Neural Image Compression with Text-guided Encoding for both Pixel-level and Perceptual Fidelity

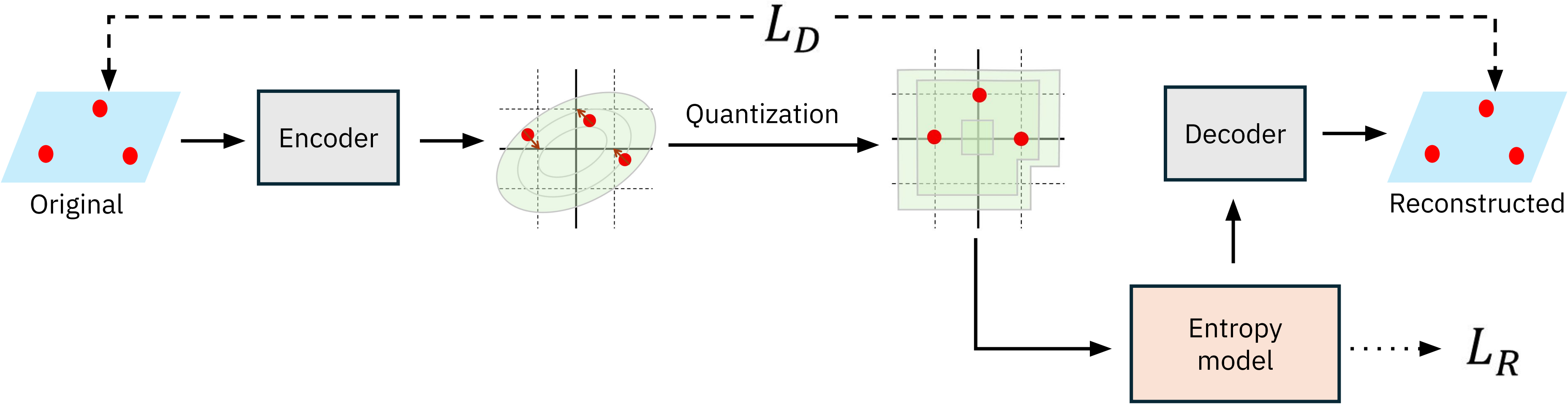
Hagyeong Lee[1*] , Minkyu Kim[1*], Jun-Hyuk Kim[2], Seungeon Kim[2], Dokwan Oh[2], Jaeho Lee[1]

(*Equal Contribution*)

[1] POSTECH, [2] Samsung Advanced Institute of Technology

Background: Neural image compression

Goal. Achieves higher pixel-level and perceptual fidelity both



$$L = L_R + \lambda \cdot L_D$$

Background: Neural image compression

Goal. Achieves higher pixel-level and perceptual fidelity both

Original (kodim04.png)

[1] LIC-TCM (bpp: 0.12)

[2] MS-ILLM (bpp: 0.13)

[1] Liu et al., "Learned Image Compression with Mixed Transformer-CNN Architectures," CVPR 2023.

[2] Muckley et al., "Improving Statistical Fidelity for Neural Image Compression with Implicit Local Likelihood Models," ICML 2023.

Motivation

Recent compression works ([1], [2]) improve perceptual quality

by using text-guided generation model. (e.g. Diffusion model)

Previous approach
(MS-ILLM)

High realistic

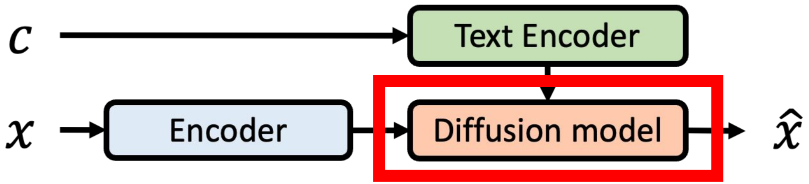
Diffusion-based approaches
(Text+Sketch, PerCo)

[1] Careil et al., "Towards image compression with perfect realism at ultralow bitrates," ICLR 2024.

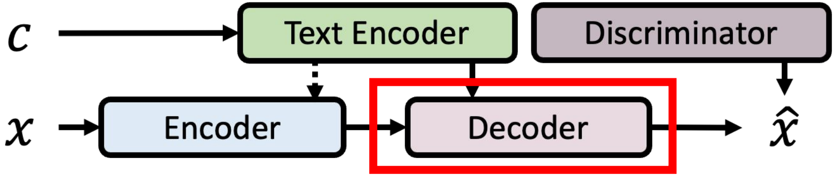
[2] Lei et al., "Text + Sketch: Image Compression at Ultra Low Rates," arXiv 2023.

Motivation

They utilize text using in decoding phase of image compression.



Text-guided decoding
with diffusion-based decoders



Text-guided decoder utilizing GAN

Motivation

Limitations of text-guided decoding are **inconsistency** and low pixel-fidelity.

`"Two parrots standing next to each other with leaves in the background".`

Original

Reconstructions by Diffusion-based approach
(PerCo)

Motivation

Limitations of text-guided decoding are **<u>inconsistency</u>** and low pixel-fidelity.

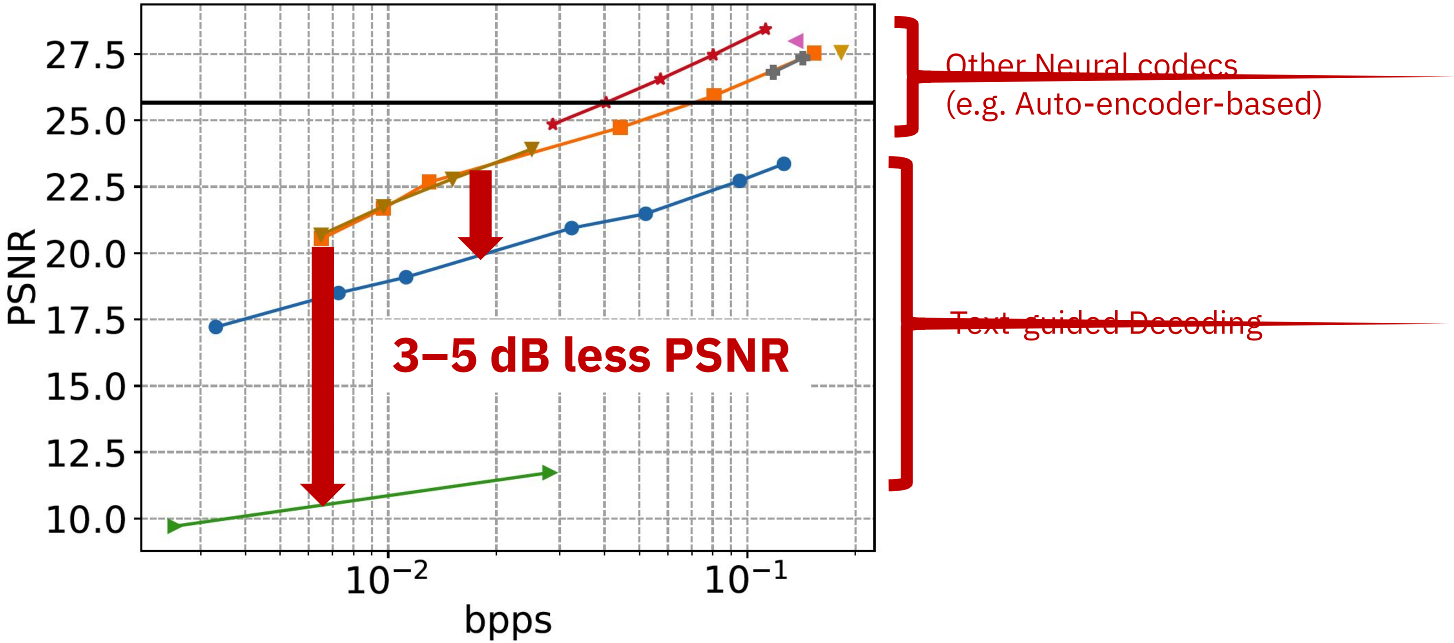
Original

$\neq$

Reconstructions by Diffusion-based approaches
(Text+Sketch, PerCo)

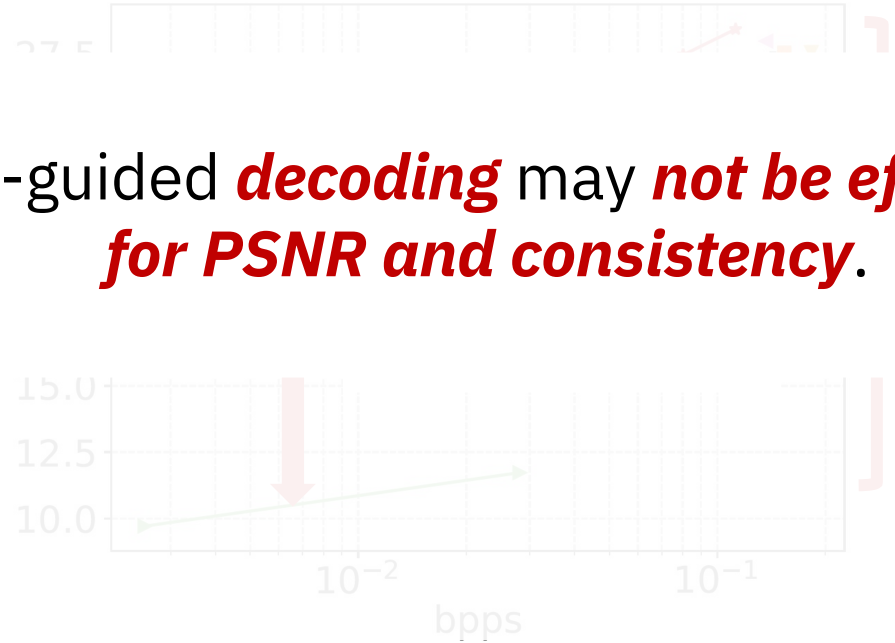
Motivation

Limitations of text-guided decoding are inconsistency and **<u>low pixel-fidelity.</u>**



Motivation

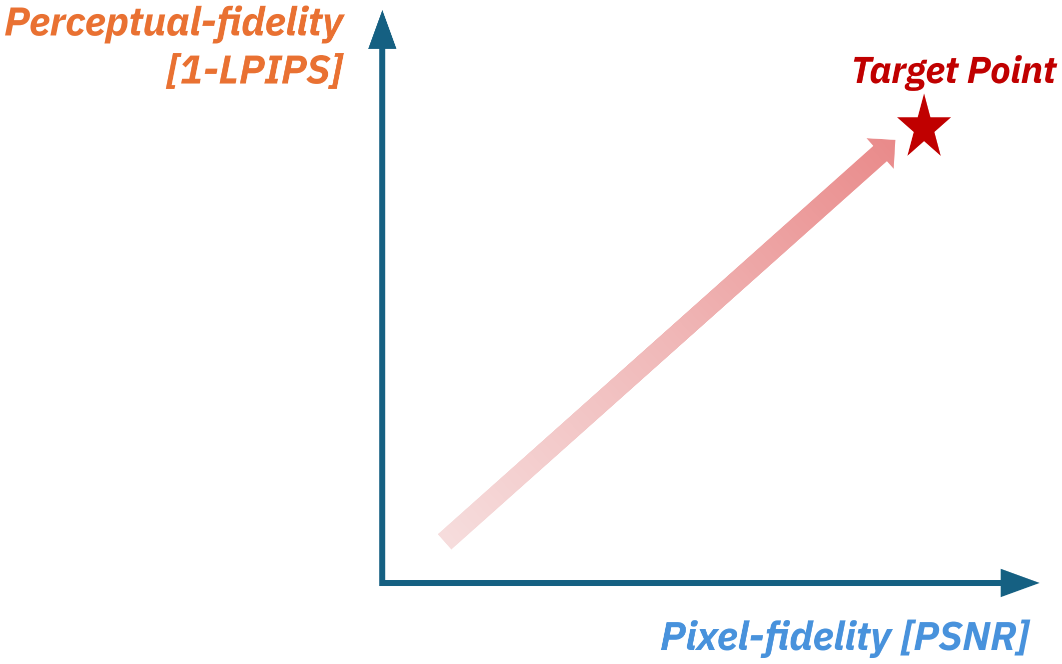
Limitations of text-guided decoding are inconsistency and __low pixel-fidelity.__



Text-guided ***decoding*** may ***not be effective for PSNR and consistency***.

Motivation

Propose a text-guided method for achieving high pixel and perceptual fidelity.

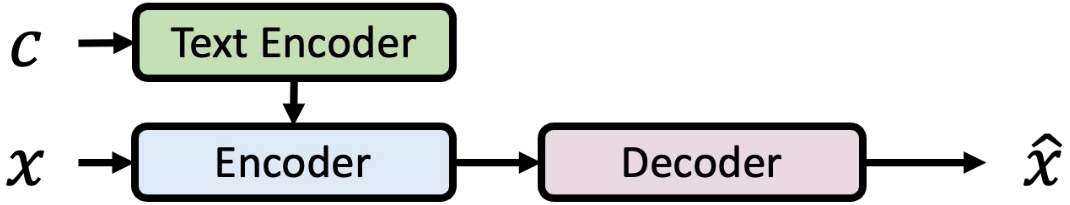


Text **A**daptive **CO**mpression

→ TACO 🌮

Text Adaptive Compression

Idea. Using text when encoding the image.



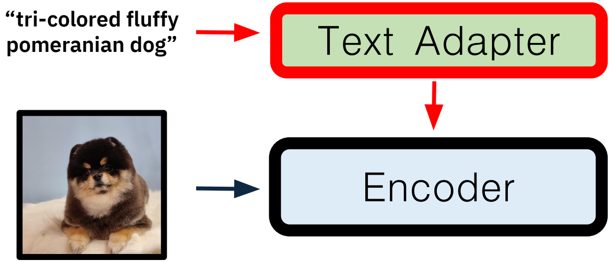
* c means text caption, x means target (original) image, $\hat{x}$ means reconstructed (compressed) image.

Overall framework

Text Adaptive Compression

Idea. Using text when encoding the image.

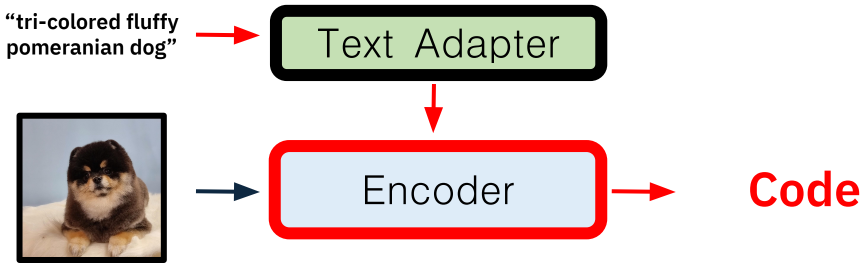
- Inspired by how humans perceive images using language.



Text Adaptive Compression

Idea. Using text when encoding the image.

- Inspired by how humans perceive images using language.
  - Encoded image feature *contains additional semantic information*.



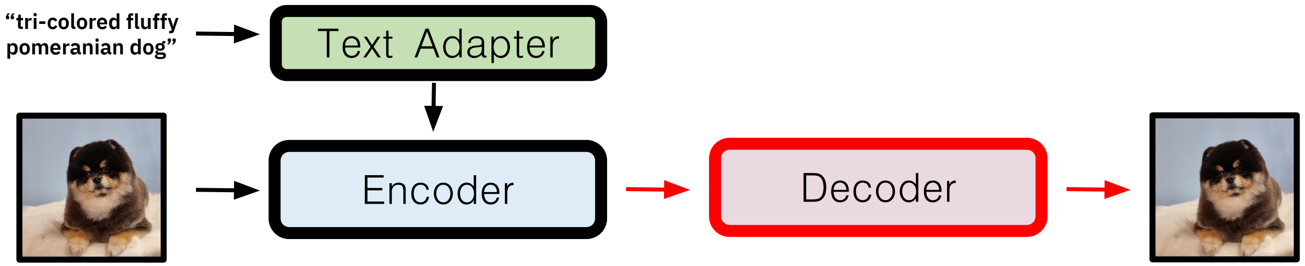
Text Adaptive Compression

Idea. Using text when encoding the image.

- During the decoding, only the image latent feature is processed.

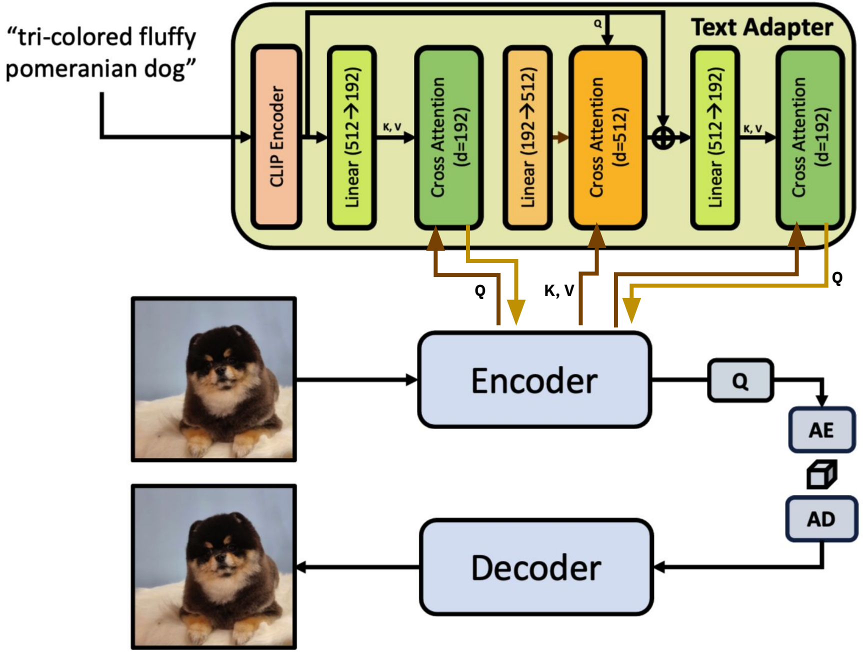
  → Reduce the <u>pixel-level distortion</u>

  → Improve the pixel fidelity



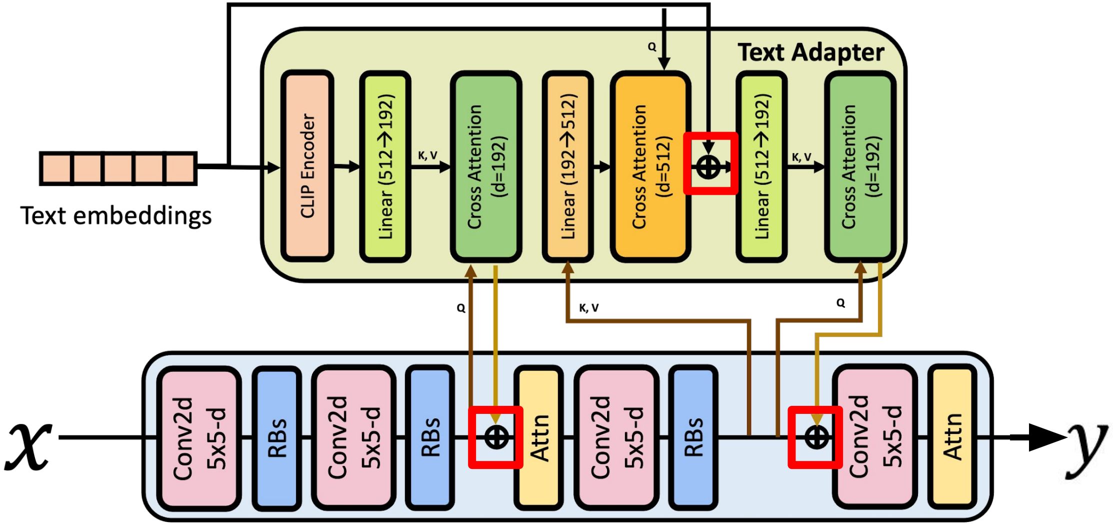
Text Adaptive Compression

TACO transforms a popular PSNR-oriented neural codec architecture into a text-guided one by augmenting the encoder with a text adapter.



Text Adapter

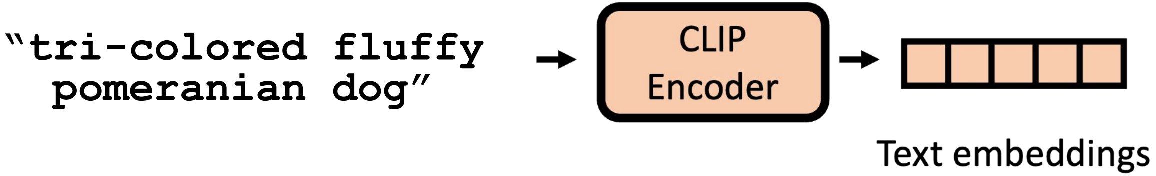
Bi-directional attention injects textual information into the latent code.



Text Adapter with encoder

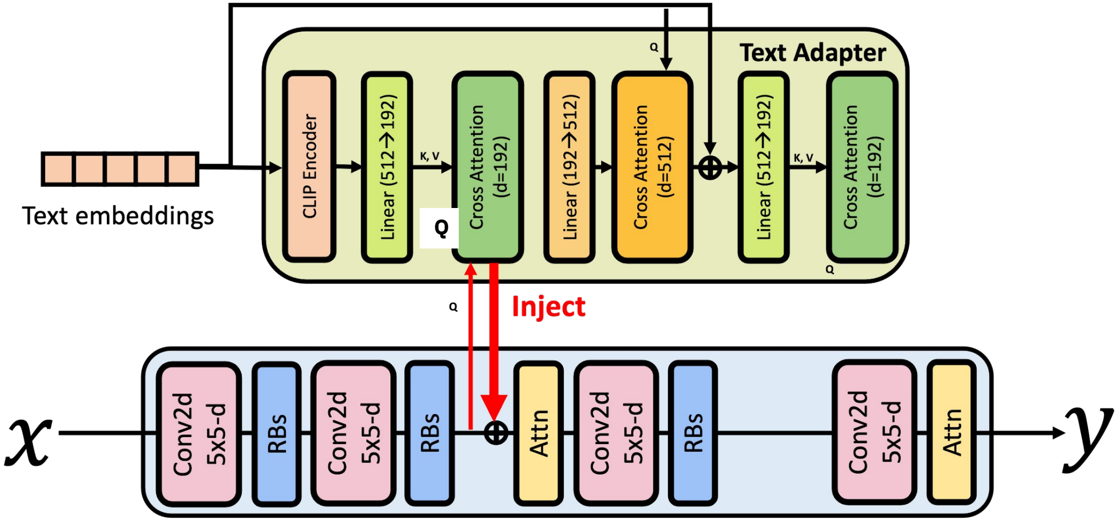
Text Adapter

Text embeddings are generated from (pre-trained) CLIP.



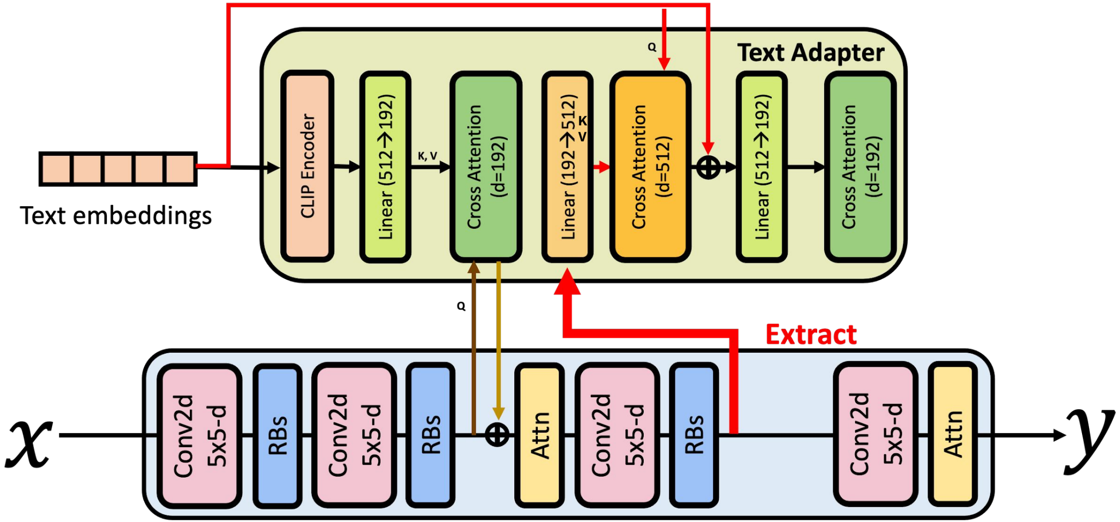
Text Adapter

Inject text information to image latent via cross-attention.
(CA computes query from image latent and key, values from text embeddings.)



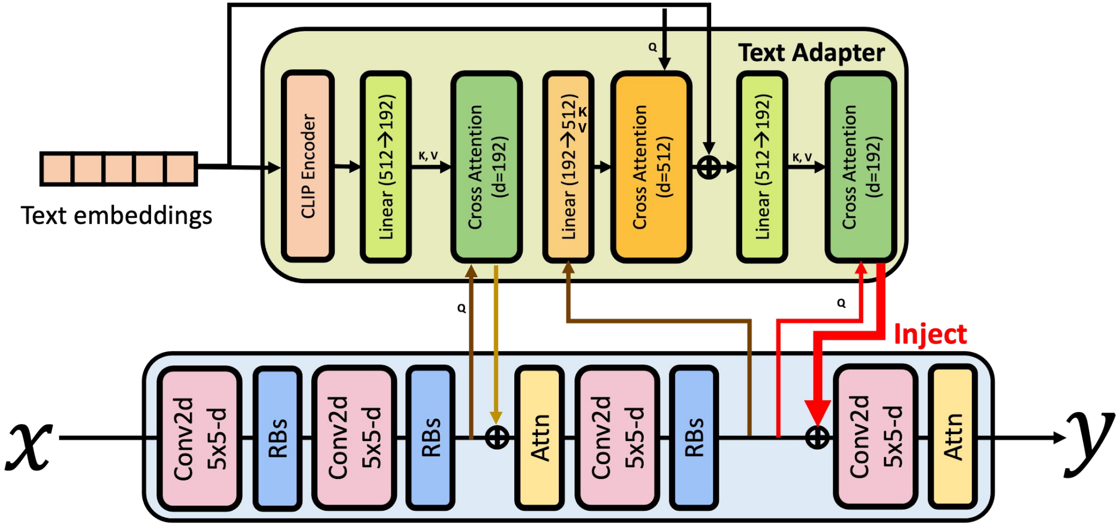
Text Adapter

Extract compressed image feature and incorporate with text using cross-attention.
(CA computes queries from the text and keys/values from the image.)



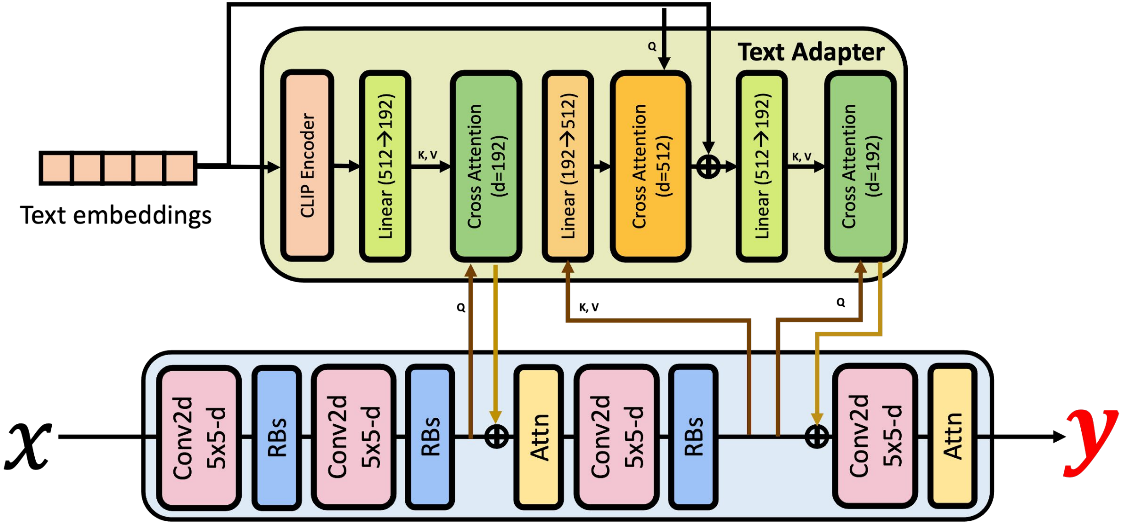
Text Adapter

Injecting the text embeddings into an image latent via cross-attention.
(Textual information is updated by image latent & image latent is down-sampled.)



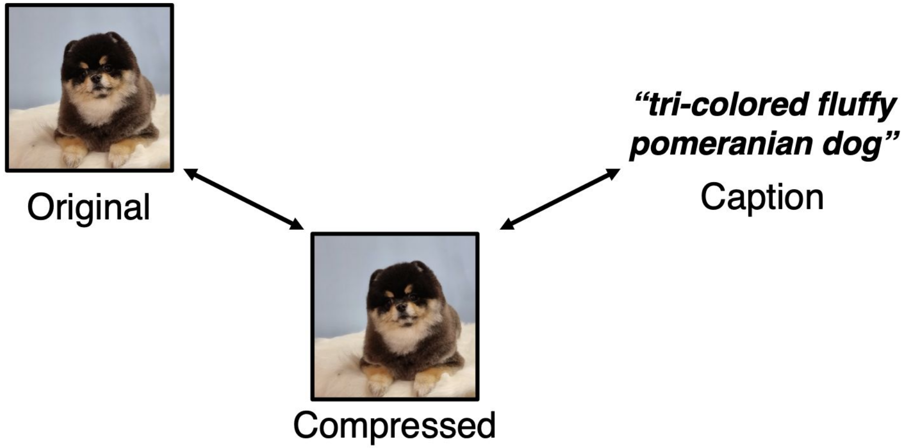
Text Adapter

Finally, the encoder generates a joint image-text latent feature (y).



Joint image-text loss

Train the model to compress the image better by leveraging text information.

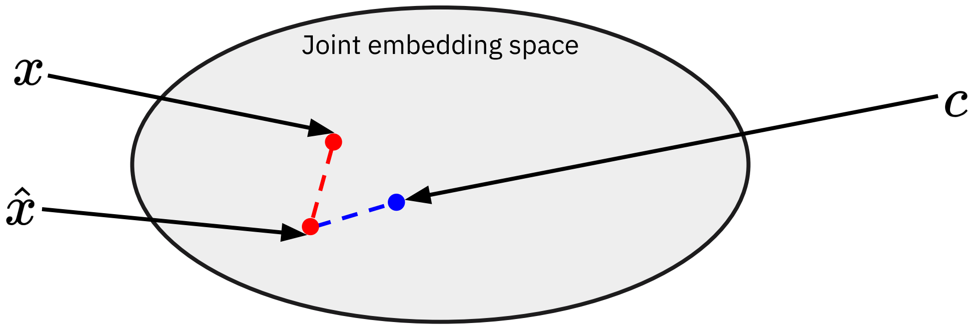


Joint image-text loss

Reduce the semantic distances by penalizing two terms:

1. Original image & Compressed image
2. Compressed image & Text description

❈ Semantic Distance is measured in the joint embedding space of CLIP.



* c means text caption, x means target (original) image, $\hat{x}$ means reconstructed (compressed) image.

Joint image-text loss

Reduce the semantic distances by penalizing two terms:

1. Original image & Compressed image
2. Compressed image & Text description

❈ Semantic Distance is measured in the joint embedding space of CLIP.

$$L_{\mathrm{j}}(x, \hat{x}, c) = L_{\mathrm{con}}\big(f_{\mathrm{I}}(\hat{x}), f_{\mathrm{T}}(c)\big) + \beta \cdot |f_{\mathrm{I}}(x) - f_{\mathrm{I}}(\hat{x})|_2$$

* c means text caption, x means target (original) image, $\hat{x}$ means reconstructed (compressed) image.

* L_{con} means contrastive loss used in CLIP.

Experimental Setup

Train Dataset. MS-COCO Train Set

- Contains 82,783 images with 5 human-annotated captions for each image

https://cocodataset.org/#explore

a cat drinking out of a glass on top of a table.
a cat is drinking something from a glass.
a cat stands on a table drinking water out of a glass
a grey colored cat that is drinking from a glass of water.
a cat drinking ice water out of a glass.

Example of train data

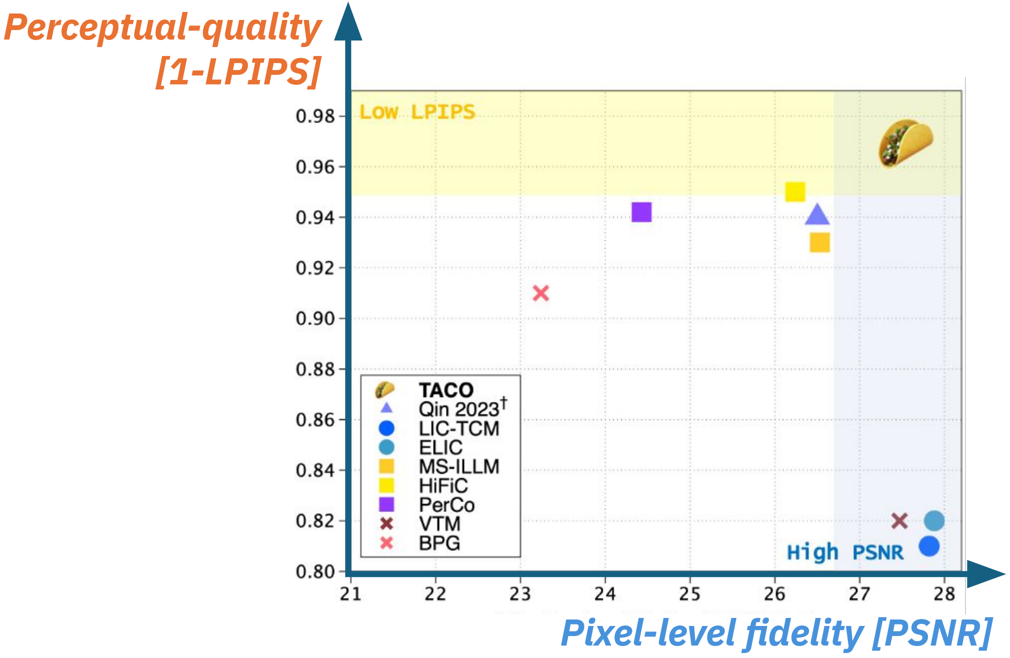
Experimental Setup

To compare with other neural image codecs, we set up the following settings:

- **Baselines**
  - **PSNR-focused.** LIC-TCM (CVPR’ 23), ELIC(CVPR’ 22)
  - **Perceptual-focused.** PerCo (ICLR’ 24), MS-ILLM (ICML’ 23), HiFiC (NeurIPS’ 20)
- **Metrics**
  - PSNR
  - LPIPS
  - FID
- **Evaluation Datasets**
  - MS-COCO 30K (Human-annotated caption)
  - CLIC (Machine-generated caption)
    - Caption is generated by OFA (ICML’ 22)

Result: Overview

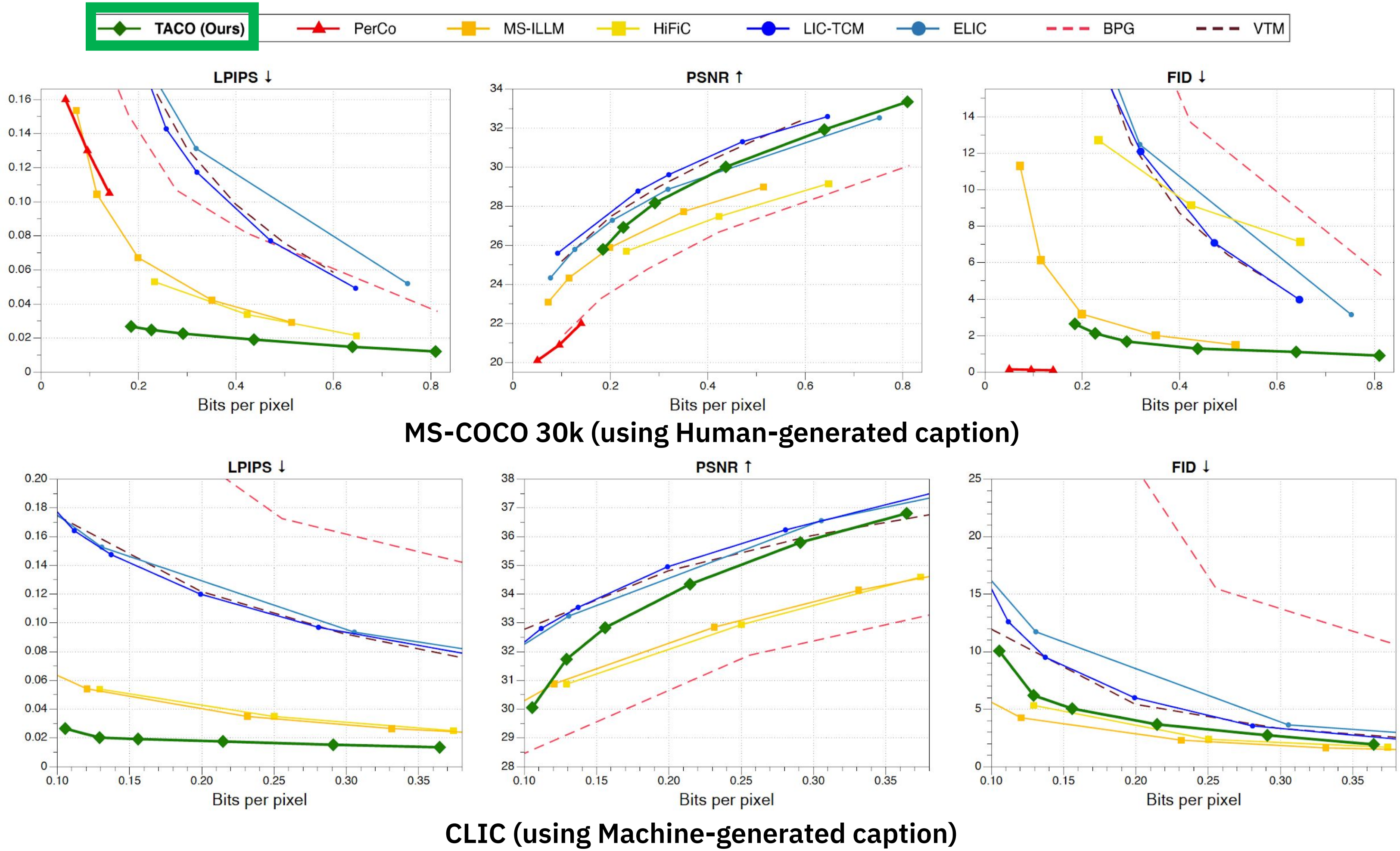
TACO achieves both high pixel-level and perceptual quality.



Result: vs. Image Compression Codecs

On all tested datasets (MS-COCO 30K, CLIC), TACO is...

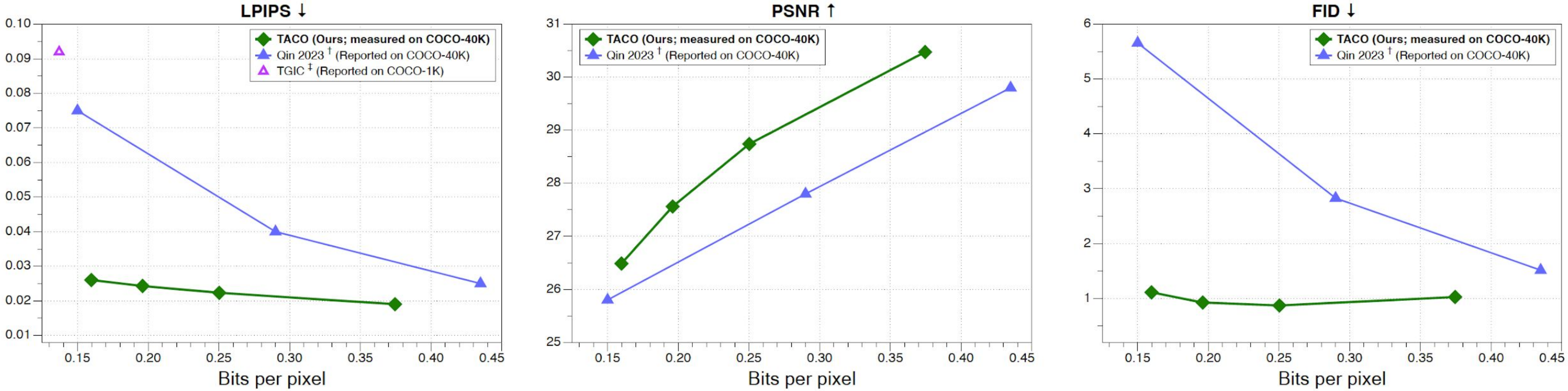
- **Perceptual-fidelity (LPIPS).** Outperforms all baselines!
- **Pixel-fidelity (PSNR).** Competitive with PSNR-focused, beats Perceptual-focused


TACO (Ours)
PerCo
MS-ILLM
HiFiC
LIC-TCM
ELIC
BPG
VTM
LPIPS ↓
PSNR ↑
FID ↓
Bits per pixel
MS-COCO 30k (using Human-generated caption)
LPIPS ↓
PSNR ↑
FID ↓
Bits per pixel
CLIC (using Machine-generated caption)

Result: vs. Image Compression Codecs

TACO achieves much better than the previous text-guided decoding baseline.

- Prevent the degradations in PSNR
- Achieve better LPIPS and FID

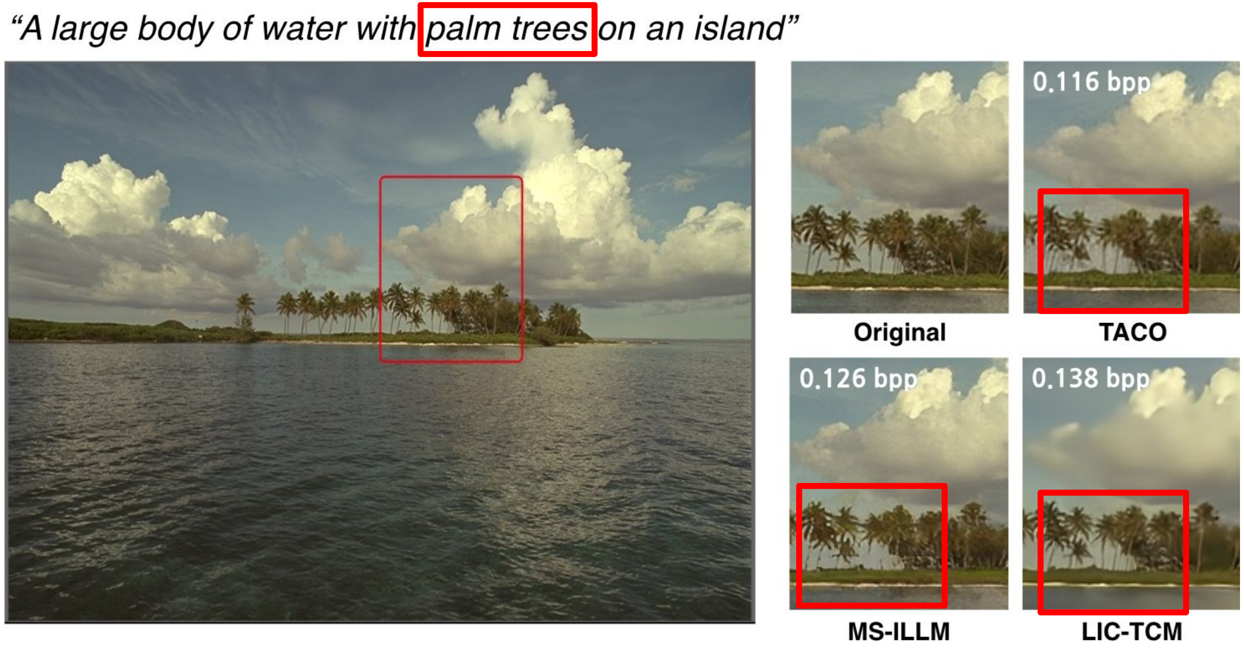


[1] Jiang et al., "Multi-Modality Deep Network for Extreme Learned Image Compression," AAAI 2023.

[2] Qin et al., "Perceptual image compression with cooperative cross-modal side information," arXiv 2023.

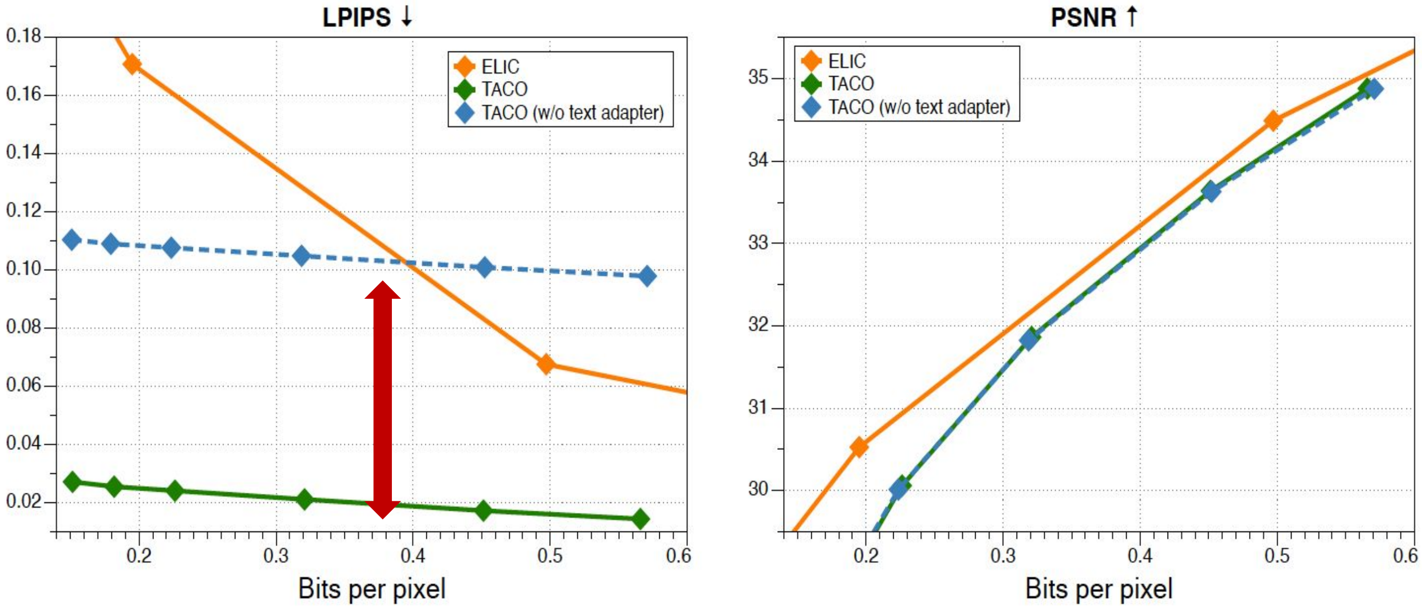
Qualitative results

TACO improves reconstruction significantly by focusing on captions.



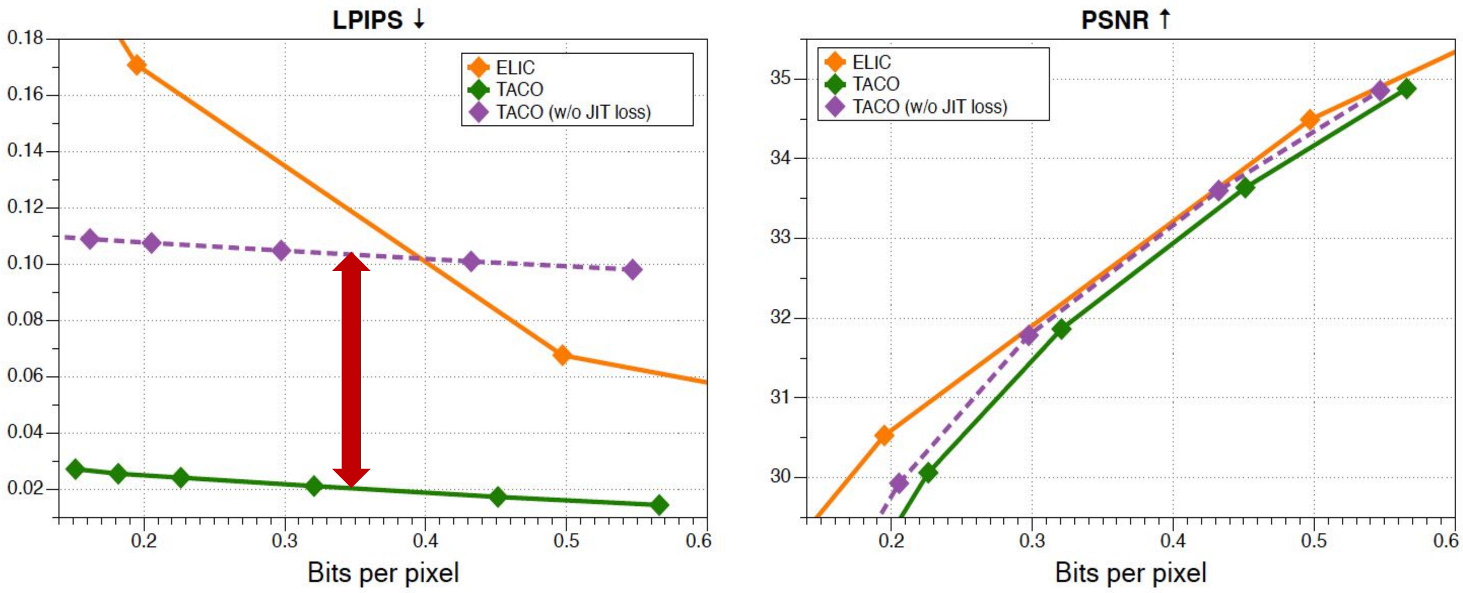
Ablation Studies

Without a text adapter, perceptual fidelity (LPIPS) is substantially degraded.



Ablation Studies

Without joint image text loss, perceptual fidelity (LPIPS) severely degrades.



Contribution

- Propose the ***first text-for-encoding-only framework***
- Achieve high pixel-level fidelity as well as high perceptual quality
- Show the **importance of using text** to focus on perceptually relevant information in images