# Diffusion Bridge: Leveraging Diffusion Model to Reduce the Modality Gap Between Text and Vision for Zero-Shot Image Captioning

## Supplementary Material

| Method | MSCOCO → Flickr30K | | | | Flickr30K → MSCOCO | | | |
|---|---|---|---|---|---|---|---|---|
| | B@4 | M | C | S | B@4 | M | C | S |
| DeCap [2] | 16.3 | 17.9 | 35.7 | 11.1 | 12.1 | 18.0 | 44.4 | 10.9 |
| CapDec [3] | 17.3 | **18.6** | 35.7 | - | 9.2 | 16.3 | 27.3 | - |
| MeaCap $_{InvLM}$ [4] | 18.5 | 19.5 | 43.9 | **12.8** | 13.1 | 19.7 | 56.4 | 13.2 |
| ViECap [1] | 17.4 | 18.0 | 38.4 | 11.2 | 12.6 | 18.3 | 54.2 | 12.5 |
| **Ours w/ ViECap** | **17.5** (+0.1) | **18.6** (+0.6) | **42.1** (+3.7) | 11.8 (+0.6) | **15.0** (+2.4) | **20.0** (+1.7) | **60.1** (+5.8) | **14.0** (+1.5) |

Table 1. Cross-domain captioning performance on the MSCOCO test and Flickr30K test set.

CapDec: A cat looking out a window at a bird.

C3: A cat looking out a window at a cat sitting on a bench.

Ours: A cat sitting on a window sill looking out.

CapDec: A man is riding a motorcycle on the street.

C3: A man in a green jacket is riding a motorcycle.

Ours: A police officer on a motorcycle with a crowd of people.

Figure 1. Qualitative comparison of generated captions.

## 1. Plug-and-Play

Early-Guidance Decoding method [2, 3, 5] and Keywords-Guidance Decoding methods [1, 4] represent parallel research directions that tackle zero-shot captioning challenges from distinct perspectives. These two approaches are inherently complementary, allowing for plug-and-play applications where the strengths of each method can be combined for enhanced performance. To demonstrate this capability, we integrate Diffusion Bridge with ViECap [1], a Keywords-Guidance Decoding method that utilizes entity extraction and hard prompts to improve caption, particularly in cross-domain scenarios.

Tab. 1 presents the results of combining Diffusion Bridge with ViECap for cross-domain captioning tasks, specifically MSCOCO → Flickr30K and Flickr30K → MSCOCO. In the MSCOCO → Flickr30K setup, Diffusion Bridge with ViECap improves performance in every metrics, indicating its ability to enhance caption quality when transitioning from a large-scale dataset to a smaller one. Similarly, in the Flickr30K → MSCOCO setup, this combination achieves the best performance across most metrics that reflect linguistic fluency and semantic relevance.

This experiment highlights the plug-and-play nature of Diffusion Bridge, showcasing its compatibility with complementary techniques like Keywords-Guidance Decoding. While Diffusion Bridge effectively reduces the modality gap at the embedding level, ViECap's keyword-driven guidance provides additional contextual specificity and robustness, especially in challenging cross-domain scenarios. This synergy underscores the adaptability of our approach and its potential to enhance performance when integrated with other decoding strategies.

## 2. Justification of the scaling factor.

Our scaling factor was selected to ensure signal dominance over noise in early diffusion steps. Scaling embeddings by $C = 5$ amplifies the signal, aligning with practices like Stable Diffusion ($C = 1/0.18215$). To assess the impact of the scaling factor, we trained models with different scales. The results showed: $C = 1$ (cross-modal cosine similarity = 0.106) fails to cross-modal align; $C = 3$ (0.581) and $C = 10$ (0.541) achieve robust alignment. $C = 5$ achieves the highest similarity (0.606), resulting in optimal performance.

## 3. Computational cost of the diffusion model.

Our analysis reveals that Diffusion Bridge incurs an additional 1.63 GFLOPs and 0.27 seconds during inference compared to the baseline method without a diffusion model, which requires 8.46 GFLOPs and 1.02 seconds. The computational cost remains manageable through the utilization of DDIM sampling with only 12 steps. Although Diffusion Bridge introduces additional computation overhead, this trade-off remains reasonable given the substantial improvements observed in zero-shot captioning performance.

## 4. Dataset-size oriented analysis.

Our experiments show that using 10% of COCO captions for training diffusion model achieves a CIDEr score of 91.8, while 50% of COCO captions results in 93.5, demonstrating that Diffusion Bridge remains effective even with limited data. In Flickr30K, excluding DeCap, our method achieves the highest performance, as shown in Tab. 2 (b). The superior performance of DeCap in Flickr30K is likely due to its memory-based projection, which utilizes the entire training dataset as a memory bank. This approach can be particularly effective for small datasets, as it directly memorizes dataset-specific projections. However, as dataset size increases, our method, which considers the intrinsic embedding distribution, generalizes better and achieves stronger performance.

## References


[1] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3146, 2023. 1

[2] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023. 1

[3] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 1

[4] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14100–14110, 2024. 1

[5] Yuhui Zhang, Elaine Sui, and Serena Yeung-Levy. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. *arXiv preprint arXiv:2401.08567*, 2024. 1