# D-Attn: Decomposed Attention for Large Vision-and-Language Model

Supplementary Material


## Abstract

*The supplementary material is organized into the following sections:*

- *Section A: Implementation Details.*
- *Section B: Embedding Distribution.*
- *Section C: Comparison on Video Understanding.*


## A. Implementation Details

In Table 1, we list key hyper-parameters for all three training stages and two LLMs, Mistral 0.3 7B and Gemma 2 9B. We use the same set of hyper-parameters for D-Attn models and their S-Attn counterparts. The weight decay and AdamW-related parameters are taken from LLaMA 2 [6] technical report.

| | **Stage 1** | **Stage 2** | | **Stage 3** | |
|---|---|---|---|---|---|
| | | Mistral 0.3 7B | Gemma 2 9B | Mistral 0.3 7B | Gemma 2 9B |
| lr adapter | 1e-3 | 5e-6 | 2e-5 | 5e-6 | 2e-5 |
| lr llm | 0.0 | 2e-6 | 1e-5 | 2e-6 | 1e-5 |
| lr vis-enc | 0.0 | 2e-7 | 1e-6 | 2e-7 | 1e-6 |
| weight decay | 0.0 | 0.1 | | 0.1 | |
| optimizer | AdamW | AdamW | | AdamW | |
| Adam $\beta_1$ | default (0.9) | 0.9 | | 0.9 | |
| Adam $\beta_2$ | default (0.999) | 0.95 | | 0.95 | |
| Adam $\epsilon$ | default (1e-8) | 1e-5 | | 1e-5 | |
| warmup ratio | 0.03 | 0.03 | | 0.03 | |
| lr scheduler | cosine | cosine | | cosine | |
| epochs | 1 | 1 | | 1 | |
| total batch size | 512 | 256 | | 128 | |
| dtype | bfloat16 | bfloat16 | | bfloat16 | |
| deepspeed | stage 2 | stage 3 | | stage 3 | |

Table 1. Hyperparameters for three training stages and two types LLMs.

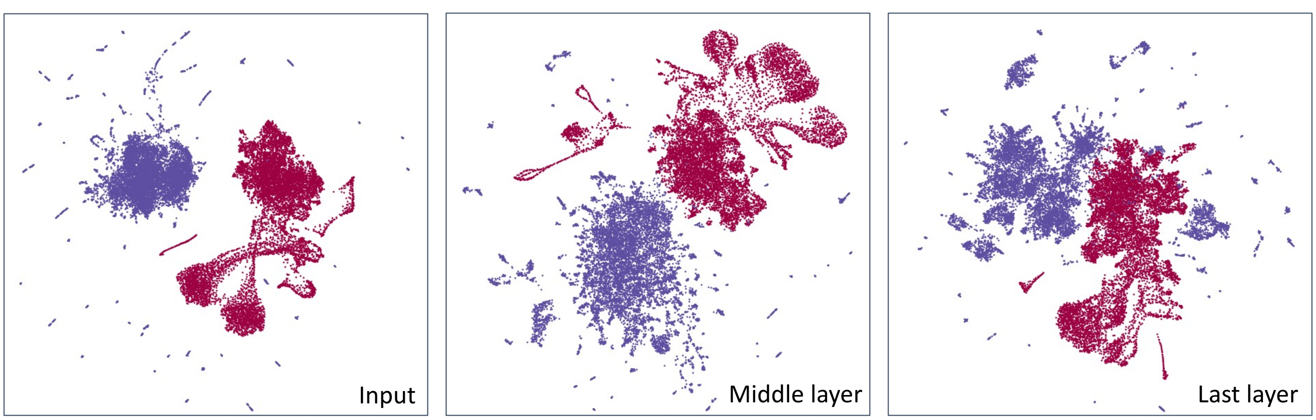


Figure 1. The UMAP visualization of visual and textual tokens.

Table 2. Main results on VideoChatGPT [3] benchmark. DO, CU, TU, CO denote correctness of information, detail orientation, context understanding, temporal understanding, and consistency.

| Method | LLM | Max #Frames | CI | DO | CU | TU | CO |
|---|---|---|---|---|---|---|---|
| S-Attn-LLaMA-Vid | Vicuna 1.5 7B | 140 | 2.2 | 2.2 | 2.7 | 2.1 | 2.3 |
| D-Attn-LLaMA-Vid | Vicuna 1.5 7B | 1156 | 2.4 | 2.2 | 2.8 | 2.0 | 2.6 |

## B. Embedding Distribution

Most recent LVLMs attempt to align the visual embeddings from vision encoders (*e.g.*, CLIP [5] or SigLIP [7]) to the text tokens of LLM by an adapter, which is usually an MLP block. In Figure 1, we visualize the feature distribution of visual and textual embeddings in LLaVA-1.6 [2] model using UMAP [4]. We find that even on the training dataset of LLaVA-1.6 [2] model, the distribution of textual and visual embeddings are not fully aligned at input, middle, and last LLM layers. This observation supports our motivation that visual and textual tokens in LVLMs are inherently different, and a specifically design attention architecture for visual tokens could potentially lead to better performance and efficieny than the tranditional self-attention.

## C. Comparison on Video Understanding

To further verify the effectiveness of D-Attn on more input visual tokens, we train a VideoQA model with the training recipe of LLaMA-VID [1], where the input is multiple frames (images) and the model has to model the contextual information across frames to answer the question. The architecture remains the same as our image model, except we encode each frame as 64 tokens and concatenate them as visual inputs. In Table 2, we again observe no performance degradation by incorporating V2V Diagonal-Attn. The D-Attn actually outperforms the S-Attn counterpart on CI (Correctness of Information) and CO (Consistency). Although the V2V attention could be important for video understanding, our explanation is that visual embeddings could exchange information indirectly via textual embeddings. Concretely, each textual embedding gathers visual information via T2V Cross-Attn at the L-th decoder layer. Therefore, at the (L+1)-th layer, visual embeddings indirectly exchange information via text embeddings in T2V Cross-Attn. Moreover, D-Attn can take much greater number of frames ($140 \rightarrow 1156$) than S-Attn, which is critical for long videos. These advantages make D-Attn suitable for video understanding tasks.

## References


[1] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2

[2] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[3] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2

[4] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 2

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[7] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 2