Distributed Image Compression with Multimodal Side Information at Extremely Low Bitrates

Supplementary Material

6. Conditional Entropy Analysis with Multimodal Side Information

6.1. Theoretical Analysis

Theorem. *Let X be the main source to be compressed, and Y denotes the correlated source. The decoder receives a representation of Y in the form of multimodal features $V = (V_X, V_Y)$ and T, where V_X denotes the components of V that are correlated with X, and V_Y denotes components specific to Y, T represents the textual feature distribution. Then for the joint distribution of (X, V_X, V_Y, T), we have*

$$H(X \mid V_X, V_Y, T) \;\le\; H(X \mid V_X, V_Y). \tag{24}$$

Proof. Using the chain rule in two different orders, we expand the joint entropy $H(X, V_X, V_Y, T)$ as

$$\begin{aligned} &H(X, V_X, V_Y, T) \\ &= H(V_X, V_Y, T) + H(X \mid V_X, V_Y, T) \\ &= H(V_X, V_Y) + H(T \mid V_X, V_Y) + H(X \mid T, V_X, V_Y) \end{aligned} \tag{25}$$

$$\begin{aligned} &H(X, V_X, V_Y, T) \\ &= H(V_X, V_Y, X) + H(T \mid V_X, V_Y, X) \\ &= H(V_X, V_Y) + H(X \mid V_X, V_Y) + H(T \mid X, V_X, V_Y) \end{aligned} \tag{26}$$

Equating Eq. (25) and Eq. (26), yielding

$$\begin{aligned} &H(T \mid V_X, V_Y) + H(X \mid V_X, V_Y, T) \\ &= H(X \mid V_X, V_Y) + H(T \mid X, V_X, V_Y). \end{aligned} \tag{27}$$

Rearranging gives

$$\begin{aligned} &H(X \mid V_X, V_Y, T) \\ &= H(X \mid V_X, V_Y) - \big(H(T \mid V_X, V_Y) - H(T \mid X, V_X, V_Y)\big) \\ &= H(X \mid V_X, V_Y) - I(X; T \mid V_X, V_Y), \end{aligned} \tag{28}$$

where $I(X; T \mid V_X, V_Y) \ge 0$ is the conditional mutual information, thus, we have

$$H(X \mid V_X, V_Y, T) \le H(X \mid V_X, V_Y). \tag{29}$$

Remark 1. Theorem 6.1 and its proof establish that incorporating the additional modality T always reduces or maintains the conditional entropy of X given the decoder side information. The reduction is quantified by the conditional mutual information $I(X; T \mid V_X, V_Y)$, which measures how much information about X is provided by T beyond what is already contained in V.

The minimum achievable coding rate of X with side information (V, T) is

$$R^*(V, T) = H(X \mid V_X, V_Y, T) \tag{30}$$

From this expression, V_Y is conditionally independent of X given (V_X, T), which does not help reduce the conditional entropy, thus

$$H(X \mid V_X, V_Y, T) = H(X \mid V_X, T). \tag{31}$$

The inequality in Eq. 29 is strict if and only if $I(X; T \mid V_X, V_Y) > 0$. This occurs when T contains components of X not already captured by V_X, or captures complementary semantic information relevant to the reconstruction of X. Components in V_Y that are specific to Y and independent of X do not contribute to reducing the conditional entropy. Hence, the compression gain primarily comes from T and V_X that correlate with X.

Remark 2. Following Remark 1, identifying the minimum achievable rates as

$$R^*(V) = H(X \mid V_X), \quad R^*(V, T) = H(X \mid V_X, T). \tag{32}$$

The expected reduction in rate due to the additional modality T is

$$\begin{aligned} \Delta R = R^*(V) - R^*(V, T) = I(X; T \mid V_X) \ge 0, \\ \Delta R = 0 \iff I(X; T \mid V) = 0 \iff X \perp T \mid V. \end{aligned} \tag{33}$$

In practice, this implies that multimodal side information can be explicitly leveraged to design more efficient distributed compression schemes. By maximizing $I(X; T \mid V)$ during feature design, one can achieve a significant reduction in the conditional entropy and, consequently, in the required bit rate for encoding X.

Extension to Lossy Compression. The previous analysis assumes lossless compression. In the lossy setting with a distortion constraint D, the minimum achievable coding rate is given by the conditional rate-distortion function:

$$R(D \mid V, T) = \min_{p_{\hat{X} \mid X, V, T}} I(X; \hat{X} \mid V, T) \;\text{ s.t. } \mathbb{E}[d(X, \hat{X})] \le D \tag{34}$$

where $d(X, \hat{X})$ denotes a distortion measure.

Analogously to the lossless case, adding the modality T never increases the minimum achievable rate:

$$R(D \mid V, T) \le R(D \mid V). \tag{35}$$

The rate reduction under distortion D is

$$\Delta R(D) = R(D \mid V) - R(D \mid V, T) \geq 0. \quad (36)$$

Similar to the lossless case, $\Delta R(D) = 0$ if and only if $X \perp T \mid V$, i.e., T provides no additional information about X given V.

In practice, designing T to maximize its information contribution relative to V helps reduce the rate under a given distortion, enabling more efficient lossy distributed compression.

6.2. Relative Experiments

In the proposed MDIC, T is constructed to capture shared multi-view semantic information that complements the visual features V. Moreover, T is utilized in the VMGM-TS to supervise the visual feature–mask generator.

Table 3. Ablation study of the Text Supervision (TS) in VMGM-TS, and textual side information z_{text}. BD-Rate measures the average bitrate difference between two rate–distortion (RD) curves over a common quality range, quantifying how much more or less bitrate a method requires, on average, to achieve the same quality level as the baseline. Lower is better (↓).

Method	BD-Rate ↓			
	LPIPS	DISTS	PSNR	MS-SSIM
MDIC w/o TS	121.49%	72.33%	97.20%	50.78%
MDIC w/o z_{text}	104.79%	82.91%	35.67%	37.61%
MDIC (Ours)	**0**	**0**	**0**	**0**

To demonstrate the impact of textual side information z_{text} on the final reconstruction results, we conduct an ablation study on both z_{text} and Text Supervision (TS) in VMGM-TS. As shown in Table 3, taking our MDIC as the baseline, removing z_{text} results in higher bitrate consumption to achieve the same reconstruction quality. In other words, both perceptual metrics (LPIPS, DISTS) and distortion metrics (PSNR, MS-SSIM) degrade without z_{text}.

To visually illustrate the impact of textual side information on reconstruction quality, we further present local reconstruction comparisons, as illustrated in Fig. 10. In terms of fine-grained color reconstruction, the absence of textual side information tends to mix colors across adjacent regions, leading to inaccurate and less faithful color details. As shown in the Fig. 10, the boundary of the lane marking is contaminated by the shadow's dark color, and the lack of certain color-transition cues results in local blurring around the window area. After incorporating textual side information, the diffusion model is guided to align its generative distribution with the specified global and local semantics, enabling it to better preserve structural details and produce more semantically accurate content.

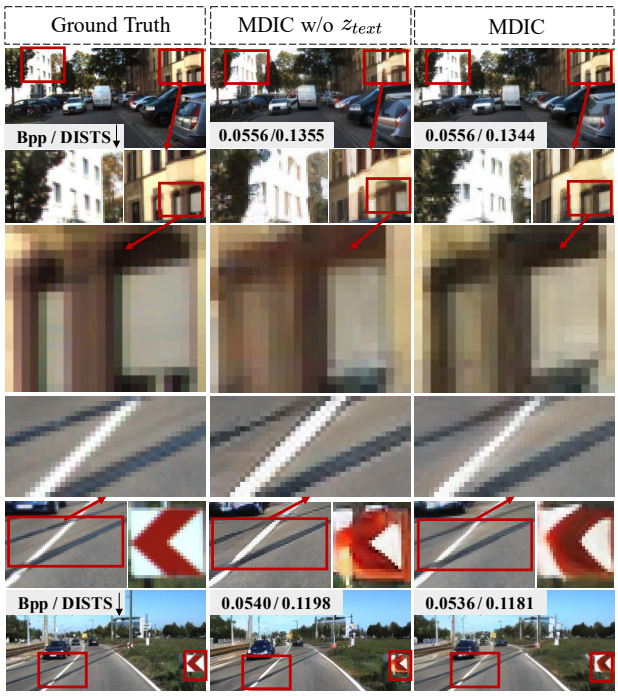


Figure 10. The visual comparison between MDIC and its variant without the textual side information z_{text}.

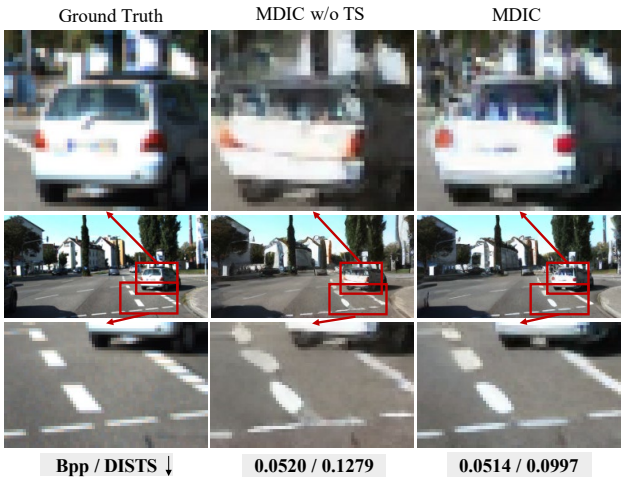
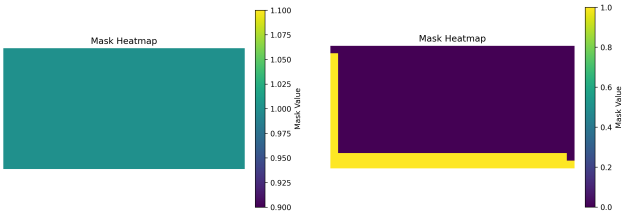


Figure 11. The visual comparison between MDIC and its variant without the Text Supervision (TS) auxiliary task.

Meanwhile, the textual side information also participates in the multimodal alignment auxiliary task, where it supervises the mask generator to focus on fine-grained details during optimization. As shown in Table 3, removing the Textual Supervision (TS) severely degrades both perceptual and distortion metrics, with the perceptual metrics being particularly affected. Under the same LPIPS score, the absence of textual supervision increases the required bitrate



(a) 0.0031 bpp (b) 0.0554 bpp

Figure 12. Visualization of the visual feature mask under different bitrates without TS.

by 121.49%.

To provide a more intuitive comparison of the reconstruction quality, we further present corresponding visualizations, as illustrated in Fig. 11. When the TS is removed, the reconstruction quality clearly deteriorates in fine-grained regions, exhibiting issues such as local blurring and boundary distortion. Moreover, as shown in Fig. 12, removing TS leads the generated masks to distribute their attention almost uniformly across the entire side-information map, rather than concentrating on specific regions. This indicates that a simple mask generator, without additional supervision, struggles to selectively filter local features, resulting in substantial redundant information and multi-view inconsistencies that ultimately degrade the reconstruction quality.

7. Additional Results

7.1. Denoising Steps

As illustrated in Fig. 13, we evaluate the proposed method under different denoising steps using both perceptual and distortion metrics. The distortion metrics achieve their best performance when the denoising step is set to 5. As the number of steps increases to 10, the model attains the highest perceptual quality while maintaining nearly optimal MS-SSIM. However, when the denoising steps exceed 10, the overall reconstruction quality gradually declines. Therefore, all experiments in this paper adopt 10 denoising steps by default.

7.2. Mask Visualization

We select the 14 most frequent object-level textual tokens from the KITTI Stereo and Cityscapes datasets to construct textual masks, and their frequency distributions are presented in Fig. 14.

We visualize the corresponding visual masks in Fig. 15. With supervision from our object-level multimodal alignment task, the model learns to generate the visual mask conditioned on each input image, assigning different importance weights to different objects or regions. This helps address a key limitation in existing cross-attention–based

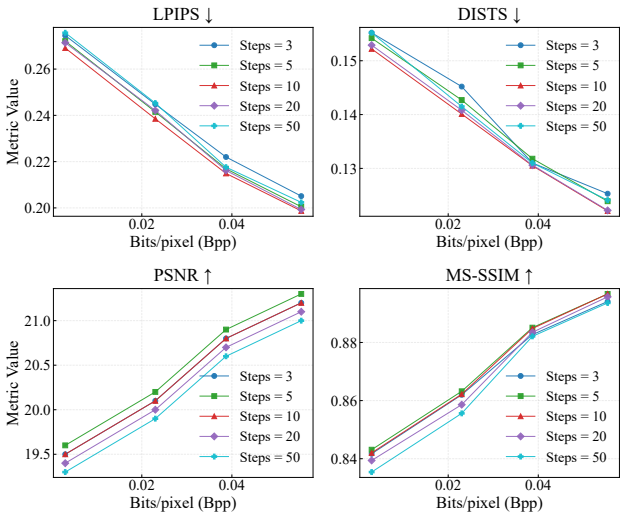


Figure 13. Evolution of LPIPS, DISTS, PSNR, and MS-SSIM across denoising steps as the number of different bitrates increases. Symbol conventions: ↑=higher better, ↓=lower better.

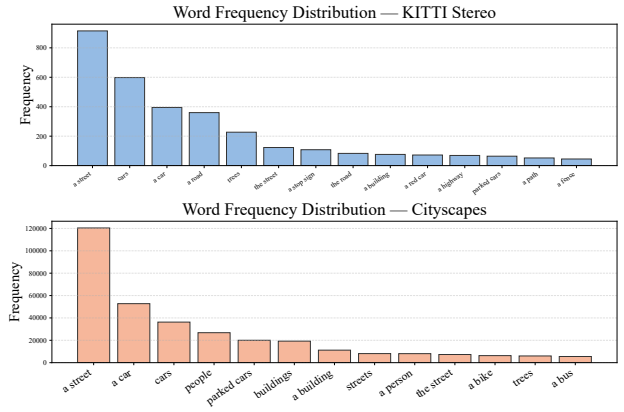


Figure 14. The high-frequency keywords on the KITTI Stereo and Cityscapes datasets.

methods, which struggle to effectively model global region-wise importance under extremely low bitrates.

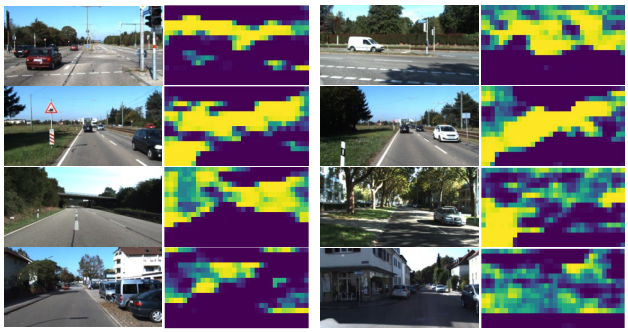


Figure 15. Visualization of the visual feature mask m_v.

Table 4. The comparison between our method and other diffusion-based approaches under different BD-Quality, as well as their encoding and decoding speeds. The BD-Quality measures the relative improvement or degradation of a given metric compared to a baseline at the same bitrate. DS represents the denoising steps.

Method	DS	BD-Quality ↑				Inference Time (s)	
		LPIPS	DISTS	PSNR	MS-SSIM	Encoding Time	Decoding Time
Perco [3]	5	-0.2501	-0.1147	-5.7031	-5.8163	0.020 ± 0.001	0.320 ± 0.002
	20	-0.2364	-0.1026	-5.8774	-5.8308	0.020 ± 0.001	0.843 ± 0.002
DiffEIC [14]	20	-0.2675	-0.1000	-5.7968	-5.9150	$0.114 + 0.005$	1.432 ± 0.016
	50	-0.2470	-0.0931	-5.9147	-5.9622	$0.114 + 0.005$	3.4371 ± 0.025
RDEIC [15]	2	-0.2295	-0.1431	-8.4241	-4.4578	0.082 ± 0.001	0.121 ± 0.001
	5	-0.2300	-0.1417	-8.7294	-4.5387	0.082 ± 0.001	0.295 ± 0.002
MDIC (Ours)	5	-0.0040	-0.0040	0.1144	0.0397	0.012 ± 0.001	0.461 ± 0.002
	10	0	0	0	0	0.012 ± 0.001	0.638 ± 0.006

As illustrated in Fig. 15, even among masked tokens, their importance is not identical. For example, the token "tree" receives consistently higher attention compared with other masked tokens such as "building" or "car". This is because trees typically contain richer structural and fine-grained texture information. Therefore, allocating more bit-budget to encode them is beneficial for reconstructing boundaries and color cues.

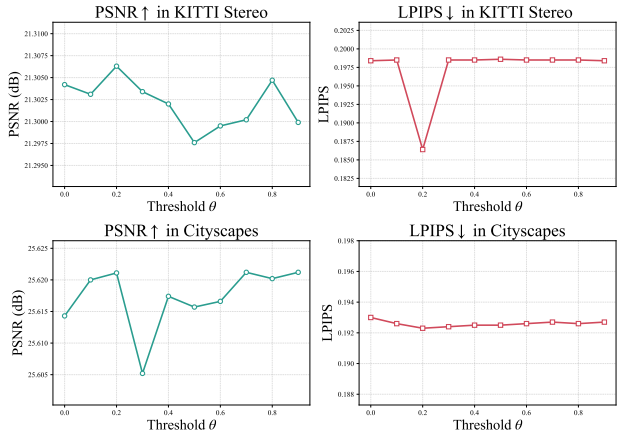


Figure 16. Variation of PSNR and LPIPS with threshold θ on the KITTI Stereo and Cityscapes datasets.

To simultaneously select regions of varying importance and sparsify the continuous mask, we evaluate LPIPS and PSNR metrics under different thresholds θ. For each threshold, metrics are averaged across all bitrates. As shown in Fig. 16, the best performance in both PSNR and LPIPS is achieved at a threshold of θ=0.2.

7.3. Generalizability and Robustness

As shown in Fig. 17 (a), applying our KITTI-trained model directly to the unseen Cityscapes dataset yields high-quality reconstructions comparable to Fig. 8 (main paper), verifying generalizability and robustness to domain shifts without fine-tuning. The reason is that the vocabulary acts strictly as training-time "Semantic Anchors" to supervise visual-semantic alignment, imposing no inference constraints. During inference, the mask generator relies solely on visual features to predict the mask without any text input. Therefore, it generalizes to unseen objects via learned structural features, akin to class-agnostic segmentation. As long as the vocabulary of size N covers the test-time semantic space, the model generalizes without imposing any class-specific constraints during inference.

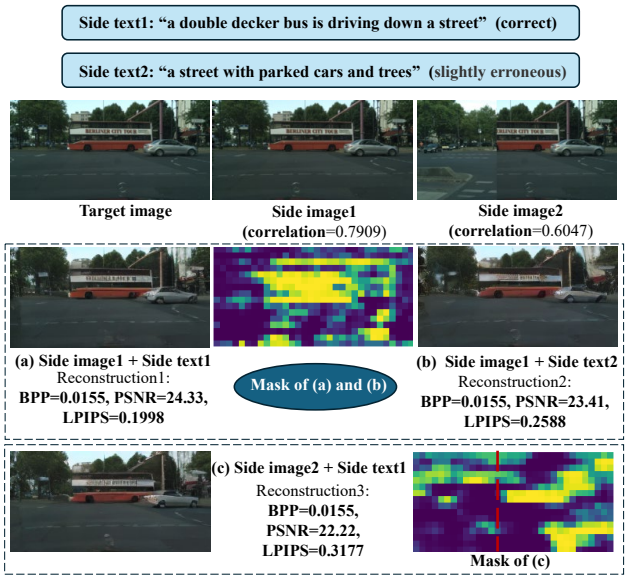


Figure 17. Applying the model trained on KITTI Stereo to the Cityscapes dataset (unseen domain) without any fine-tuning.

Furthermore, we use two side images with different degrees of correlation (measured by SSIM). In Fig. 17 (b) and (c), despite the expected quality degradation caused by incorrect captions or imperfect side information alignment,

the model avoids collapse and continues to produce plausible images without severe artifacts. The mask successfully suppressed irrelevant regions (wrong road layout) and retained partial attention to the trees in the irrelevant region (sharing similar texture features) due to the latent similarity calculation and attention mechanism.

7.4. Complexity Comparison

As shown in Table 4, we compare our MDIC with representative diffusion-based image compression methods under different denoising steps in terms of Rate–Distortion-Perception (R-D-P) performance and inference speed.

For a fair comparison, we adopt the results of each method under its optimal denoising step as reported in the original papers. We then take our 10-step MDIC model as the baseline and compute the average metrics to measure relative performance variation at the same bitrate. Specifically, the BD-Quality is calculated by integrating the difference between the RD curves or RP curves of two methods over the overlapping bitrate range, where a positive value indicates improvement and a negative value denotes degradation compared with our baseline. The overall calculation process is as follows:

$$BD = \frac{1}{r_2 - r_1} \int_{r_1}^{r_2} [D_A(r) - D_B(r)]\, dr, \qquad (37)$$

where $D_A(r)$ and $D_B(r)$ denote the distortion or perception at bitrate r for methods A and B, respectively, and $[r_1, r_2]$ represents their common bitrate range.

In implementation, we fit the R-D or R-P points of each method using cubic interpolation and compute the definite integral of the curve differences over the overlapping bitrate range. This yields a unified and quantitative comparison across both perceptual and distortion metrics.

Diffusion-based LIC methods generally exhibit higher decoding complexity than VAE-based approaches due to their reliance on Stable Diffusion backbones. Although MDIC introduces an additional side-information interaction module compared to other diffusion-based LICs, it increases the inference time by only about 0.1s relative to Perco, while achieving notably better R-D-P performance at comparable bitrates.

7.5. Qualitative Comparison

As shown in Fig. 18, we visualize the reconstruction results of all compared methods under approximately the same compression rate. Existing distortion-oriented DIC methods (NDIC [20], ATN [21], LDMIC [40]) achieve high pixel-level fidelity and maintain good semantic consistency with the original images. However, they still suffer from noticeable blurring and artifacts in both local and global regions. Joint encoding methods (SASIC [33], ECSIC [34], BiSIC [17]) enable additional interaction among multi-view images during the encoding stage, thereby improving the overall perceptual quality compared to conventional DIC approaches. Nonetheless, due to the extremely low bitrate and the limited amount of transmitted information, these methods lose significant fine-grained details, leading to local blurriness and texture degradation. Perception-oriented LIC methods (Perco [3], DiffEIC [14], RDEIC [15]) demonstrate better perceptual results in both global and local aspects. However, the absence of semantic consistency constraints and the lack of effective side-information utilization often result in semantically inconsistent details and distorted edge structures that deviate from the original image content. In contrast, the proposed MDIC achieves the best global and local perceptual quality. Even under extremely low bitrates, it maintains high perceptual fidelity and fine-grained semantic consistency with original image, clearly demonstrating the superiority of MDIC in effectively leveraging side information for reconstruction.

Figure 18. Visualization results on KITTI Stereo and Cityscapes datasets. Each row corresponds to a different method. Numbers below the images indicate (BPP, PSNR, LPIPS). Symbol conventions: ↑=higher better, ↓=lower better.