DLF: Extreme Image Compression with Dual-generative Latent Fusion

Naifu Xue[1], Zhaoyang jia[2], Jiahao Li[3], Bin Li[3], Yuan Zhang[1], Yan Lu[3]

[1]Communication University of China, [2]University of Science and Technology of China, [3]Microsoft Research Asia

ICCV HONOLULU HAWAII OCT 19-23, 2025

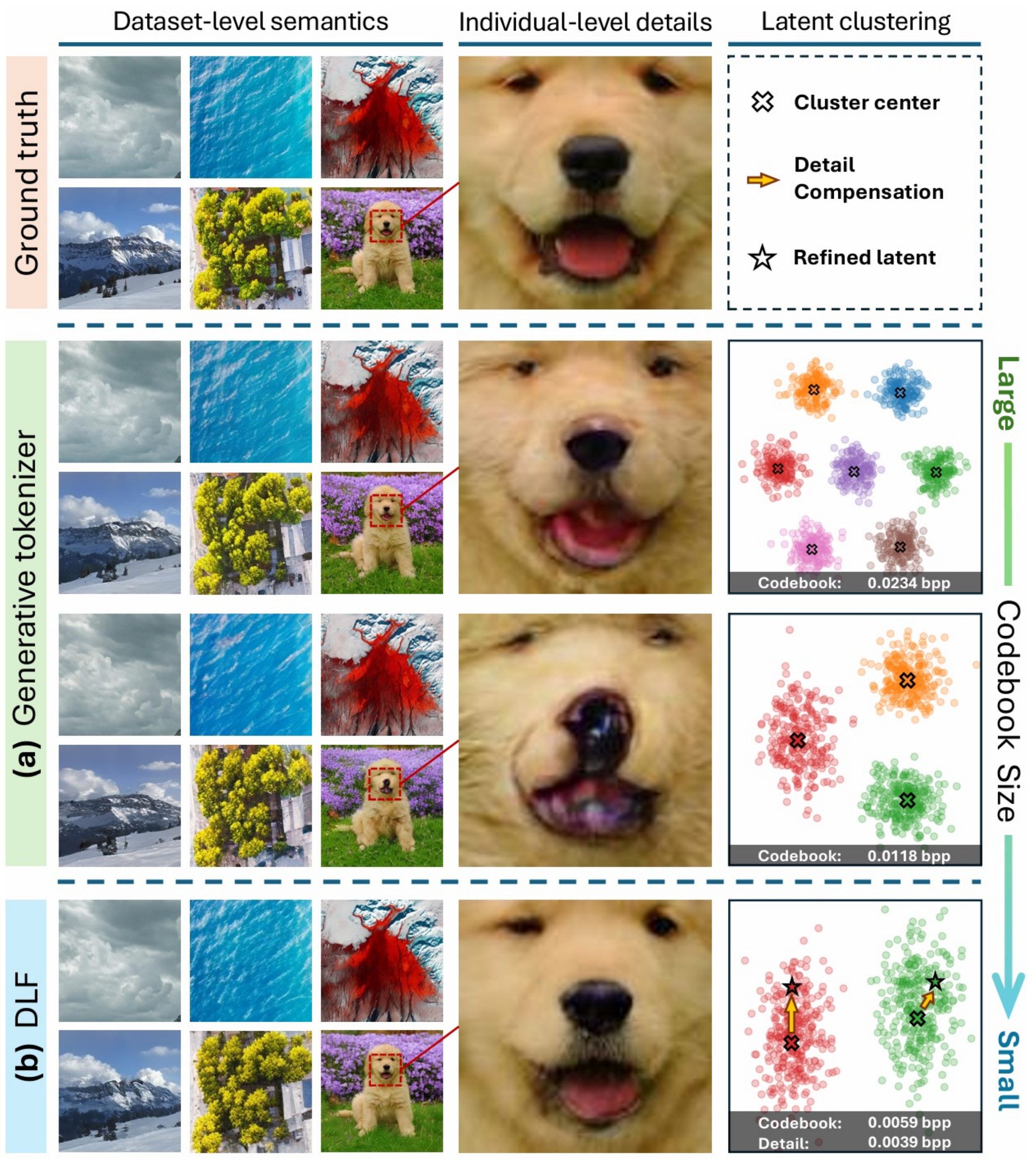
Introduction

Motivation: Increasing demand for efficient image compression methods.

- Traditional / MSE-optimized codecs produce blurry images at low bitrates.
- Recent generative codecs apply **visual tokenizer** for higher compression ratio but sacrifice detail fidelity at low bitrates.

Tokenizer Analysis

- Visual tokenizers compress images into latent features and generate reconstructions from compact representations.
- They prioritize clustering common semantics on the dataset.
- But failing at capturing fine-grained details on single image.
- Leads to **suboptimal fidelity** at extremely low bitrates!

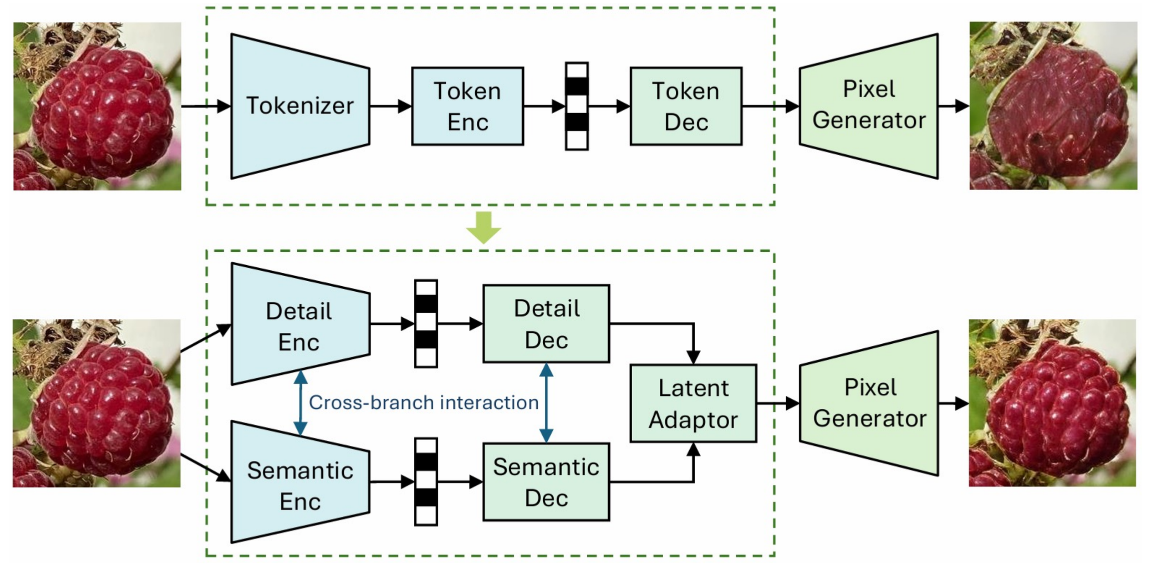


Our Solution: DLF

(**D**ual-generative **L**atent **F**usion)

- DLF decomposes the image into **semantic** and **detail** parts.
- **Semantic branch** inherits the clustering capability of generative visual tokenizer.
- **Detail branch** represents the diverse details through a large quantization space.
- **Cross-branch interaction** optimizes bit allocation between bottlenecks, there by reducing redundancy.

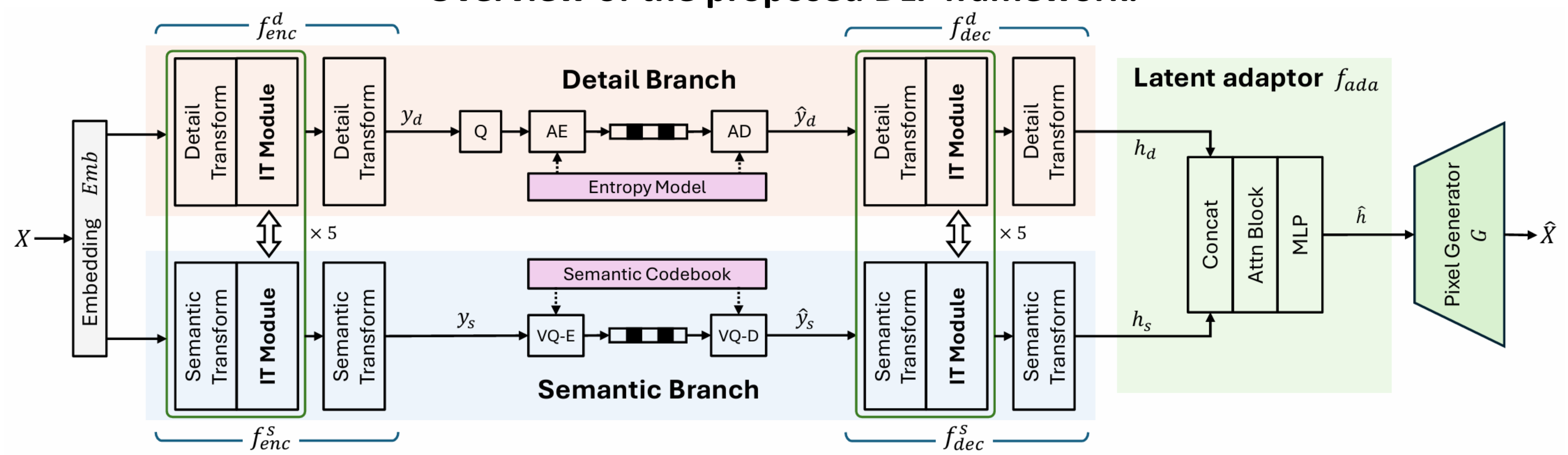
Existing generative tokenizer-based framework



Our **Dual**-generative Latent Fusion framework

Method

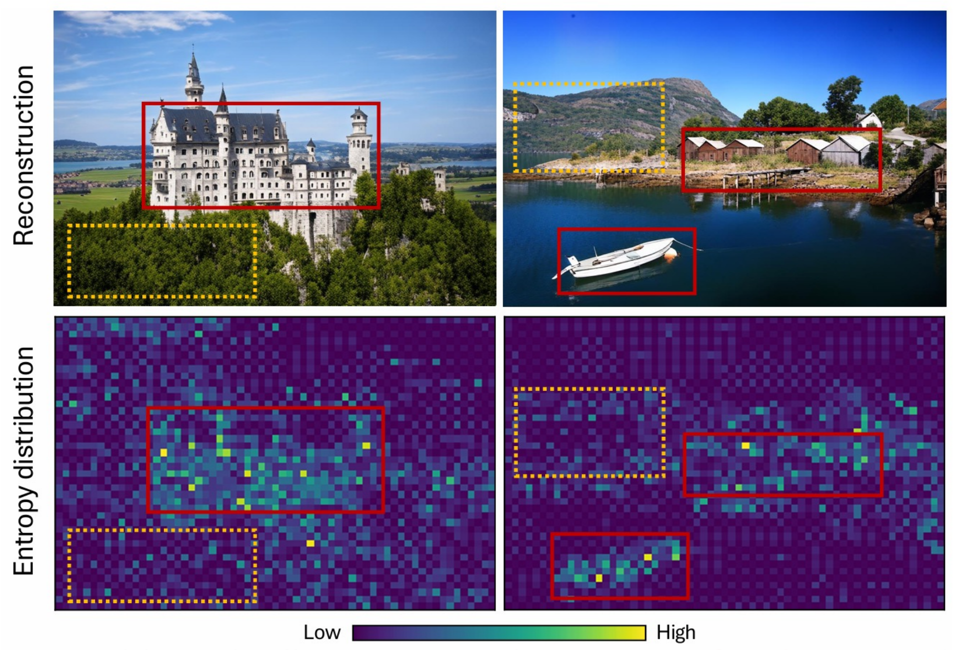
Overview of the proposed DLF framework.



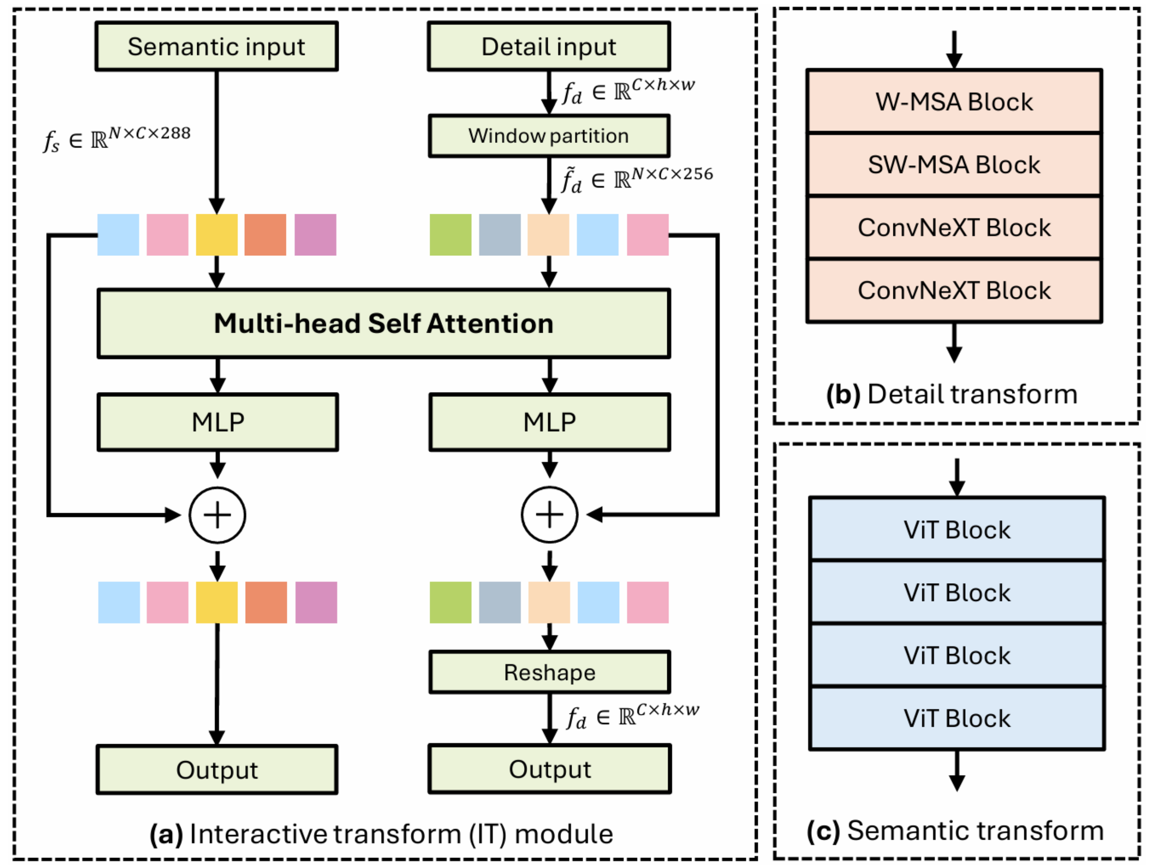
Q: scalar quantization; AE/AD: arithmetic codec; VQ-E/VQ-D: VQ-indices coding.

Key components

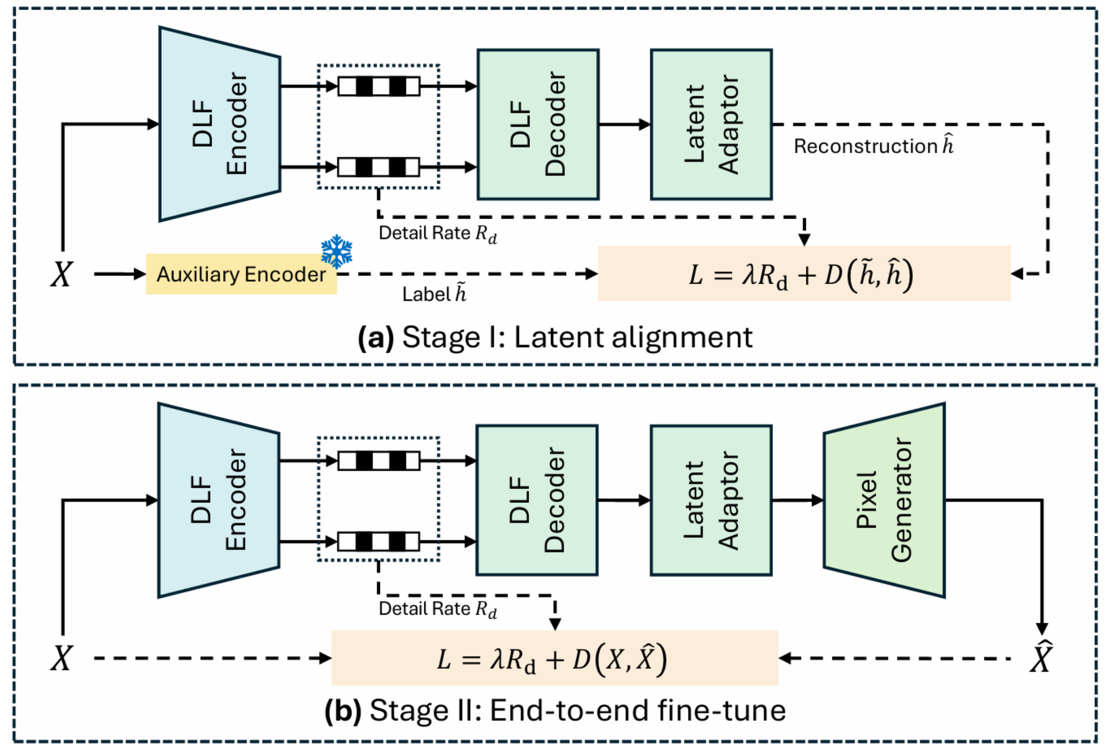
- Semantic transform adopts 1-D tokenizer to compress each 256×256 image patch into 32 tokens, while detail transform adopts Swin-Conv blocks to extract details.
- Two branches interact via multiple IT modules to reduce redundancy and improve compression efficiency.
- Semantic latents are quantized with vector quantization (VQ), while the detail branch uses scalar quantization (SQ) to represent finer content with a larger quantization space.
- Latent Adaptor fuses two latents into VQGAN's latent space. The image is decoded with a pretrained pixel generator.



Detail branch allocates more bits to specific objects and fewer bits to common content.



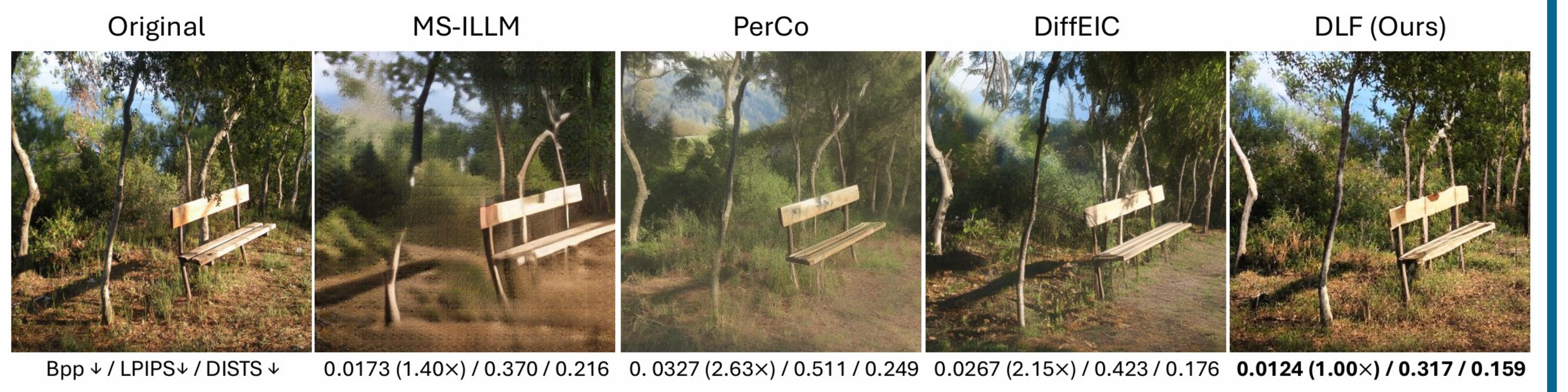
Architecture of Interactive Transform (IT) Module.



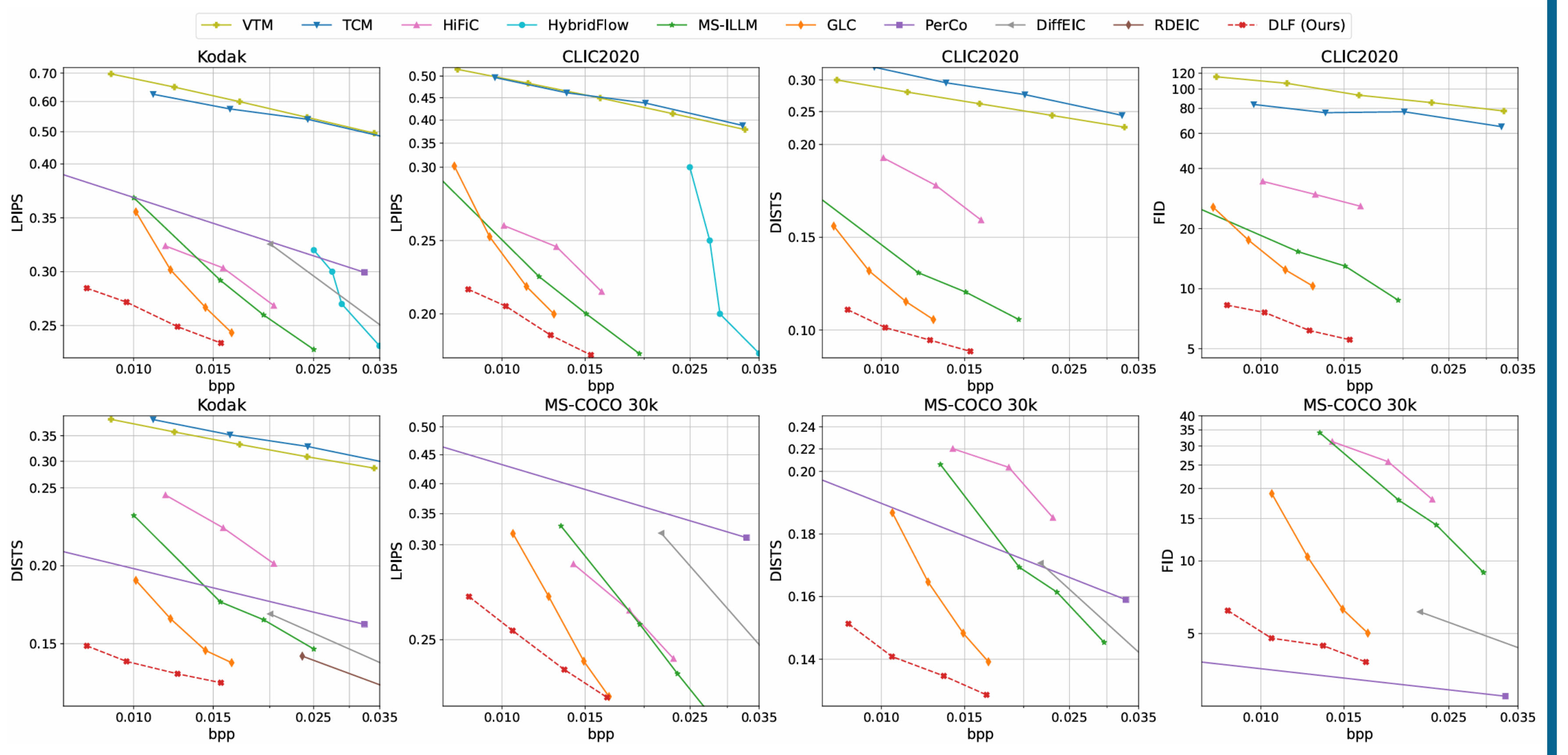
Two-stage training: **Stage I** aligns the latent to accelerate convergence; **Stage II** fine-tunes the entire model in the pixel domain, enhancing reconstruction quality.

Experiment

Qualitative Evaluation: DLF delivers best visual quality with lowest bitrate.



Quantitative Evaluation: DLF achieves SOTA rate-distortion performance.



Ablation studies validate the effectiveness of the dual-branch architecture, cross-branch interactive design, and SQ-based detail quantization.

Model variants	Kodak		CLIC2020	
	LPIPS	DISTS	LPIPS	DISTS
w/o detail	17.5%	20.2%	47.9%	47.6%
w/o interactive	64.1%	73.6%	68.8%	61.8%
w/ VQ detail	18.3%	40.7%	27.3%	58.1%
w/ SQ detail (**DLF**)	0.0%	0.0%	0.0%	0.0%

Complexity analysis shows DLF achieves faster coding and better quality than recent diffusion-based codecs (PerCo, DiffEIC. The larger model leads slower speed compared to MS-ILLM but ensures superior generation quality.

Model	Enc. Time (s)	Dec. Time (s)	BD-Rate
MS-ILLM [35]	0.064 ± 0.010	0.070 ± 0.011	0.00%
PerCo [7]	0.461 ± 0.017	2.443 ± 0.011	-4.02%
DiffEIC [25]	0.152 ± 0.014	4.093 ± 0.042	14.67%
DLF	0.178 ± 0.015	0.252 ± 0.014	-67.82%