacts (Sec.F).
- We provide the code for reproducibility (Sec.G).

A More Intuition on Hybrid Frequency Aware Diffusion Models to Text-Guided Image Inpainting

Due to page limitation, we offer more visual results of the mid-frequency band of the denoising process. Fig.1(a) illustrates the text-guided denoising results for mid-frequency band. To validate its stability, we visualize the layout information as bounding box for each step, which, as indicated by [2, 3, 11], is closely related to the mid-frequency band. The denoised mid-frequency band also changes during the initial early stage owing to text prompts with high-level noise, yet quickly converges by the end of the early stage e.g., 60-th step, which further leads to the stable mid-frequency band with *nearly* no influence by text prompt across the whole late stage with low-level noise. We further exhibit additional visual results by performing the experiments on the same image based on three similar text prompts. Despite the different text prompts, the mid-frequency can quickly converges by the end of the early stage e.g., 60-th. Following the above, we propose to exploit the mid-frequency band during denoising process, which plays the pivotal role of achieving the semantics consistency upon text prompts, while can better achieve the semantics consistency between masked and unmasked regions, while preserve its own frequency band well, owing to its robustness to the text prompt during the denoising process.

B Additional Quantitative and Qualitative analysis in the Sec.3.2 of the Main body

Evaluation metric. We adopt five metrics to evaluate the inpainted results below: Peak Signal-to-Noise Ratio (**PSNR**), structural similarity index (**SSIM**) [9] and Mean Squared Error (**MSE**) evaluate low-level pixel-wise differences between generated images and their ground truth counterparts. Additionally, Learned Perceptual Image Patch Similarity (**LPIPS**) [13] measures perceptual similarity by computing the distance between deep features extracted from a pre-trained neural network, offering a robust metric for assessing perceptual alignment; CLIP Similarity (**CLIP Score**) [10] measures text-image consistency by projecting both the generated images and their corresponding text prompts into a shared embedding space using the CLIP model. It then evaluates the similarity between their embeddings.

As indicated in the main body, we further exhibit additional quantitative results by performing the experiments on the EditBench [8]; see Table.1. It is observed that **NTN-Diff** enjoys larger PSNR, SSIM and CLIP Score, together with smaller MSE and LPIPS than the competitors. Notably, **BrushEdit**[5] remains the large performance margins (at most, 3.86% for PSNR, 0.42% for MSE, 1.85% for LPIPS, 0.86% for SSIM, and 0.1% for CLIP Score) compared to **NTN-Diff** in Table.1.

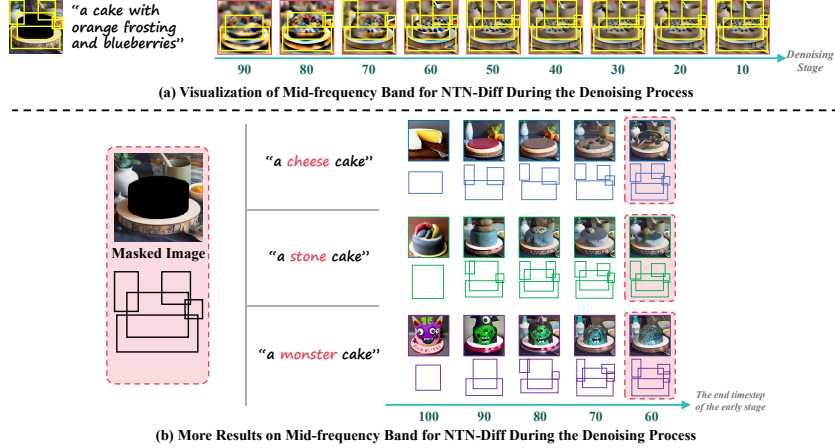


Figure 1: We investigate the text-guided denoising process for the mid-frequency bands. For each step, we employ the bounding boxes to visualize the variations for layout information in (a) and the changes of layout information for the same image based on three similar text prompts. the denoised mid-frequency band changes during the initial early stage owing to text prompts with high-level noise, yet quickly converges by the end of the early stage, which further leads to the stable states with nearly no influence by text prompt across the whole late stage.

Table 1: Quantitative comparisons between NTN-Diff* and other diffusion-based inpainting models over EditBench for inpainting are shown, where all models use Stable Diffusion V1.5 as the baseline model. **red** and **blue** stand for the best and second best result. NTN-Diff* achieves the best result.

Metrics		Masked Region Preservation				Text Alignment
Models	Venue	PSNR \uparrow	MSE $\times 10^3\downarrow$	LPIPS $\times 10^3\downarrow$	SSIM $\times 10^2\uparrow$	CLIP Score \uparrow
SDI [7]	CVPR' 22	23.25	6.94	24.30	90.13	28.00
BLD [1]	TOG' 23	20.89	10.93	31.90	85.09	28.62
CNI [12]	ICCV' 23	12.71	69.42	159.71	79.16	28.16
CNI* [12]	ICCV' 23	22.61	35.93	26.14	94.05	27.74
PP [14]	ECCV' 24	23.34	20.12	24.12	91.49	27.80
BrushNet* [4]	ECCV' 24	33.66	0.63	10.12	98.13	28.87
BrushEdit* [5]	ArXiv' 24	32.97	0.70	7.24	98.60	29.62
HDP [6]	ICLR' 25	23.07	6.70	24.32	92.56	28.34
NTN-Diff* (Ours)	-	36.83	0.28	5.39	99.46	29.72

* with blending operation of BrushNet

We also present the quantitative results of **NTN-Diff** with the pixel-level blending operation of [4], named **NTN-Diff***, to preserve the unmasked regions, which demonstrates the ability to tame the hybrid frequency issue, the results further verifies the intuition in Sec.1 of the main body – *NTN-Diff can achieve the semantics consistency between mid-and-low frequency bands across masked and unmasked regions, while preserving unmasked regions.*

To shed more light on the advantages of our method, we further perform the visual analysis on the inpainted results with *the higher resolution*. Fig.2 delivers the following: due to the discrepancy between the diffusion process for unmasked regions substitution and the denoising process for masked region alignment upon text prompt. BLD [1] inevitably generate the content out of the masked regions (the first column of the Fig.2(a)). Note that, PP [14] and BrushNet [4] often exhibits the unreasonable inpainted result, *e.g.*, The extra woman(the 4th row of PP in Fig.2(a)) and the disappeared dinosaur (the 1st row of BrushNet in Fig.2(b)), owing to the unmasked regions are disrupted, thus fails to reconstruct the masked region as per text prompt, which attributes to the *the low-frequency bands for both masked and unmasked regions are easily to be changed by text prompts, especially under the early stage of the text-guided denoising process with high-level noise.* As opposed to that, *mid-frequency band can better achieve the semantics consistency between masked and unmasked regions,*

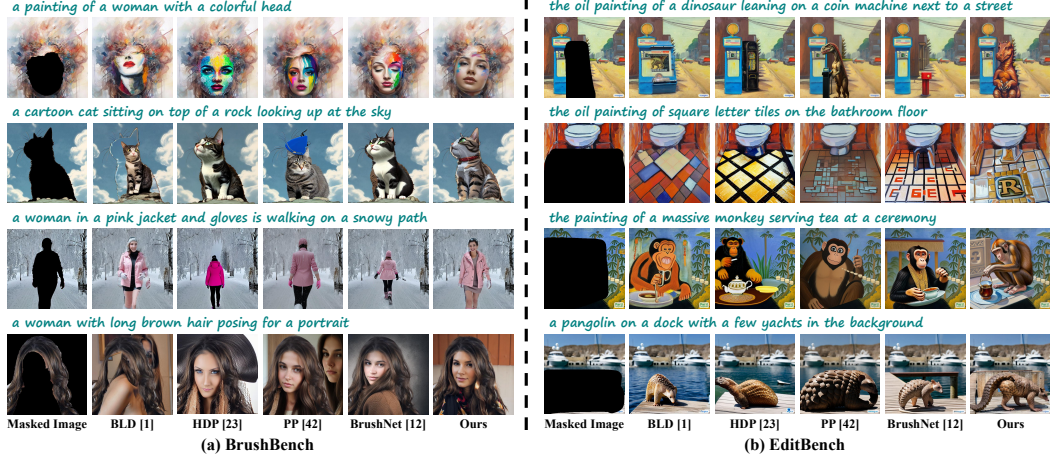


Figure 2: Comparison of the text-guided inpainted results with the state-of-the-arts on BrushBench [4] and EditBench [8]. NTN-Diff delivers the superior inpainted results over others, which can simultaneously preserve the unmasked regions while achieve the semantics consistency between unmasked and inpainted masked regions.

while preserve its own frequency band well, than low-frequency band, owing to its robustness to the text prompt during the denoising process, while the previous works fail to disentangle all frequency bands during the denoising process especially for its early stage with high-level noise (see Sec.1 of the main body).

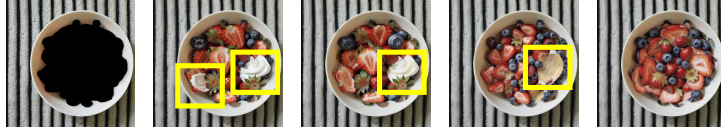
C More Ablation Studies on the Impact of Adaptively Extracting the Low-and-Mid Frequency Bands

As mentioned in Sec.2.3.2 and Sec.2.3.3 of the main body, we set th_{lp} in Eq.(7), and th_{mp1} and th_{mp2} in Eq.(10) to correlate with the ratio of unmasked regions. Due to page limitations, we provide further ablation studies on the impact of adaptively extracting low-and-mid frequency bands on the BrushBench and EditBench datasets using three variants: **Case I** adopts fixed thresholds to extract both mid-and-low frequency bands for different images; **Case II** adaptively extracts low-frequency band while using fixed thresholds for mid-frequency bands; **Case III** adaptively extracts mid-frequency band while using a fixed threshold for low-frequency extraction. As shown in Fig.3, due to the use of fixed thresholds for extracting mid-frequency band, **Case I** and **Case II** inevitably generate inconsistent content between masked and unmasked regions (e.g., the smiling cat in the fourth row of Fig. 3), which can be attributed to the fact that the larger unmasked regions require more mid-frequency band to encode the information from the low-frequency band substituted from the first null-text denoising process so that the low-frequency band under mid-frequency guidance for masked regions can achieve the consistency to the substituted unmasked regions for ground truth, as explained in Sec.2.4 of the main body. In contrast, **Case III** shows substantial performance degradation (e.g., the light yellow color blocks in the bowl in the first row of Fig.3), confirming that larger unmasked regions demand more low-frequency information from the null-text denoising process to replace the low-frequency bands in the text-guided denoising process, which is consistent with the findings in Sec.2.3.2 of the main body. These observations are also illustrated in Table.2.

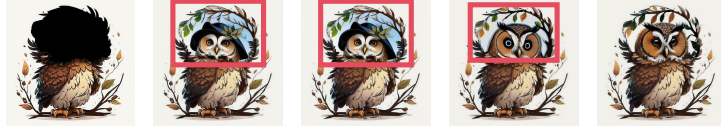
D More Ablation Studies on the Impact of Text and Null-Text Prompts on Low-and-Mid Frequency Bands during Denoising Process

As mentioned in Sec.3.3.1 of the main body, due to page limitation, we further provide more ablation studies on the impact of text and null-text prompts on low-and-mid frequency bands during denoising process on BrushBench and EditBench datasets with three variants: **TTN-Diff**: replacing the first null-text denoising process with the text-guided denoising process; **NTT-Diff**: replacing the last null-text denoising process with the text-guided denoising process; **TTT-Diff**: replacing both the

a bowl of strawberries and blueberries on a striped table



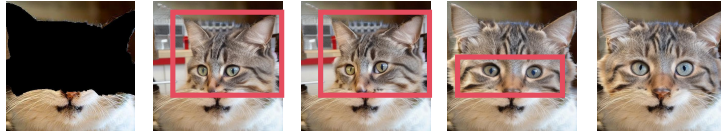
a cute owl sitting on a branch with leaves



two tents in the desert with a car parked in the background



a cat with a big smile on its face



Masked Image **Case I** **Case II** **Case III** **Ours (NTN-Diff)**

Figure 3: The inpainted output about the impact of adaptively extracting the low-and-mid frequency bands, NTN-Diff can achieve the better inpainted results (marked as the **box**) than others.

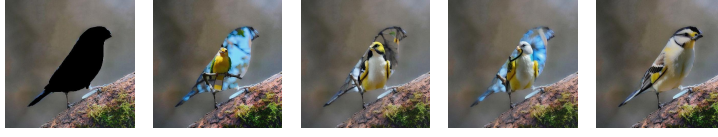
Table 2: Ablation studies on the impact of adaptively extracting the low-and-mid frequency bands, With setting th_{lp} in Eq.(7), and th_{mp1} and th_{mp2} in Eq.(10) to correlate with the ratio of unmasked regions, the best performance of image quality, unmasked region preservation and text alignment are achieved. **red** and **blue** stand for the best and second best result.

Dataset	Metric	Case I	Case II	Case III	Ours
EditBench	$IR_{\times 10} \uparrow$	2.98	3.02	3.07	3.10
	PSNR \uparrow	22.53	22.50	22.62	22.65
	LPIPS $\times 10^3 \downarrow$	25.12	25.20	24.55	24.21
	CLIP Score \uparrow	28.91	28.87	28.93	28.95
BrushBench	$IR_{\times 10} \uparrow$	10.04	10.10	10.95	11.12
	PSNR \uparrow	28.08	28.12	28.08	28.10
	LPIPS $\times 10^3 \downarrow$	44.18	44.25	44.15	44.09
	CLIP Score \uparrow	25.91	25.89	26.07	26.09

Table 3: Ablation studies on the impact of text and null-text prompts on low-and-mid frequency bands during denoising process: TTN-Diff, NTT-Diff and TTT-Diff for the early stage, based on the same text-guided denoising process (Sec.2.4 of the main body) for the late stage. **red** and **blue** stand for the best and second best result.

Dataset	Metric	TTN-Diff	NTT-Diff	TTT-Diff	Ours(NTN-Diff)
EditBench	$IR_{\times 10} \uparrow$	1.57	2.12	1.34	3.10
	PSNR \uparrow	22.57	22.58	22.51	22.65
	LPIPS $\times 10^3 \downarrow$	25.12	24.99	25.24	24.21
	CLIP Score \uparrow	28.87	28.89	28.81	28.95
BrushBench	$IR_{\times 10} \uparrow$	9.61	10.32	9.22	11.12
	PSNR \uparrow	28.06	28.08	28.01	28.10
	LPIPS $\times 10^3 \downarrow$	44.63	44.55	44.88	44.09
	CLIP Score \uparrow	25.94	25.97	25.89	26.09

a bird with a yellow and white face sitting on a tree branch



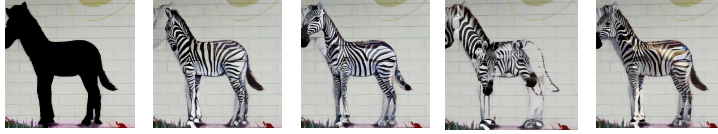
a painting of fish swimming in a coral reef



a teacup and saucer with a spoon and a spoon



a bee is shown in this poster



Masked
Image

TTN-Diff

NTT-Dif

TTT-Dif

Ours
(NTN-Diff)

Figure 4: The inpainted output about the impact of text and null-text prompts on low-and-mid frequency bands during denoising process, with the first null-text denoising process (Sec.2.3.1 of the main body), the second text-guided denoising process (Sec.2.3.2 of the main body) and the last null-text denoising process (Sec.2.3.3 of the main body), NTN-Diff can achieve the better inpainted results than others.

first and last null-text denoising processes with the text-guided denoising processes; Table.3 suggests that our **NTN-Diff** outperforms **NTT-Diff**, despite the denoised mid-frequency band well aligns with the text prompts especially for masked regions, while also encodes the information related to low-frequency band from the first null-text denoising process, which verifies that *the last null-text denoising process can achieve semantics consistency between mid-and-low frequency bands across masked and unmasked regions, by denoising the low-frequency band throughout the path to be semantically consistent to mid-frequency band, with no influence from text prompts*, which is in line with Sec.2.3.3 of the main body; **TTN-Diff** exhibits a substantial performance degradation, confirming that *the null-text denoising process conditioned on null-text prompt can avoid being influenced by text prompts even under the high-level noise, focusing primarily on low-frequency band*, hence validating the important the first null-text denoising process (Sec.2.3.1 of the main body). We illustrate the above intuitions in Fig.4.

E Additional Visual Results for the Ablation Study in Sec.3.3.2 of the Main Body

As mentioned in Sec.3.3.2 of the main body, due to page limitation, we further provide more visual results to analyse the parameter λ in the denoising process (Sec.2.2 of the main body), which is utilized to divide the denoising process into *early* and *late* stages, demarcated by the critical step λT . see Fig.5. When $\lambda = 0.9$, the performance is the worst with the shortest early stage, such as blue background in the 4th row. When $\lambda = 0.6$ with the balance of early and late stage, the best performance of image quality, unmasked region preservation and text alignment are achieved, such as the face of the cat in the first rows, confirming the rational that *the denoised low-frequency band for*

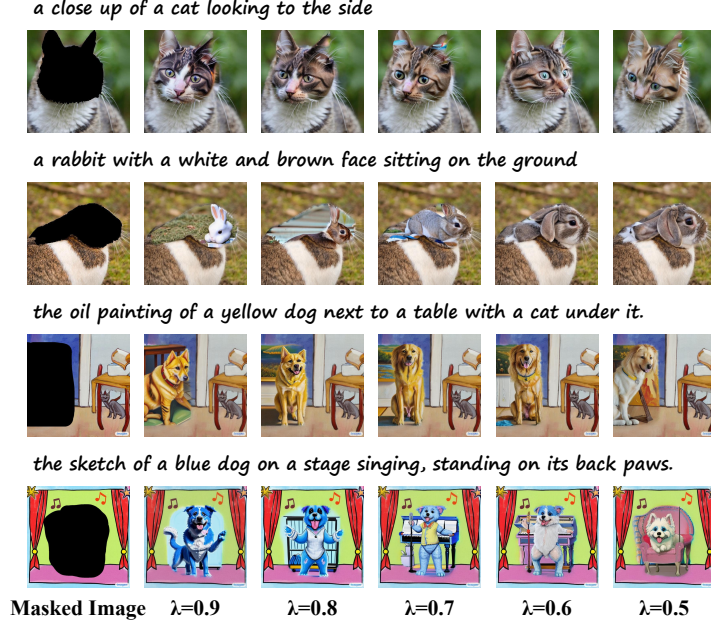


Figure 5: The inpainted output about hyperparameter sensitivity analysis of λ for the length of the early and late stage, when $\lambda = 0.6$ with the balance of early and late stage, NTN-Diff can achieve the better inpainted results than others.

the masked regions is guided by the denoised mid-frequency band within the whole early stage, which guided by text prompts for a few number of steps from text-guided late denoising process to avoid the large influence from text prompts to be inconsistent for unmasked regions, thus the desirable length of early stage can make the low-frequency band under the mid-frequency guidance for masked regions achieve the consistency to the substituted unmasked regions for ground truth, which is consistent to Sec.2.4 of the main body.

F Limitations and Broader Impacts

Limitations. Unlike previous state-of-the-art methods that fine-tune Stable Diffusion (v1.5) on large-scale datasets for text-guided image inpainting, which typically require extensive time and computational resources for training, our method introduces a plug-and-play frequency-aware null-text-null diffusion framework. This framework substitutes mid-and-low frequency bands during the early stage of the denoising process and, as a result, involves three parallel null-text-null denoising processes in the early stage, resulting in a larger computational cost than single process. Nevertheless, it shares the same order of magnitude as the previous methods.

Broader Impacts. Our NTN-Diff can achieve the masked regions to be generated according to the text prompt, while preserve the unmasked regions, our method may raise certain ethical concerns. The inpainted image could potentially be misused in the creation of misleading information. We strongly advocate for the establishment of clear accountability in the use of such technologies, along with enhanced legal and technical oversight, to ensure they are applied responsibly.

G Code

We implement the proposed NTN-Diff in pytorch framework under the running environment as: python3.6, pytorch1.7 and cuda10.0. The codes are available in the package **code.zip**.

References

- [1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.
- [2] Xiang Gao and Jiaying Liu. Fbsdiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4101–4109, 2024.
- [3] Xiang Gao, Zhengbo Xu, Junhan Zhao, and Jiaying Liu. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1824–1832, 2024.
- [4] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion, 2024.
- [5] Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, , Junhao Zhuang, Ying Shan, Yuexian Zou, and Qiang Xu. Brushedit: All-in-one image inpainting and editing, 2024.
- [6] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2312.14091*, 2023.
- [7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [8] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18359–18369, 2023.
- [9] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [10] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.
- [11] Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, and Yunhe Wang. Learning frequency-aware dynamic network for efficient super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4308–4317, 2021.
- [12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [14] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*, 2023.