# One Stone with Two Birds: A Null-Text-Null Frequency-Aware Diffusion Models for Text-Guided Image Inpainting

Haipeng Liu†, Yang Wang† *, Meng Wang

School of Computer Science and Information Engineering
Hefei University of Technology, China

† Equal contribution, * Yang Wang is the corresponding author

Paper Code

NEURAL INFORMATION PROCESSING SYSTEMS

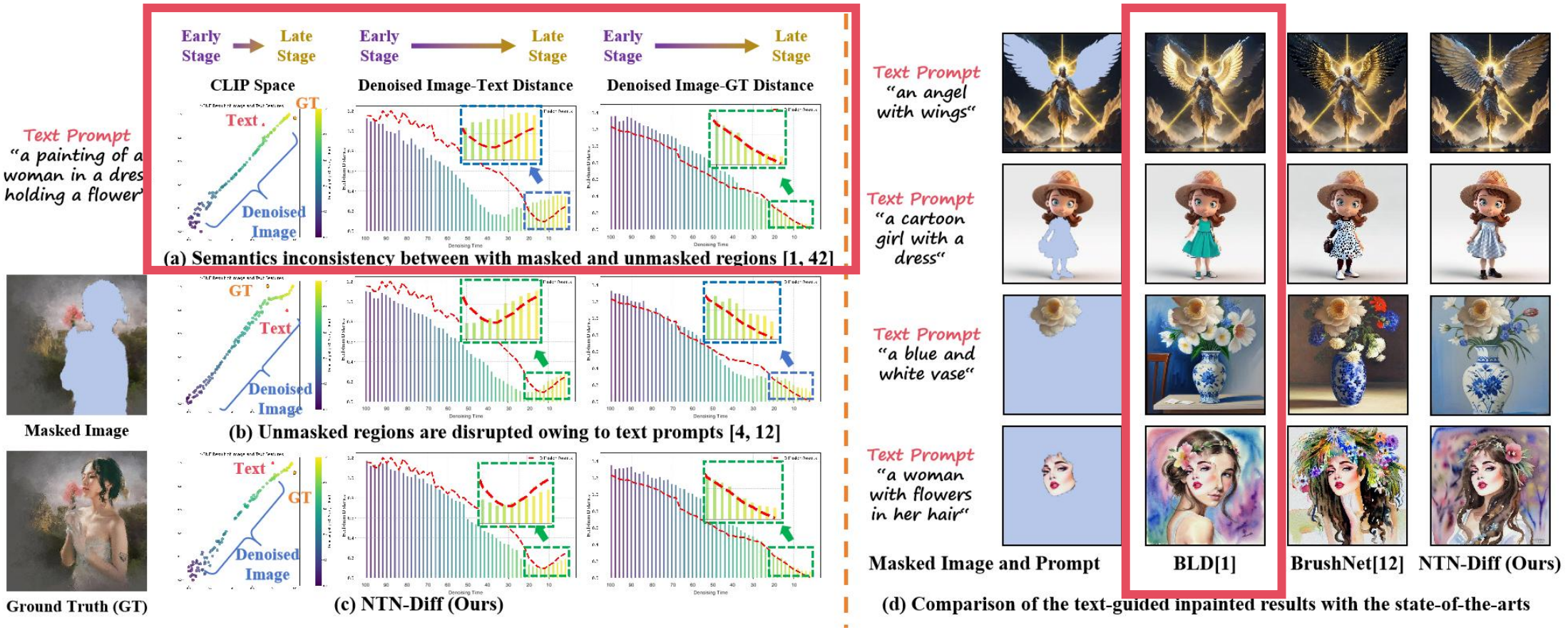
# Background

- **Text-guided Image Inpainting** -- *Inpainting the masked regions of the image according to the text prompt.*

**Challenges**: Upon the alignment between the generated content for masked regions and the text prompts:

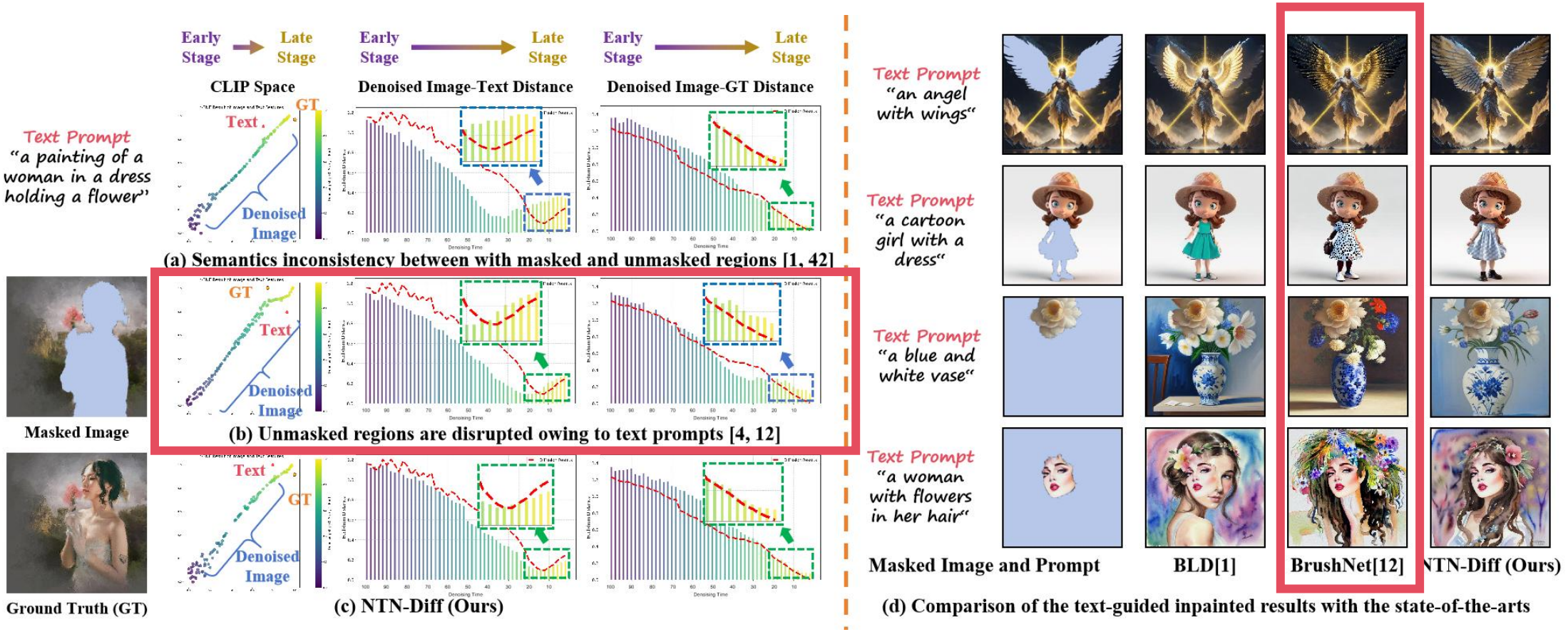
- The preservation for unmasked regions
- Achieving the semantics consistency between unmasked and masked regions as inpainted.

# Motivation




- Suffering from the semantics inconsistency between with masked and unmasked regions, owing to the discrepancy from the diffusion process and the inpainted maksed regions from the text-guided denoising process.

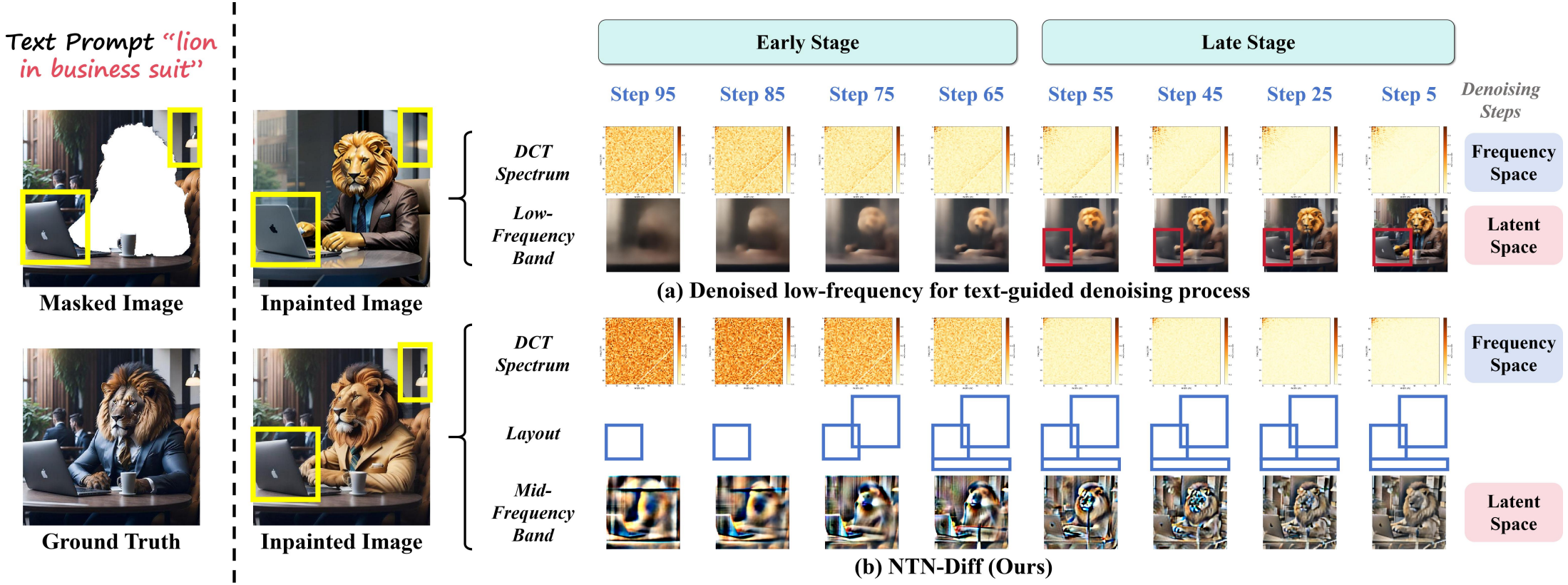
# Motivation




- The unmasked regions fail to be preserved, which is incurred by the other text-guided denoising process to inpaint masked regions.

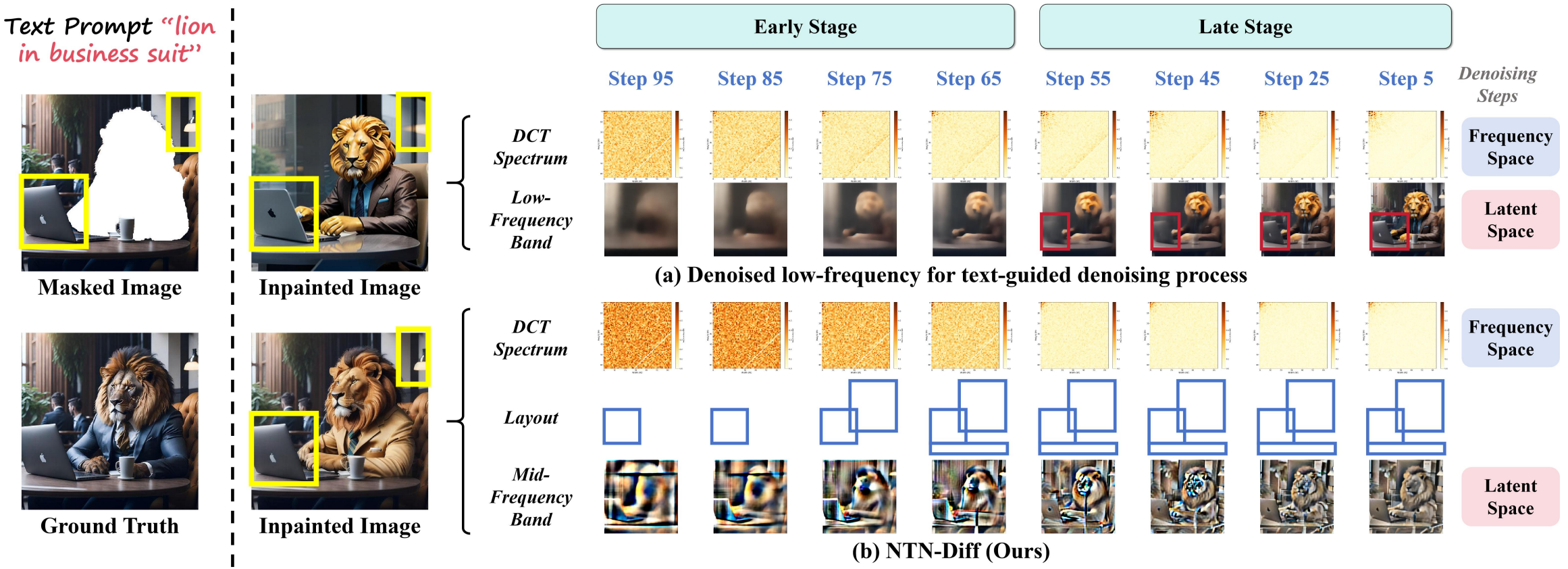
# Motivation

- The low-frequency band for both masked and unmasked regions are easy to be modulated by text prompts, as illustrated in (a).
- To be contrary, as seen in (b), the mid-frequency band across all regions is robust to the text prompts while aligns well with text prompts,which may better preserve the unmasked regions than low-frequency band upon text prompts.



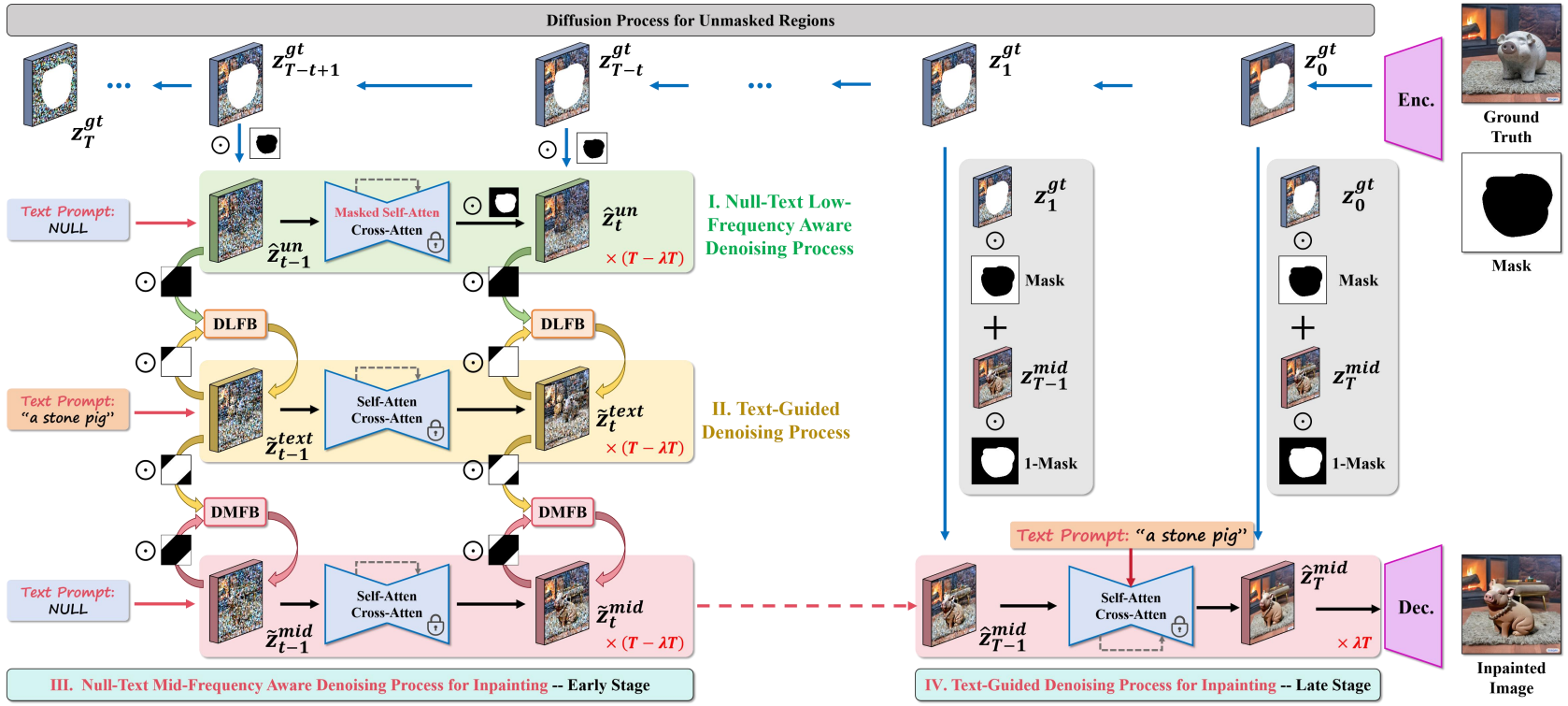
# Motivation

- The low-frequency band for both masked and unmasked regions are easy to be modulated by text prompts, as illustrated in (a).
- To be contrary, as seen in (b), the mid-frequency band across all regions is robust to the text prompts while aligns well with text prompts,which may better preserve the unmasked regions than low-frequency band upon text prompts.



# Methodology

- How to disentangle different frequency bands, particularly the early stage of the denoising process with high-level noise?
- How to exploit the hybrid frequency bands for diffusion models to simultaneously achieve the above two goals?

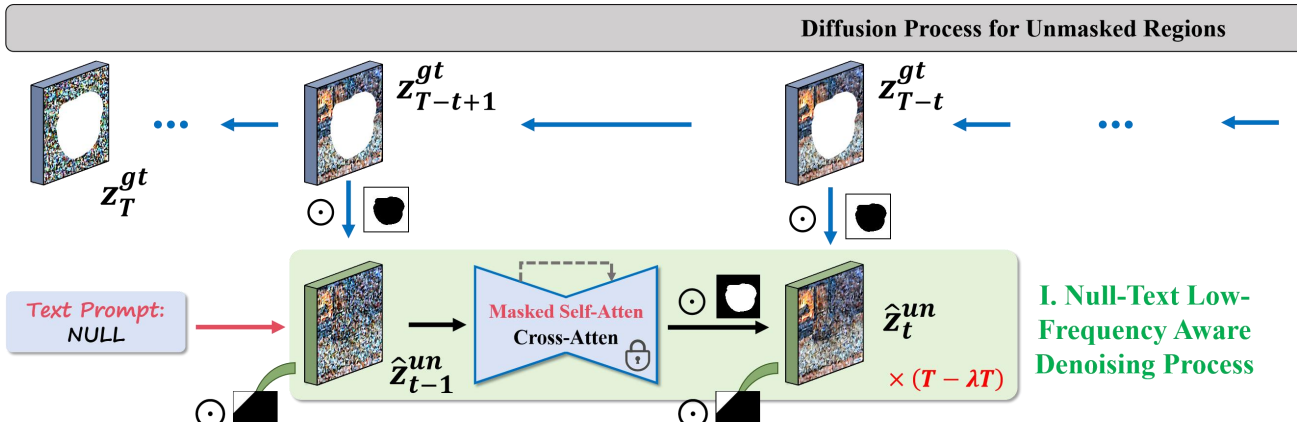



**The frequency-aware null-text-null diffusion models**

# Methodology

1. **Null-Text Low-Frequency Aware Denoising Process**: ***avoiding the low-frequency band is influenced by text prompts under the high-level noise***, *then replace its denoised result with unmasked regions from the forward diffused results.*

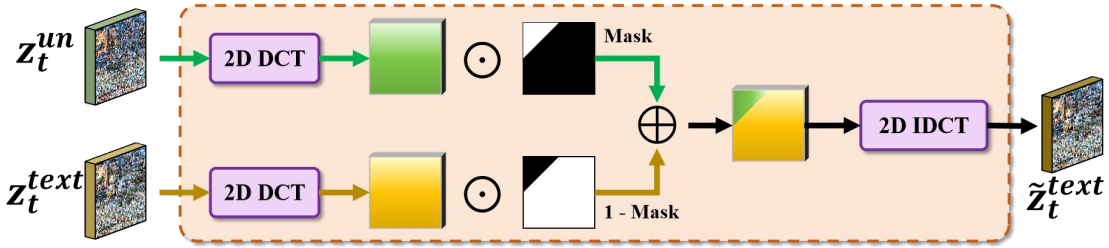
$$\hat{z}_t^{un} = z_{T-t}^{gt} \odot m_z + z_t^{un} \odot (1 - m_z),$$



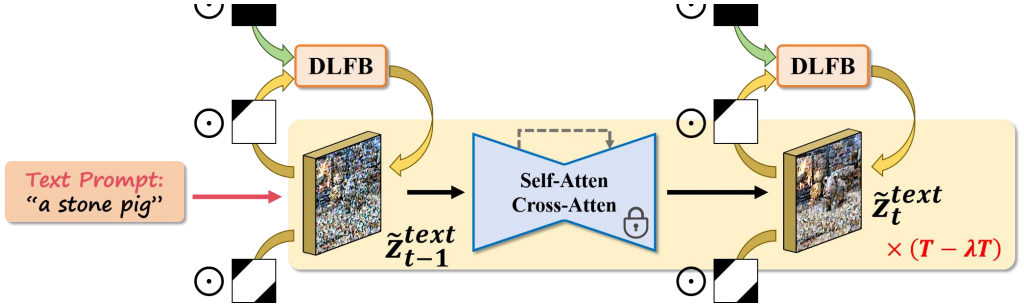
# Methodology

2. **Text-Guided Denoising Process**: *Resorting to the text-guided denoising process, together with low-frequency band substitution from null-text denoising process.*

$$\tilde{z}_t^{text} = \mathrm{IDCT}\left(\mathrm{DCT}(z_t^{un}) \odot m_{low} + \mathrm{DCT}(z_t^{text}) \odot (1 - m_{low})\right)$$




(a) Denoised Low-Frequency Band layer




II. Text-Guided Denoising Process

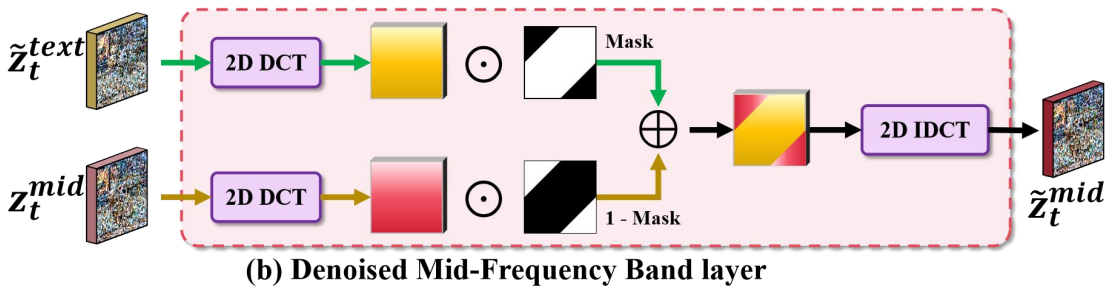
**Remark1**: *the low-frequency band across both masked and unmasked regions are preserved even under text prompts owing to the substitutions from the null-text denoising process, with only mid-frequency aligned with text prompts for both regions,*

➢ ***hence still failed to achieve the consistency between masked and unmasked regions.***

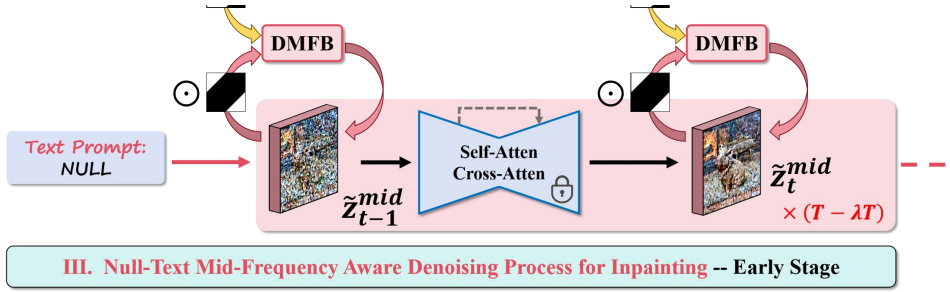
# Methodology

3. **Null-Text Mid-Frequency Aware Denoising Process**: *exploit the above denoised mid-frequency to guide the last null-text denoising process, by substituting mid-frequency band from this null-text denoising process,*

$$\tilde{z}_t^{in} = \mathrm{IDCT}\left(\mathrm{DCT}(z_t^{text}) \odot m_{\mathrm{mid}} + \mathrm{DCT}(z_t^{in}) \odot (1 - m_{\mathrm{mid}})\right)$$




(b) Denoised Mid-Frequency Band layer



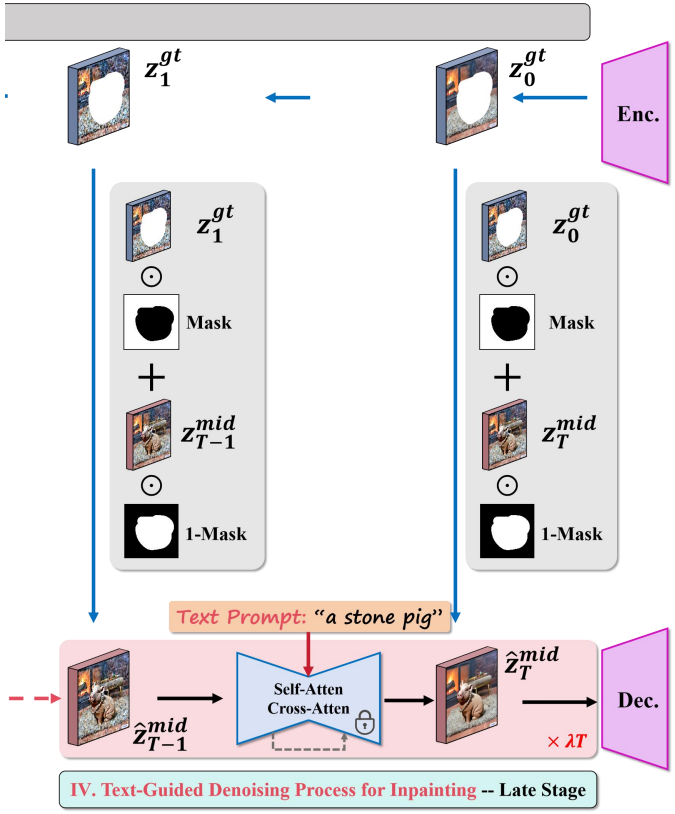

III. Null-Text Mid-Frequency Aware Denoising Process for Inpainting -- Early Stage

**Remark2**: *The above null-text mid-frequency guided denoised process can denoise* ***the low-frequency band for the masked regions to be aligned with text prompt***

- *However, cannot be served as the final text-guided inpainting output for masked regions, as the denoised mid-and-low frequency band is not semantically strong as text prompt.*

# Methodology

## Late Stage of Text-Guided Denoising Process




$$\hat{z}_t^{in} = z_{T-t}^{gt} \odot m_z + z_t^{in} \odot (1 - m_z)$$

**Remark3**: ***How the denoised results within this late stage can achieve semantics consistency between masked and unmasked regions?***

- *the denoised mid-frequency band from the early text-guided denoising process also encodes the information from the low-frequency band substituted from the null-text denoising process to match diffusion process (ground truth) including both masked and unmasked regions*

# Experiments

## Qualitative Comparisons


(a) BrushBench


(b) EditBench

# Experiments

## Quantitative Comparisons

| Metrics | | | Image Quality | | Masked Region Preservation | | | Text Align |
|---|---|---|---|---|---|---|---|---|
| Task | Models | Venue | IR $_{\times 10}$ ↑ | HPS v2 $_{\times 10^2}$ ↑ | PSNR↑ | MSE $_{\times 10^3}$ ↓ | LPIPS $_{\times 10^3}$ ↓ | CLIP Score↑ |
| Inside Inpainting | BLD[1] | TOG' 23 | 9.78 | 25.87 | 21.33 | 9.76 | 49.26 | 26.15 |
| | CNI[47] | ICCV' 23 | 9.9 | 26.02 | 12.39 | 78.78 | 243.62 | 26.47 |
| | PP[49] | ECCV' 24 | 11.46 | 27.35 | 21.43 | 32.73 | 48.43 | **26.48** |
| | BrushNet[12] | ECCV' 24 | **12.36** | **27.40** | 21.65 | **9.31** | 48.28 | **26.48** |
| | HDP[25] | ICLR' 25 | 11.68 | 26.90 | **22.61** | 9.95 | **43.50** | 26.37 |
| | NTN-Diff (Ours) | - | **12.45** | **27.57** | **23.51** | **6.50** | **40.79** | **26.54** |
| | CNI*[47] | ICCV' 23 | 11.21 | 26.92 | 22.73 | 24.58 | 43.49 | 26.22 |
| | BrushNet*[12] | ECCV' 24 | **12.64** | **27.78** | **31.94** | **0.80** | **18.67** | **26.39** |
| | NTN-Diff* (Ours) | - | **12.69** | **27.82** | **40.70** | **0.11** | **0.88** | **26.49** |
| Outside Inpainting | BLD[1] | TOG' 23 | 7.81 | 26.77 | 15.85 | 35.86 | 21.40 | 26.73 |
| | CNI[47] | ICCV' 23 | 9.26 | 27.68 | 11.91 | 83.03 | 58.16 | 27.29 |
| | PP[49] | ECCV' 24 | 7.45 | 28.01 | 18.04 | 31.78 | 15.13 | 26.72 |
| | BrushNet[12] | ECCV' 24 | **10.82** | **28.02** | **18.06** | 22.86 | **15.08** | **27.33** |
| | HDP[25] | ICLR' 25 | 9.66 | 27.79 | 18.03 | 22.99 | 15.22 | 26.96 |
| | NTN-Diff (Ours) | - | **11.54** | **28.22** | **18.47** | **20.44** | **14.46** | **27.54** |
| | CNI*[47] | ICCV' 23 | 9.57 | 27.76 | 17.50 | 37.72 | 19.95 | 26.92 |
| | BrushNet*[12] | ECCV' 24 | **10.88** | **28.09** | **27.82** | **2.25** | **4.63** | **27.22** |
| | NTN-Diff* (Ours) | - | **11.61** | **28.36** | **31.08** | **1.23** | **1.24** | **27.30** |

# Conclusion

In this paper, we propose a null-text-null frequency-aware diffusion models, named NTN-Diff, for text-guided image inpainting, by decomposing the semantics consistency across masked and unmasked regions into the consistencies as per each frequency band, while preserving the unmasked regions, to simultaneously address two challenges of unmasked regions preservations, along with its semantics consistency with inpainted masked regions.

# Thanks for Listening

Paper

Code

NEURAL INFORMATION PROCESSING SYSTEMS