Browse and Concentrate: Comprehending Multimodal Content via prior-LLM Context Fusion

Ziyue Wang[*,1], Chi Chen[*,1], Yiqi Zhu[1], Fuwen Luo[1], Peng Li[✉2,4], Ming Yan[3], Ji Zhang[3], Fei Huang[✉3], Maosong Sun[1], Yang Liu[1,2,4,5]

[1]1. Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[2]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[3]Institute of Intelligent Computing, Alibaba Group
[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[5]Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China

Abstract

With the bloom of Large Language Models (LLMs), Multimodal Large Language Models (MLLMs) that incorporate LLMs with pre-trained vision models have recently demonstrated impressive performance across diverse vision-language tasks. However, they fall short to comprehend context involving multiple images. A primary reason for this shortcoming is that the visual features for each images are encoded individually by frozen encoders before feeding into the LLM backbone, lacking awareness of other images and the multimodal instructions. We term this issue as prior-LLM modality isolation and propose a two phase paradigm, browse-and-concentrate[1], to enable in-depth multimodal context fusion prior to feeding the features into LLMs. This paradigm initially "browses" through the inputs for essential insights, and then revisits the inputs to "concentrate" on crucial details, guided by these insights, to achieve a more comprehensive understanding of the multimodal inputs. Additionally, we develop training strategies specifically to enhance the understanding of multi-image inputs. Our method markedly boosts the performance on 7 multi-image scenarios, contributing to increments on average accuracy by 2.13% and 7.60% against strong MLLMs baselines with 3B and 11B LLMs, respectively.

1 Introduction

Multimodal Large Language Models (MLLMs) have recently garnered attention for their surging popularity and impressive performance across diverse Vision-Language (VL) tasks (Team et al., 2023; OpenAI, 2023; Qi et al., 2023). Among these MLLMs, the paradigm that extending Large Language Models (LLMs) with pre-trained vision encoders has shown remarkable abilities in visual reasoning and visual instruction-following (Wu et al., 2023; Yin et al., 2023). These models also draw attention for their feasibility and flexibility in adapting to varied scenarios and demands (Liu et al., 2023b; Zhu et al., 2023).

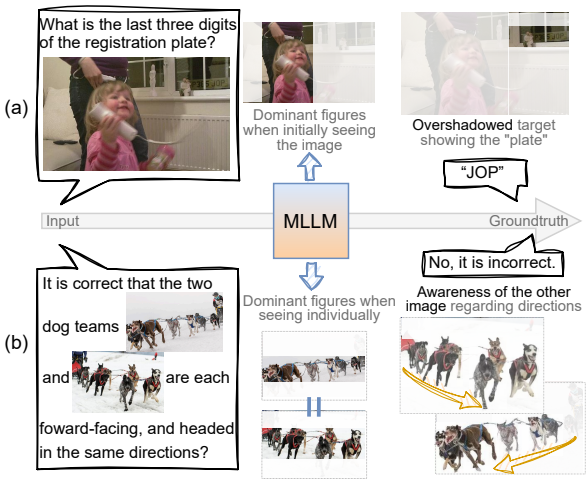

Figure 1: Examples of the modality isolation issue. (a) illustrates image-text isolation, where the child figure dominates the image while the "registration plate", which should have been focused on, is overshadowed. (b) illustrates inter-image isolation, where the two images lack information regarding "directions" of each other. Both situations undergo absence of awareness regarding the global multimodal context.

Despite its impressive abilities, this paradigm faces challenges that obscure a deeper understanding of multi-image and interleaved inputs (Dai et al., 2023; Luo et al., 2023; Zhao et al., 2024; Li et al., 2023d). The approach of simply gluing up pre-trained vision and language models via intermediate components (Li et al., 2023e; Liu et al., 2023d) potentially neglects essential cross-modality and inter-image interactions, leading to the LLM being presented with isolate visual and textual features without recognition of interleaved multimodal inputs. We refer to this problem as *prior-LLM modality isolation*, which further divides two issues *image-text isolation* and *inter-image isolation*. These challenges have received

[*]These authors contribute equally.
[✉]Corresponding authors: Peng Li and Fei Huang.
[1]Code is released at https://github.com/THUNLP-MT/Brote

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11229–11245

considerable attention but remain unresolved.

Firstly, *image-text isolation* happens when frozen vision encoders produce generic visual features, overlooking crucial target-specific information. For instance, in Figure 1 (a), the emphasis should be on the "registration plate". This plate, occupying only a minor area of the image, is prone to being overshadowed by predominant elements due to inadequate image-text interaction. To tackle this problem, Dai et al. (2023) and Luo et al. (2023) integrate textual instructions into visual feature extraction to enhance the responsiveness of these features to the given instructions. Moreover, some researchers propose to alter the internal structure of LLMs to bridge the gap between visual and linguistic spaces (Wang et al., 2023). While these methods are effective in single-image scenarios, they do not address the concurrent fusion of multiple images.

Secondly, *inter-image isolation* arises from encoding images separately, disrupting semantic links among images and conveying misinformation of the multi-image context. This issue is particularly prevalent in scenarios involving interleaved and multiple images. As illustrated in Figure 1 (b), the moving direction regarding the other image should be considered. However, such relational information remains isolated and fails to transmit across images. Consequently, the lack of awareness regarding relevant content from other images can lead to the exclusion of essential visual information. To handle this issue, recent studies have developed context schemes that aim to improve image-text correlations and the connections between multiple images (Zhao et al., 2024; Li et al., 2023b). Nevertheless, the prior-LLM fusion of multimodal context are overlooked.

To mitigate the two outlined issues, we utilize a cognitive strategy that mirrors the process through which humans typically understand new content: by first grasping the main ideas during an initial browsing and then revisiting the material to deepen their understanding with the browsing insights (Garner, 1987). Inspired by this approach, we propose a novel paradigm named **Bro**wse-and-Concentra**te** (**Brote**). This paradigm begins with a browsing phase to generate a condition context vector, serving as a collection of browsing insights, encapsulating the main intent and visual information derived from images. Subsequently, a concentrating phase is employed to comprehend multimodal inputs, guided by the condition context vector. Furthermore, to enhance the effectiveness of the browsing insights, we have developed training strategies that prompt the model to implicitly leverage these insights for more precise extraction of image features, allowing for the possibility of bypassing explicit browsing in some scenarios. Our contributions can be summarized as follows:

- We address the challenge of prior-LLM modality isolation by proposing the browse-and-concentrate paradigm, alongside training strategies to encourage the model to leverage and explore the browsing insights.
- We explore two method to implement our paradigm, demonstrating that Brote not only learns to concentrate on interleaved inputs via explicit context vectors, but also integrates this ability directly into the model implicitly.
- We conduct comprehensive evaluations on 7 multi-image scenarios and exhibits notable advancements, improving the average accuracy by 2.13% and 7.60% against baselines with 3B and 11B LLMs, respectively.

2 Related Work

2.1 Empowering LLMs with Visual Abilities via Pre-trained Vision Models

With the surging of LLMs, MLLMs that empower LLMs with visual abilities have also witnessed a rapid growth. Following the initial effort (Tsimpoukelli et al., 2021) to convert visual features into readable embeddings for LLMs, researchers have proposed to bridge vision and language modalities via diverse visual prompt generators (VPG), such as Resampler (Alayrac et al., 2022), Q-Former (Li et al., 2023e; Dai et al., 2023), and linear projections (Liu et al., 2023d; Huang et al., 2023). They utilize image features from frozen vision models (Dosovitskiy et al., 2021; Radford et al., 2021), and subsequently integrate these features into pre-trained LLMs. These MLLMs inherit cognitive and perceptual abilities from vision models and the emergent ability from LLMs, exhibiting impressive performance without intensive training. However, they bear the modality isolation issue that obscures a deeper understanding multimodal context.

2.2 Enhancing Visual Features with Textual Instructions

Recent studies have concentrated on augmenting the capability for MLLMs to follow visual instructions (Liu et al., 2023d; Dai et al., 2023; Luo et al.,

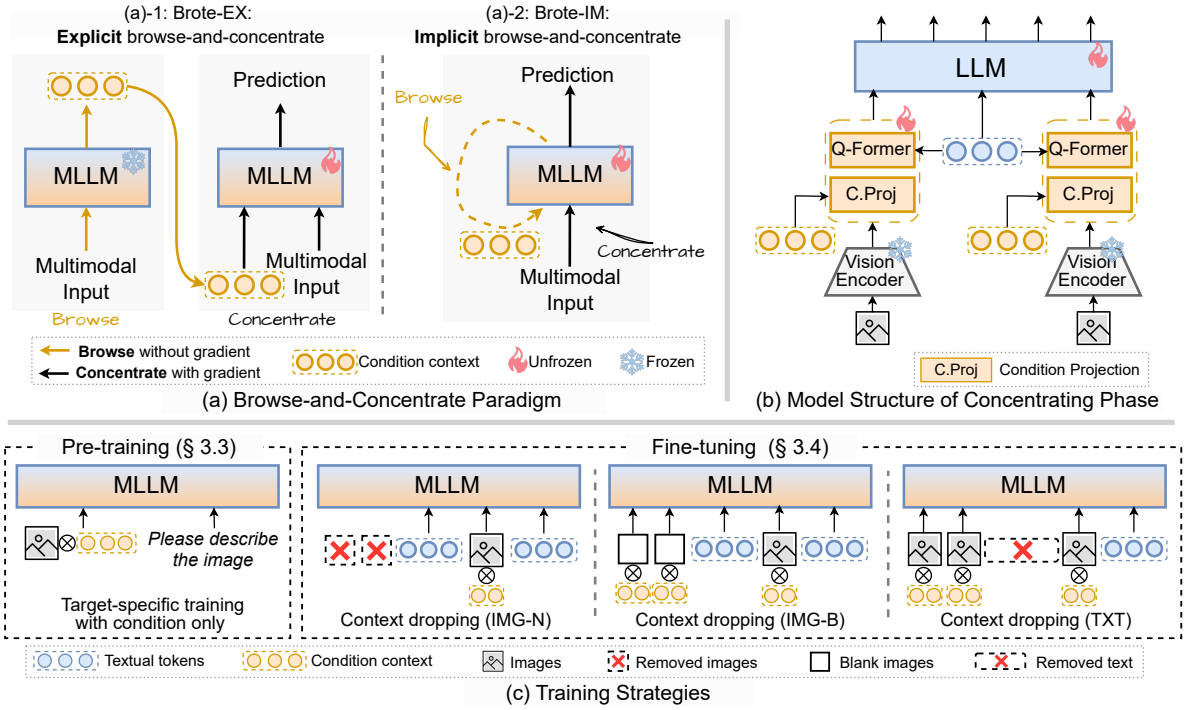

Figure 2: The illustration of browse-and-concentrate paradigm (a), model architecture of the concentrating phase (b), and our proposed training strategies (c). (a) shows the pipelines of Brote models, (a)-1 for Brote-EX and (a)-2 for Brote-IM. (c) depicts the strategies described in §3.3 and §3.4.

2023; Wang et al., 2023; Ye et al., 2023). Some researchers have focused on the fine-tuning LLMs to better response to visual instructions (Ye et al., 2023; Liu et al., 2023c), employing techniques such as LoRA (Hu et al., 2022a). While other studies target the issue of image-text isolation by manipulating with the visual features. For instance, Dai et al. (2023) enhance Q-Former with textual instructions to obtain instruction-aware visual features. Luo et al. (2023) integrate learnable instruction features directly into the vision encoders. Despite these innovations, they primarily incorporate instructions into the vision modules pay less attention to the complexity of multi-image senarios.

2.3 MLLMs Enhanced for Comprehending Multiple Images

The ability to comprehend multiple images simultaneously draws considerable attention (Alayrac et al., 2022; Zhao et al., 2024; Li et al., 2023b; Shukor et al., 2023). Multi-image scenarios can be categorized into interleaved image-text formats and multimodal ICL settings. To improve ICL preformance, Shukor et al. (2023) analyse prompt-based approaches, introducing three templates for multimodal ICL. Meanwhile, some researchers work on methods requiring model-tuning (Alayrac et al., 2022; Shukor et al., 2023; Sun et al., 2023). Additionally, some scholars broaden the exploration of multi-image scenarios to include both ICL and interleaved inputs. Li et al. (2023d) insert middle-layer LLM outputs into the VPG as additional guidance for spotting the differences between images. Zhao et al. (2024) and Li et al. (2023b) construct datasets targeting the multi-image issue, and propose context schemes to improve the understanding of interleaved inputs. Despite these advancements, the prior-LLM multimodal context fusion is not sufficiently explored.

3 Method

3.1 Overview

To stimulate prior-LLM multimodal context fusion and improve the awareness of multimodal context of the LLM, we propose a paradigm, **Bro**wse-and-concentra**te** (Brote). It progressively comprehends images via two phases, **browsing** and **concentrating**. As illustrated in Figure 2 (a), in the browsing phase, the MLLM browses the entire input and generates a condition context as the browsing result, denoted as $\mathcal{C}$ in the rest of this paper. Then, in the concentrating phase, the model comprehends multimodal inputs under the guidance of $\mathcal{C}$. We refer to the model of browsing phase as $\mathcal{M}_B$ and the model of concentrating phase as $\mathcal{M}_C$.

Our proposed Brote can be further divided into two modes, *explicit* and *implicit*, regarding the distinct approaches of incorporating $\mathcal{C}$. The explicit browse-and-concentrate (Figure 2 (a)-1), denoted as *Brote-EX*, operates with separated parameters ($\mathcal{M}_B \neq \mathcal{M}_C$). This explicit mode first generates $\mathcal{C}$ using $\mathcal{M}_B$, followed by $\mathcal{M}_C$ to infer the final outcomes. In contrast, for the implicit browse-and-concentrate (Figure 2 (a)-2), denoted as *Brote-IM*, employs shared parameters for both phases ($\mathcal{M}_B = \mathcal{M}_C$), permitting $\mathcal{M}_C$ to directly predict the answer without the need to explicitly produce intermediate vectors from the other model. Along with the proposed paradigm, we devise training strategies for the explicit browse-and-concentrate mode. This strategies encourage the model to leverage and explore the generated condition context vectors. The explicit mode serves as a precursor to the implicit mode, preparing the model with fundamental and essential ability to understand $\mathcal{C}$.

We will elaborately describe the workflow of Brote in §3.2, followed by the proposed strategies for pre-training (§3.3) and fine-tuning (§3.4).

3.2 Browse-and-Concentrate Paradigm

We represent the interleaved multimodal input as $\boldsymbol{x}$, defined as $\boldsymbol{x} = [x_0^m, x_1^m, \ldots, x_n^m, \ldots, x_{N-1}^m]$ for N tokens, with $n = 0, 1, \ldots, N-1$. Each token is associated with modality m, where $m \in \{\text{image}, \text{text}\}$. Images are individually encoded by vision encoder $g_{\phi_v}(\cdot)$ with parameters ϕ_v, which provides image features $\boldsymbol{v} = g_{\phi_v}(x_n^m)$, for $m = \text{image}$. Referring to $\text{Emb}(\cdot)$ as the embedding mapping, $\boldsymbol{v}$ is subsequently integrated with textual instructions $\boldsymbol{h}^{\text{text}} = \text{Emb}(x_n^m)$, for $m = \text{text}$, via a Q-Former $f_{\phi_Q}(\cdot,\cdot)$ parameterized by ϕ_Q,

$$\boldsymbol{h} = [h_0^m, h_1^m, \ldots, h_n^m, \ldots, h_{N-1}^m] \tag{1}$$

$$\boldsymbol{h}_n^m = \begin{cases} \text{Emb}(x_n^m) & \text{if } m = \text{text} \\ f_{\phi_Q}(\boldsymbol{v}, [\boldsymbol{Q}; \boldsymbol{h}^{\text{text}}]) & \text{if } m = \text{image} \end{cases} \tag{2}$$

where $\boldsymbol{h}$ denotes the multimodal embeddings, and $\boldsymbol{Q}$ is the learnable query tokens in Q-Former.

The LLM component of $\mathcal{M}_B$ is denoted by $f_{\phi_L}(\cdot)$ with parameters ϕ_L. The browsing phase produces $\mathcal{C}$ by extracting the last hidden states of the LLM $f_{\phi_L}(\cdot)$, denoting as follows:

$$\mathcal{C} = f_{\phi_L}^{(l)}(\boldsymbol{h}). \tag{3}$$

where l represents the last layer of $f_{\phi_L}(\cdot)$.

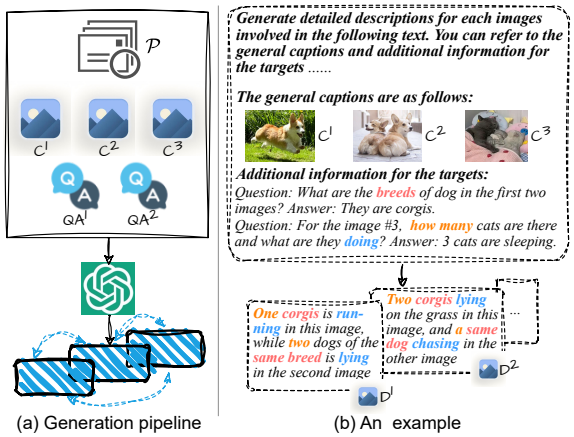

Figure 3: A illustration of data construction process (a) described in §3.3, with a detailed example (b). The generated descriptions should be aware of both the targets ("*breeds*", "*how many*", "*doing*") and another images.

In the concentrating phase, the images undergo alterations conditioned on $\mathcal{C}$. We add $\mathcal{C}$ to query tokens Q, and obtain the altered visual token embeddings $\tilde{\boldsymbol{h}}^{\text{image}}$ as,

$$\tilde{\boldsymbol{h}}^{\text{image}} = f_{\phi_Q}(\boldsymbol{v}, [\boldsymbol{Q} + \text{Linear}(\mathcal{C}); \boldsymbol{h}^{\text{text}}]), \tag{4}$$

where $\text{Linear}(\cdot)$ denotes the linear projection, and $[\cdot;\cdot]$ denotes concatenation. In this phase, Q-Former accepts an extra input $\mathcal{C}$ compare to the browsing phase. Finally, the prediction $\boldsymbol{y}$ with T tokens is formulated as follows:

$$\begin{aligned} \boldsymbol{y} &= \mathcal{M}_C(\boldsymbol{x}, \mathcal{C}) \\ &= \underset{\boldsymbol{y}}{\text{argmax}}\, p(y|\boldsymbol{x}, \mathcal{C}; f_{\phi'_L}, f_{\phi'_Q}, g_{\phi'_v}) \\ &= \underset{\boldsymbol{y}}{\text{argmax}} \prod_{t=1}^{T} p(y_t|y_{<t}, \boldsymbol{x}, \mathcal{C}; f_{\phi'_L}, f_{\phi'_Q}, g_{\phi'_v}), \end{aligned} \tag{5}$$

where y_t is the t-th token of the prediction, $y_{<t} = y_1, \cdots, y_{t-1}$, $f_{\phi'_L}$, $f_{\phi'_Q}$, $g_{\phi'_v}$ are the LLM, Q-Former, and vision encoder components of $\mathcal{M}_C$ parameterized by ϕ'_L, ϕ'_Q, and ϕ'_v, respectively.

3.3 Context-Enhanced Pre-Training

Condition context-enhanced pre-training. The pre-training stage aims at adapting the model to utilize $\mathcal{C}$ and enhancing visual feature extraction with its conveyed multimodal context. To this end, we propose a training task that challenges the model to generate task-specific descriptions without direct exposure to the question. Initially, we obtain $\mathcal{C}$ by feeding the intact inputs into $\mathcal{M}_B$. Then, in the concentrating phase, $\mathcal{M}_C$ is required to generate image descriptions specialised for the questions

that are not explicitly presented but instead implicitly encoded within $\mathcal{C}$. As depicted in Figure 2 (c) "Pre-training", the model is presented with only the text "*Please describe the image*" alongside altered visual tokens $\tilde{\boldsymbol{h}}^{\text{image}}$. This strategy urges the model to explore $\mathcal{C}$ for target information. Additionally, we combine the task-specific training targets together with the general ones, enabling the model to discern between inputs with and without $\mathcal{C}$. The objective for pre-training is as follows:

$$\mathcal{L}_{\mathcal{M}_C} = -\sum_{t=1}^{T} \hat{y}_t \log p(y_t|\boldsymbol{x}, \mathcal{C}; f_{\phi'_L}, f_{\phi'_Q}, g_{\phi'_v}), \quad (6)$$

where $\hat{y}_t$ is the t-th groundtruth token.

Data construction. In alignment with the task-specific training strategy, we design a data generation method to secure task-specific supervisions as mentioned above. Inspired by PromptCap (Hu et al., 2022b), we leverage LLMs to craft target-aware image descriptions. Our approach is extended from the producing of individual image descriptions to addressing multiple interleaved inputs, enabling a more profound understanding of multi-image and interleaved context. We obtain the image-target related descriptions as demonstrated in Figure 3. The LLM receives a triplet $(\mathcal{P}, C^K, QA^J)$, comprising task instruction prompt $\mathcal{P}$, general image descriptions C^K and question-answer pairs QA^J, where K and J represent the counts of images and targeted question-answer pairs respectively. Also noticing that K is not necessarily equal to J. For each k-th image ($k = 1, 2, \ldots, K$), the LLM is required to generate image description D^k that satisfies the target clarified in $\mathcal{P}$. Accordingly, D^k contains specific messages for questions in QA^J and information about the other related images o, $k \neq o$.

We construct a total of 56k data for the pre-training stage, and manually assess the quality of the generated captions by randomly sampling 230 generated captions. We detect that 36 (out of 230) captions contain hallucination or minor incorrect information, while the rest 84% are of good quality, containing desired and correct question-aware information. Please refer to Appendix A for details of the generated data.

3.4 Condition-Aware Task Fine-Tuning

To encourage further exploration of information from $\mathcal{C}$ for VL tasks, we propose a new training strategy named *context-dropping training*. The strategy intentionally omits particular inputs yet requiring the model to infer for answers solely with the assistant of $\mathcal{C}$. It motivates the model to compensate for the missing information from the provided condition context $\mathcal{C}$. We propose different dropping strategies as illustrated in Figure 2 (b):

- Drop images: This involves two approaches, removing certain images (Figure 2 (b), "Context Dropping (IMG-N)"), and replacing original images by blank placeholders (Figure 2 (b), "Context Dropping (IMG-B)").
- Drop text: We remove the text before the last image as shown in Figure 2 (b), "Context Dropping (TXT)".
- Drop ALL: A combination of the above settings denoted as "ALL", applied with the same probabilities.

To ensure integration with $\mathcal{C}$, we preserve the last image across all dropping strategies. Notice that the "drop images" approaches are not applicable to inputs with only one image. These strategies compel the model to infer indispensable information from $\mathcal{C}$ that should have been given in the input.

As mentioned in §3.1, we investigate two modes for incorporating $\mathcal{C}$, Brote-EX and Brote-IM. For Brote-EX, we apply context-dropping strategies to the concentrating phase with $\mathcal{C}$ provided by frozen model $\mathcal{M}_B$. The training objective for explicit mode is $\mathcal{L}_{\mathcal{M}_C}$ as described in Equation 6. While for Brote-IM, parameters of $\mathcal{M}_B$ are shared with $\mathcal{M}_C$. When optimizing the shared parameters, we also take into account the loss for $\mathcal{M}_B$ as follows:

$$\mathcal{L}_{\mathcal{M}_B} = -\sum_{t=1}^{T} \hat{y}_t \log p(y_t|\boldsymbol{x}; f_{\phi_L}, f_{\phi_Q}, g_{\phi_v}). \quad (7)$$

For the training of Brote-IM, we sum up the two losses, for $\mathcal{M}_B$ and $\mathcal{M}_C$ respectively, as $\mathcal{L}_{\mathcal{M}_B} + \mathcal{L}_{\mathcal{M}_C}$, denoted by *dual-loss*. Details of the training process are documented in Appendix C.

4 Experiments

4.1 Implementation

We implement our method upon InstructBLIP (Dai et al., 2023) with FlanT5 (Chung et al., 2022) as the language backbone. We pre-train our model on the 56k generated data as described in §3.3, and then extract about 490k data from the MIC dataset (Zhao et al., 2024) for model fine-tuning. The fine-tuning data is sampled according to the data-balanced sampling algorithm suggested by

	Model	#Param LLM	In-context Learning		Multi-image / Video Tasks					AVG
			VQAv2	A-OKVQA	NLVR2	DEMON	SEED	MSVD QA	MSRVTT QA	
#LLM ≤ 10B	KOSMOS-1	1.3B	51.8	-	-	-	-	-	-	-
	InstructBLIP-XL	3B	31.76*	39.13*	52.59*	32.59*	52.7	43.40*	12.12*	37.77
	MMICL-XL$^{\diamond}$	3B	69.16	53.43*	71.48*	**38.14***	54.69*	53.68	42.36*	54.71
	Otter	7B	45.39*	38.42*	49.54*	24.51	39.7	25.87*	9.78*	33.32
	VPG-C-LLaMA2	7B	-	34.29*	53.82*	37.22	-	6.03*	-	-
	Flamingo-9B	7B	56.3	-	-	-	-	30.2	13.7	-
	Brote-EX-XL (ours)	3B	**69.97**	56.00	71.41	37.33	57.51	53.02	43.14	55.48
	Brote-IM-XL (ours)	3B	68.94	**56.43**	**76.02**	37.34	**57.86**	**56.06**	**45.08**	**56.84**
#LLM > 10B	InstructBLIP-XXL	11B	48.21*	45.92*	64.54*	33.00*	50.81*	44.30*	15.49*	43.18
	MMICL-XXL$^{\diamond}$	11B	70.56	54.85*	56.16*	36.30*	56.66*	52.19	39.46*	52.18
	EMU-2	33B	67.0	-	-	-	**62.8**	49.0	31.4	-
	Flamingo-80B	70B	63.1	-	-	-	-	35.6	17.4	-
	Brote-EX-XXL (ours)	11B	70.86	59.94	70.42	38.70	59.31	54.52	45.24	57.00
	Brote-IM-XXL (ours)	11B	**71.71**	**60.31**	**80.71**	**38.94**	61.64	**57.29**	**45.94**	**59.78**

Table 1: Results for multi-image settings. The best results for models larger/smaller than 10B are separately **bolded** and the seconds are underlined. $^{\diamond}$: the InstructBLIP version. We evaluate results which are not officially announced using public checkpoints and mark them by *. SEED refers to SEED-Bench that contains both images and videos.

	Model	#Param LLM	VQAv2	A-OKVQA	ScienceQA -IMG	MME Perception	MME Cognition	MMBench	AVG
#LLM ≤ 10B	InstructBLIP-XL	3B	36.77	**54.57**	70.40	1093.70*	281.43*	69.68*	68.52
	MMICL-XL	3B	69.13	52.12*	**72.58***	1184.54*	277.86*	73.11*	75.81
	LLaVA†	7B	-	-	-	457.82	214.64	36.2	-
	Otter†	7B	57.89*	41.92*	63.10	**1292.26**	**306.43**	48.3	69.51
	Brote-EX-XL (ours)	3B	69.90	52.93	71.15	1203.87	301.79	73.27	**77.18**
	Brote-IM-XL (ours)	3B	**70.24**	53.40	**72.58**	1181.95	266.79	**74.29**	75.90
#LLM > 10B	InstructBLIP-XXL	11B	63.69	57.10	70.60	1212.82*	291.79*	70.34*	75.99
	MMICL-XXL	11B	70.30	51.35*	74.92*	1313.88*	311.79*	76.58*	80.41
	MMICL-XXL (BLIP-2)†	11B	69.99	-	-	**1381.74**	**428.93**	65.24	-
	Brote-EX-XXL (ours)	11B	71.58	56.47	77.69	1279.73	310.01	76.67	81.31
	Brote-IM-XXL (ours)	11B	**73.02**	**57.83**	**78.38**	1284.13	300.00	**77.34**	**81.66**

Table 2: Zero-shot results for single-image settings. The best results for models larger/smaller than 10B are separately **bolded** and the seconds are underlined. †: results of these models are taken from Zhao et al. (2024). We evaluate results which are not officially announced using public checkpoints and mark them by *. For "AVG", we first average the MME scores over its subtasks, then calculate the average scores of all benchmarks in this table. We include closely related baselines in this table, and refer readers to Appendix F for detailed results of other models.

Dai et al. (2023). Please refer to Appendix B for details of the training data and Appendix C for more information of the training process.

4.2 Evaluation Settings

Baselines. We primarily employ models designed for accepting multiple images or interleaved image-text inputs as baselines, such as MMICL (Zhao et al., 2024), Otter (Li et al., 2023b) and VPG-C (Li et al., 2023d). Additionally, MLLMs that are used to develop these baselines are also considered, such as BLIP-2 (Li et al., 2023e) and InstructBLIP (Dai et al., 2023). Please refer to Appendix D for detailed information of the employed baselines. For models whose results are not officially reported, we utilize the publicly available checkpoints for evaluation.[2]

Benchmarks and Metrics. We investigate diverse VL benchmarks and focus on multi-image tasks, including visual reasoning (NLVR2 (Suhr et al., 2019)), few-shot ICL for image QA (VQAv2 (Goyal et al., 2017) and A-OKVQA (Schwenk et al., 2022)), video QA (MSVD QA (Xu et al., 2017), MSRVTT QA (Xu et al., 2017), SEED-Bench (Li et al., 2023c)), and

[2]We use the public checkpoints to obtain the missing results for MMICL (https://huggingface.co/BleachNick), InstructBLIP (https://huggingface.co/Salesforce), and Otter (https://huggingface.co/luodian), together with official scripts and required environments.

	Dataset	Settings	MMICL	Brote-EX		Brote-IM	
				Ours	Ours-None	Ours	Ours-None
XL	A-OKVQA	0-shot	52.12	52.93	49.40 -3.53	53.40	51.88 -1.52
		4-shot	53.43	56.00	55.10 -0.90	56.53	56.28 -0.25
	SEED	Image	57.99	61.90	59.79 -2.11	61.82	61.48 -0.34
		Video	41.94	40.50	39.06 -1.44	42.52	42.11 -0.41
	NLVR2	0-shot	71.48	71.41	69.27 -2.14	76.02	75.59 -0.43
	Average		55.39	56.55	54.52 -2.03	58.06	57.47 -0.59
XXL	A-OKVQA	0-shot	51.35	56.47	55.32 -1.15	57.83	57.61 -0.22
		4-shot	54.85	59.94	58.70 -1.24	60.65	60.25 -0.40
	SEED	Image	59.17	63.70	63.26 -0.44	65.58	64.65 -0.93
		Video	46.90	42.27	41.88 -0.39	46.37	46.25 -0.12
	NLVR2	0-shot	53.62	70.42	68.60 -1.82	80.71	79.69 -1.02
	Average		53.18	58.56	57.55 -1.01	62.23	61.69 -0.54

Table 3: Ablation study of the condition context vectors. "Ours-None" indicates the none condition setting (replacing the condition by all-zero vectors when testing).

multi-image instruction following (DEMON (Li et al., 2023d)). Note that SEED-Bench comprises of both images and videos. For video benchmarks, following Zhao et al. (2024), we uniformly extract eight frames from the given video clips for answering the questions. For few-shot ICL, we employ the widely used four-shot setting. Additionally, we conduct experiments on single-image tasks to fairly compare with models that are not designed for multi-image settings. These tasks include zero-shot setting for VQAv2, A-OKVQA and MME (Fu et al., 2023). ScienceQA (Saikh et al., 2022) (SciQA) is designed for Chain-of-Thought (CoT) (Wei et al., 2022) scenario with accompany hints, and we adopt the zero-shot CoT (ZS-CoT) setting for this dataset. Details of these evaluation benchmarks, including data scale, the type of tasks and evaluation metrics, are listed in Appendix E.

4.3 Results

We report results for multi-image settings in Table 1 and single-image settings in Table 2. Drawing conclusions from these tables, our method presents significant improvement for multi-image settings, while concurrently improves the performance of 3 single-image tasks.

Our models exhibit notable advancements over models in Table 1, showing profound comprehending ability for multi-image and interleaved inputs. We outperform strong baselines, such as InstructBLIP, MMICL and VPG-C, which include shallow prior-LLM instruction-image fusion. Our method goes beyond merely cross-modality integration between image and text to also include intra-modality fusion among images. Impressively, our models show consistent advantage over benchmarks involving videos and multiple images, and for few-shot ICL of QA tasks as well. For the average scores of models following InstructBLIP paradigm, our models achieve improvements of 2.13% and 7.60% for XL and XXL models respectively, over MMICL.

For single-image tasks reported in Table 2, our models continue to manifest progress, presenting higher average scores. We improve the performance for two zero-shot VQA tasks and one MLLM benchmark, MMBench. However, our models only show modest performance on MME.

Models	PT	FT	Drop	AVG-Multi	AVG
InstructBLIP	-	-	-	42.56	50.34
Ours-sampled	✘	✓	✘	46.51 (+3.96)	51.85 (+1.51)
Ours-sampled	✓	✓	✘	47.12 (+4.56)	50.94 (+0.50)
Ours-sampled	✓	✓	IMG-N	48.06	52.09
Ours-sampled	✓	✓	IMG-B	48.06	51.90
Ours-sampled	✓	✓	TXT	48.08	52.08
Ours-sampled	✓	✓	ALL	**48.87** (+6.31)	**52.39** (+2.05)
MMICL	-	-	-	47.05	51.68

Table 4: Ablation study of different training strategies on XL-sized (3B LLM) models. "PT" refers to pre-training, and "FT" denotes fine-tuning. "Ours-sampled" is described in §4.4. "AVG-Multi" is the average score for multi-image settings, including A-OKVQA 4-shot, NLVR2, SEED video split and MSVD QA. "AVG" refers to the average score over 7 tasks, with detailed results presented in Appendix G.

4.4 Ablation Study

Impact of condition context vectors. To determine whether condition context vectors $\mathcal{C}$ contribute to the improvement, we conduct ablation study by removing $\mathcal{C}$ and observe a decline in accuracy across various tasks, evaluation settings, and model scales. In detail, we replace these vectors by zero vectors to simulate the absence of $\mathcal{C}$. Experiments are conducted on zero-shot and few-shot VQA, and multi-image visual reasoning tasks. As shown in Table 3, the models augmented by $\mathcal{C}$ ("Ours") consistently outperform those with zero vectors ("Ours-None"). The most substantial average discrepancy is observed in Brote-EX at the XL scale (2.03%), while the smallest gap is presented by Brote-IM at the XXL scale (0.54%). We notice that Brote-EX tends to gain more directly from $\mathcal{C}$ compared to Brote-IM, and conclude that Brote-IM directly integrates additional benefits provided by $\mathcal{C}$ into the model through dual-loss training. More sophisticated analysis are documented in §5.1.

Impact of different training strategies. For efficient iteration and validation, we create a subset by sampling one-third of the training data for ablation

Model	A-OK	NLVR2	MSVD	SEED	AVG	Gain
Brote-EX	56.00	71.41	53.02	57.51	59.49	-
Brote-EX (+*2epoch*)	55.83	75.07	55.60	57.60	61.03	1.54
Brote-IM	**56.53**	**76.02**	**56.06**	**57.86**	**61.62**	**2.13**

Table 5: Results of continue training with XL models. "Brote-EX (+*2epoch*)" is training Brote-EX for 2 extra epochs using dual-loss without providing $\mathcal{C}$ for $\mathcal{M}_C$. "A-OK" is A-OKVQA for short. "Gain" implies the increment from extra epochs over original Brote-EX.

studies on different training strategies, denoting the resulting models as *Ours-sampled*. For fair comparison, we also reproduce MMICL-XL (with InstructBLIP backbone) using this subset[3]. We evaluate the average scores of training strategies described in §3.4 and §3.3, with a special focus on prior-LLM multimodal context fusion for multi-image scenarios. The averaged scores for multi-image tasks and the overall tasks are reported in Table 4, with detailed results provided in Appendix G. The performance of InstructBLIP serves as the baseline and is used to indicate the contribution of each of the designed strategies. Summarised from Table 4, the models equipped with context-dropping strategies yield higher average scores. Notably, the dropping-ALL strategy presents the highest average scores of both multi-image and overall tasks, showing profound multimodal context handling ability. We consequently adopt this strategy for training our Brote models.

5 Discussions and Analysis

5.1 Explicit Versus Implicit

As discussed in §4.4, Brote-EX exhibits a more significant benefit from $\mathcal{C}$ compared to Brote-IM. We propose two potential reasons for this observation:

- Brote-IM gains advantages from extra training steps rather than insights provided by $\mathcal{C}$;
- Brote-IM effectively incorporates the capabilities afforded by $\mathcal{C}$ into the parameters of LLM and Q-former during the training process.

For further investigation, we extend the training of Brote-EX with the same configurations and objectives as applied to Brote-IM, except that we zero out $\mathcal{C}$ for the concentrating phase. Specifically, we replace $\mathcal{C}$ in Equation 5 by zero vectors. Brote-IM is trained for two epochs based on Brote-EX as detailed in Appendix C. Hence, we train Brote-EX for additional two epochs, denoting this model as Brote-EX (+*2epoch*). We report the results in Table 5. The results reveal that the observed improvements of Brote-IM over Brote-EX are not solely attributable to increased training steps. Rather, the improvements stem from the integration with $\mathcal{C}$ during training. Although Brote-EX (+*2epoch*) presents an average increase of 1.54% over Brote-EX, Brote-IM exhibits an additional 0.59% average improvement over Brote-EX (+*2epoch*) with the participant of $\mathcal{C}$ during training, culminating in a total increment of 2.13% over Brote-EX.

Furthermore, as detailed in Table 3, the absence of $\mathcal{C}$ does not prevent Brote-IM models from outperforming Brote-EX. This finding supports the conclusion that Brote-IM integrates the function of $\mathcal{C}$ into the model parameters themselves without explicitly generate $\mathcal{C}$ from the other model, facilitating a more profound comprehension of multimodal inputs. In contrast, Brote-EX relies on an extra explicit representation, condition context $\mathcal{C}$, to obtain multimodal comprehension and achieve good performances. The superior performance of Brote-IM affirms the efficacy of the dual-loss training strategy. Brote-IM markedly benefits from $\mathcal{C}$, thereby enabling the development of a more apt parameter set for multimodal context.

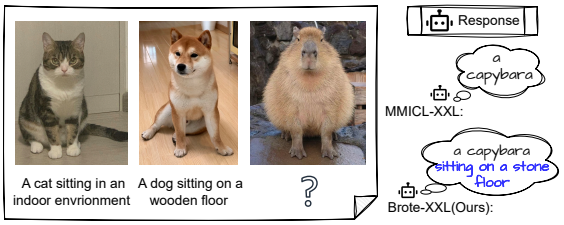

Figure 4: A case showing that our method is more coherent to the given multimodal context.

5.2 Case Study

In Figure 4, we illustrate a case study on multimodal ICL, highlighting the coherent performance of our model in response to the input. Specifically, our model demonstrates an acute awareness of the target information conveyed through the multimodal inputs, capturing an intra-image connection characterized by "an *animal* sitting on/in *a certain place*". Compared to MMICL, our model produces a response that precisely aligns with the input, showcasing its profound ability to comprehend multimodal contexts.

[3]We use the published code from https://github.com/HaozheZhao/MIC

5.3 Efficiency of Inference

We investigate the efficiency of inference with our models in terms of both time and GPU memory, as our methods involve two forward iterations and an additional $\mathcal{C}$ compared to other InstructBLIP-based models. We conduct experiments with XL models using single NVIDIA A100 GPU, with batch size 10 and data type float32. Results indicate that Brote-EX requires almost equal GPU memory (18G) and inference time (around 2.3 second per batch) compared to MMICL. However, Brote-IM exhibits an increase of GPU memory from 18G to 24G for an additional "browsing" iteration, and doubles the time cost to 5 second per batch.

6 Conclusion

In this paper, we address the prior-LLM modality isolation issue for both image-text and inter-image context, which lacks sufficient investigation in previous works. To mitigate this issue, we propose browse-and-concentrate paradigm that leverages the initial browsing insights for the prior-LLM multimodal context fusion to stimulate more profound comprehending of multi-image and interleaved inputs. We present in-depth analysis on our proposed training strategies and the two approaches for implementing our proposed paradigm. The two approaches, explicitly or implicitly browse through and then concentrate on the context, exhibits comprehensive multimodal context understanding. Our method demonstrates remarkable improvements on 7 multi-image tasks against strong baselines that enable prior-LLM image-text fusion.

Limitations

We conlude the limitations of our method as follows: First, although presenting improved results for multi-image scenarios, our method does not achieve equally impressive performances across all single-image tasks evaluated. This discrepancy can be attributed to the employed backbone models (InstructBLIP), which already incorporate the textual instructions into the visual feature extraction process, partially addressing the challenge of prior-LLM modality isolation we aim to overcome. Our future work includes validating the proposed paradigm on broader backbone models. Second, we do not specifically incorporate datasets designed for visual instruction tuning, such as LLaVA (Liu et al., 2023d), which could be a reason for the modest performance on MME benchmark. In this paper, we primarily focuse on multi-image scenarios, such as question-answering and visual reasoning, without a particular emphasis on following visual instruction. Third, as we introduce a two-phase paradigm, the time cost and the required GPU memory for inference with Brote-IM are also increased.

Acknowledgements

This work is supported by the National Key R&D Program of China (2022ZD0160502) and the National Natural Science Foundation of China (No. 61925601, 62276152). We appreciate all the reviewers for their insightful suggestions. We thank Siyu Wang for her participation in this work, and Haozhe Zhao for his technical support. We thank Tong Su for providing photos presented in Figure 4, and Alan (the cat) for being the model.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4290–4300. IEEE.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias

Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *The Nineth International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models.

Ruth Garner. 1987. *Metacognition and reading comprehension.* Ablex Publishing.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022b. PromptCap: Prompt-guided task-aware image captioning. *ArXiv preprint*, abs/2211.09699.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *ArXiv preprint*, abs/2306.05425.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A multi-modal model with in-context instruction tuning. *ArXiv preprint*, abs/2305.03726.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023c. SEED-Bench: Benchmarking multimodal llms with generative comprehension.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023d. Fine-tuning multimodal llms to follow zeroshot demonstrative instructions. *ArXiv preprint*, abs/2308.04152.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023e. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. 2023b. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. *ArXiv preprint*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023e. Mmbench: Is your multi-modal model an all-around player?

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *ArXiv preprint*, abs/2305.15023.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Zhangyang Qi, Ye Fang, Mengchen Zhang, Zeyi Sun, Tong Wu, Ziwei Liu, Dahua Lin, Jiaqi Wang, and Hengshuang Zhao. 2023. Gemini vs GPT-4V: A preliminary comparison and combination of vision-language models through qualitative cases.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. ScienceQA: a novel resource for question answering on scholarly

articles. *International Journal on Digital Libraries*, 23(3):289–301.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: a benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer.

Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2023. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *ArXiv preprint*, abs/2310.00647.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative multimodal models are in-context learners.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 200–212.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogVLM: Visual expert for pretrained language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1666–1677. IEEE.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv preprint*, abs/2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *ArXiv preprint*, abs/2306.13549.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024. MMICL: Empowering vision-language model with multi-modal in-context learning. In *The Twelfth International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MIMIGPT-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, abs/2304.10592.

A Details of Our Constructed Data for Pre-training

In this section, we provide details of the data generation process mentioned in §3.3. Inspired by PromptCap (Hu et al., 2022b), we employ LLM APIs to generate target-aware image descriptions, but explore broader types of tasks. Extending from single-image descriptions, we require the LLM[4] to generate descriptions regarding other images as well, enabling a more profound understanding of multi-image and interleaved content. We utilize

[4]We use the GPT-4 API with version "gpt-4-1106-preview"

datasets targeting difference aspects of visual reasoning, including the maintenance of general world knowledge, the spatial and temporal information, the OCR ability, and the ability to distinguish differences of images. These datasets are as follows:

- **VQAv2** (Goyal et al., 2017) is a single-image visual captioning dataset. It is utilized to consolidate the general ability of our model.
- **ST-VQA** (Biten et al., 2019) and **IconQA** (Lu et al., 2021) are two single-image VQA datasets. They primarily conrtibute the tasks of OCR, object identification and counting.
- **VSR** (Liu et al., 2023a) is a VQA dataset addressing the spatial relation between two objects. It is employed to enhance the spatial reasoning ability.
- **VCR** (Zellers et al., 2019) is a visual reasoning dataset with images extracted from video scenes, focusing on the relations between presenting figures and objects.
- **NLVR2** (Suhr et al., 2019) and **MIMIC-IT(CGD)** (Li et al., 2023b) both contain two interleaved images with text. They help with the ability to distinguish the differences between two images.
- **iVQA** (Yang et al., 2021) is a video question answering dataset. We utilize this dataset to promote the ability to deal with the sequential information.

For datasets with naturally interleaved formats, VCR, NLVR2, and MIMIC-IT, we directly employ them to prompt LLMs, aiming to generate task-specific and multi-image aware descriptions. For other datasets, VQAv2, ST-VQA, IconQA, VSR, and iVQA, we adopt their few-shot versions from the MIC dataset (Zhao et al., 2024). In these versions, single-image instances are reconfigured into a few-shot format, featuring one or multiple images for zero to eight shots. These adapted instances serve as QA^J, as detailed in §3.3, with the corresponding questions designed as targeted tasks for LLM responses. The statistics of our generated data are listed in Table 6. We split instances containing multiple images into single image paired with the corresponding descriptions, and use them for pre-training as described in §3.3.

Original Dataset	Task Type	Format	#Generated Data	#Training Pairs
♢IconQA	VQA	S + M	1.8k	5.4k
♡VSR	VQA	S + M	3.0k	14.4k
♢VQAv2	VQA	S + M	6.5k	19.3k
△STVQA	VQA	S + M	10.0k	27.2k
□NLVR2	Reasoning	I	10.0k	20.0k
♠CGD*	Reasoning	I	10.2k	20.5k
♣VCR	Reasoning	I	10.0k	62.0k
△iVQA	Video-QA	I	5.2k	22.4k
	Total		56.7k	191.2k

Table 6: The statistics of our pre-training data. "S": **single**-image input; "M": **multi**-image input, wherein this case, the multiple images come from the few-shot examples of sinlge-image QA pairs; "I": naturally **interleaved** data, such as VCR and videos. "CGD*": MIMIC-IT CGD task (Li et al., 2023a). ♢: datasets licensed under CC-BY 4.0. ♡: datasets licensed under Apache License, Version 2.0. ♠: datasets licensed under MIT License. ♣: datasets licensed under Custom License. △: datasets with unknown license. □: datasets with CC-BY 4.0 License for annotations and unknown license for images.

A.1 Prompt Templates for Pre-training Data Generation

To be consistent with §3.3, we represent our employed prompt including task instruction $\mathcal{P}$, the general image descriptions C^K, and the question-answer pairs QA^J, where K and J are the number of images and targeted question-answer pairs. We provide prompt templates for the included datasets as belows:

Prompt Template for Data Generation

General task instruction $\mathcal{P}$
========
The general descriptions <caption> for each image are as follows:
C^1, C^2, ..., C^K
========
Here are the additional information of <Question>-<Answer> you should focus on!
QA^1, QA^2, ..., QA^J

Detailed task instruction of $\mathcal{P}$ for different datasets are as follows:

- Task Instruction for VQAv2 & VCR & IconQA & ST-VQA & VSR:
  Generate detailed captions of each image involved in the following text according to the original caption and the given <Question>-<Answer> pairs. You should pay attention to the information in the <Answer>. Your

output should be in the json format, as {"image0":"", "image1":"", "image2":""}. Your output should also be natural as an original caption and not include words like "answer" or "caption"!

- Task Instruction for NLVR2:
  Generate detailed captions of each image involved in the following text according to the original caption and the given <Question>-<Answer> pairs. You should pay attention to the information in the <Answer>. Your output should be in the json format, as {"image0":"", "image1":"", "image2":""}. Your output should also be natural as an original caption and not include words like "answer" or "caption"! You should also notice that <image0> is the left image and <image1> is the right image.

- Task Instruction for iVQA:
  Generate detailed captions of each image involved in the following text according to the original caption and the given <Question>-<Answer> pairs. You should pay attention to the information in the <Answer>. Your output should be in the json format, as {"image0":"", "image1":"", "image2":""}. Your output should also be natural as an original caption and not include words like "answer" or "caption"! You should notice that there exists sequential information between images!

- Task Instruction for Task Instruction for MIMIC-IT(CGD):
  Generate detailed captions of each image involved in the following text according to the original caption and the given <Option>-<Answer> pairs. You should pay attention to the information in the <Answer>. Your output should be in the json format, as {"image0":"", "image1":"", "image2":""}. Your output should also be natural as an original caption and not include words like "answer" or "caption"! Your output should also not clearly contain comparison while the information in <Option>-<Answer> pair should be presented!

Here is a detailed example from ST-VQA:

An Example for Data Generation

```
Generate detailed captions of each image
involved in the following text according
to the original caption and the given
<Question>-<Answer> pairs.  You should
pay attention to the information in the
<Answer>.  Your output should be in the
json format, as {"image0":"", "image1":"",
"image2":""}.  Your output should also be
as natural as an original caption and not
include words like "answer" or "caption"!
========
The original caption for each image are as
follows.
<image0>:  a sign with chinese characters
on it;
<image1>: a man is walking down a hallway
with a television above him.
========
Here are the additional information that
you should focus on!

The image 0:  <image0> is the primary
source of information for answering the
questions.  Please refer to it carefully
when answering question:  What does the
street sign say? Answer: anping jie

Answer each question based on the
information presented in image 1: <image1>.
Given the picture <image1>, what is the
answer to the question:  What does the
green sign say? Answer: exit
```

The corresponding outputs for the two images are "The green street sign displays the words 'anping jie' in Chinese characters." and "A man is strolling through a hallway while a television monitor is mounted above him alongside an indication of an 'exit' on a green sign."

B Training data for Model Fine-tuning

The select 17 datasets targeting different tasks from MIC dataset. Following Dai et al. (2023) and Zhao et al. (2024), We sampled about 490k instances from MIC according to this equation:

$$p_d = \frac{\sqrt{N_d}}{\sum_D^{i=1}\sqrt{N_i}}, \tag{8}$$

where p_d refers to the probability to select N instances of dataset d, from a total of D datasets. We list the involved datasets in Table 7.

C Model Training

This section describes detailed pre-training and fine-tune setting.

Dataset	Task	Format
♠COCO	Captioning	paired & few-shot
△Flickr	Captioning	paired & few-shot
△MSRVTT	Captioning	interleaved
♡VSR	Visual Reasoning	paired & few-shot
□NLVR2	Visual Reasoning	interleaved
♠VCR	Visual Reasoning	interleaved
△OKVQA	VQA	paired & few-shot
♢VQAv2	VQA	paired & few-shot
△GQA	VQA	paired & few-shot
△STVQA	VQA	paired & few-shot
♢TextVQA	VQA	paired & few-shot
♡RefCOCO	VQA	paired & few-shot
♣WikiART	VQA	paired & few-shot
♢IconQA	VQA	paired & few-shot
△iVQA	Video QA	interleaved
△MSVD	Video QA	interleaved
△MiniImageNet	Classification	paired & few-shot

Table 7: An overview of our fine-tuning data. ♢: datasets licensed under CC-BY 4.0. ♡: datasets licensed under Apache License, Version 2.0. ♠: datasets licensed under Custom License. ♣: datasets licensed under Non-commercial. △: datasets with unknown license. □: datasets with CC-BY 4.0 License for annotations and unknown license for images.

C.1 Pre-training

We initially acquire all the condition context vectors for the data outlined in Appendix A by MMICL models, where MMICL-XL and MMICL-XXL models are employed for Brote-XL and Brote-XXL, respectively. Leveraging these vectors, we bypass the forward iteration stage during pre-training and directly proceed the concentrating phase. We set the learning rate for the condition projection at 1×10^{-4} and for both the Q-Former and the language projection at 1×10^{-5}, applying a cosine learning rate scheduler. These experiments are conducted on the NVIDIA A100 GPU, with the pre-training configurations detailed in Table 8. As a complement to Figure 2, we provide a detailed models structure in Figure 5.

C.2 Fine-tuning

In the pre-training stage, we adapt the parameters of the Q-Former and the condition projection to effectively integrate $\mathcal{C}$, enhancing the models' ability to interpret multimodal contexts. Based on this, the subsequent fine-tuning encompasses both browsing and concentrating phases. Following Dai et al. (2023) and Zhao et al. (2024), we fine-tune our model on multiple originated datasets to enable the ability to accomplish practical and diverse tasks. As described in §3.1, we develop two approached for incorporating $\mathcal{C}$, each predicated on differing objectives: For Brote-EX, condition context vectors are derived from the frozen MMICL model, whereas Brote-IM generates these vectors internally. Training specifics for Brote-EX-XL involve four epochs focused on the objective $\mathcal{L}_{\mathcal{M}_C}$, while Brote-IM-XL extends this with two epochs under a dual-loss objective, $\mathcal{L}_{\mathcal{M}_B} + \mathcal{L}_{\mathcal{M}_C}$, starting from the Brote-EX-XL foundation. For the XXL models, the training duration for both explicit and implicit training modes are adjusted to half that of their XL counterparts. Detailed configurations are listed in Table 9, not that the settings of learning rates identical to that of pre-training stage.

Model Scale	Epoch	Batch Size	Gradient Accu. Steps	Warmup Portion	GPUs
XL	4	10	8	0.2	4
XXL	4	2	4	0.2	4

Table 8: The pre-training settings. "Gradient Accu. Steps" refers to the gradient accumulation steps.

Model Scale	Epoch	Batch Size	Gradient Accu. Steps	Warmup Portion	GPUs
Brote-EX-XL	4	10	8	0.2	4
Brote-IM-XL	2	10	8	0.2	4
Brote-EX-XXL	2	2	4	0.2	4
Brote-IM-XXL	1	1	4	0.2	4

Table 9: The fine-tuning settings. "Gradient Accu. Steps" refers to the gradient accumulation steps.

D Baselines

We compare to MLLMs who also notice the multi-image scenarios, including MMICL (Zhao et al., 2024), Otter (Li et al., 2023b), VPG-C (Li et al., 2023d), KOSMOS-1 (Huang et al., 2023) and EMU (Sun et al., 2023). For models that are not initially designed for accepting multiple images, such as InstructBLIP, we concatenate the visual embeddings for all the input image together to enable the multi-image processing ability. The details of the baselines are listed together with the results in Table 11 and Table 12.

E Benchmarks and Metrics for Evaluation

The employed benchmarks and corresponding metrics are listed in Table 10. We investigate diverse conventional VL benchmarks and recently proposed MLLM benchmarks, including VQAv2 (Goyal et al., 2017), A-OKVQA (Schwenk et al., 2022), ScienceQA (Saikh et al., 2022), NLVR2 (Suhr et al., 2019), MSVD QA (Xu

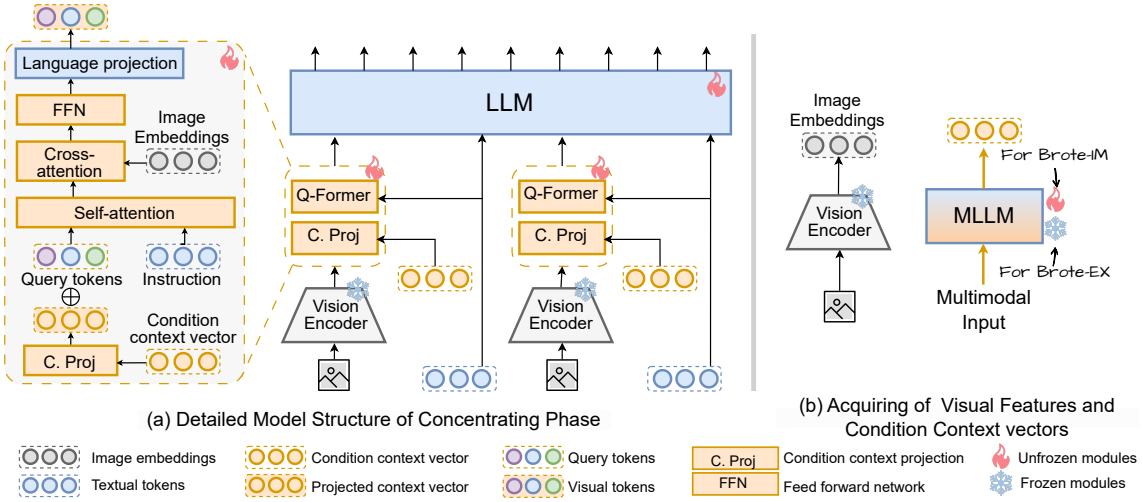

Figure 5: The detailed model structure for concentrating phase.

Benchmark	Data Format	Answer Type	Setting	Data Split	Metrics
NLVR2	Multi-image	True/False	Zero-shot	Test	Accuracy
DEMON-Core	Multi-image	Multiple-choice & Open-ended	Zero-shot	-	I4-score
MSVD QA	Video	Open-ended	Zero-shot	Test	Accuracy
MSRVTT QA	Video	Open-ended	Zero-shot	Test	Accuracy
SEED Bench	Video & Single image	Multiple-choice	Zero-shot	-	Accuracy
VQAv2	Single-image	Open-ended	Zero- & few-shot	Test	Soft accuracy
A-OKVQA	Single-image	Open-ended	Zero- & few-shot	Val	Soft accuracy
ScienceQA-IMG	Single-image	Multiple-choice	Zero-shot CoT	Test	Accuracy
MMBench	Single-image	Multiple-choice	Zero-shot	Dev	Accuracy+
MME	Single-image	Yes/No	Zero-shot	-	Accuracy+

Table 10: An overview evaluation benchmarks and metrics.

et al., 2017), MSRVTT QA (Xu et al., 2017), SEED-Bench (Li et al., 2023c)), DEMON (Li et al., 2023d)), MME (Fu et al., 2023) and MM-Bench (Liu et al., 2023e).

For VQAv2 and A-OKVQA, we test the zero-shot question answering ability given a single image, and also evaluate the ability to gain information from other related images under few-shot ICL setting. ScienceQA is proposed for Chain-of-Thought (CoT) (Wei et al., 2022) scenario, and we adopt the corresponding zero-shot CoT setting. SEED-Bench (Li et al., 2023c) is a recently proposed benchmark that also aims at question answering, which comprises both images and videos. We evaluate the zero-shot ability on it because no training set is available for extracting few-shot examples. As NLVR2 contains naturally interleaved image-text instances, we only conduct zero-shot evaluation. With recently proposed benchmarks for MLLMs, such as MME (Fu et al., 2023), MM-Bench (Liu et al., 2023e), and DEMON (Li et al., 2023d), we employ the zero-shot setting.

Following previous works (Antol et al., 2015; Saikh et al., 2022; Suhr et al., 2019; Huang et al., 2023), we report the soft-accuracy scores (Antol et al., 2015) for A-OKVQA and VQAv2, and calculate the accuracy scores on ScienceQA, NLVR2, MSVD QA and MSRVTT QA. The accuracy+ is employed as the metric for MME, and I4-score (Li et al., 2023d) is used for DEMON-Core.

F Full results of other popular MLLMs

As we evalution the performance on a variaty of task, where some results are missing for certain closel related baselines. We use the public checkpoints to obtain the missing results for MMICL, InstructBLIP and Otter [5], together with official scripts and required environments. Apart from the MLLMs that focus on interleaved and instruction-

[5]The detailed model versions with links are as follows: MMICL (https://huggingface.co/BleachNick/MMICL-Instructblip-T5-xl and https://huggingface.co/BleachNick/MMICL-Instructblip-T5-xxl); InstructBLIP (https://huggingface.co/Salesforce/instructblip-flan-t5-xl and https://huggingface.co/Salesforce/instructblip-flan-t5-xxl); Otter: (https://huggingface.co/luodian/OTTER-Image-MPT7B).

Model	LLM Backbone	#Param LLM	In-context Learning		Multi-image / Video Tasks				
			VQAv2	A-OKVQA	NLVR2	DEMON	SEED	MSVD QA	MSRVTT QA
Models SMALLER than 10B									
KOSMOS-1	MAGNETO	1.3B	51.80	-	-	-	-	-	-
KOSMOS-2	MAGNETO	1.3B	-	-	-	-	50.00	-	-
InstructBLIP-XL	FlanT5	3B	31.76*	39.13*	52.59*	32.59*	52.70	43.40*	12.12*
MMICL-XL$^{\diamond}$	FlanT5	3B	69.16	53.43*	71.48*	**38.14***	54.69*	53.68	42.36*
Otter	MPT	7B	45.39*	38.42*	49.54*	24.51	39.70	25.87*	9.78*
VPG-C-LLaMA2	LLaMA	7B	-	34.29*	53.82*	37.22	-	6.03*	-
Flamingo-9B	Chinchilla	7B	56.3	-	-	-	-	30.2	13.7
Brote-EX-XL(ours)	FlanT5	3B	**69.97**	56.00	71.41	37.33	57.51	53.02	43.14
Brote-IM-XL(ours)	FlanT5	3B	68.94	**56.43**	**76.02**	37.34	**57.86**	**56.06**	**45.08**
Models LARGER than 10B									
InstructBLIP-XXL	FlanT5	11B	48.21*	45.92*	64.54*	33.00*	50.81*	44.30*	15.49*
MMICL-XXL$^{\diamond}$	FlanT5	11B	70.56	54.85*	56.16*	36.30*	56.66*	52.19	39.46*
VPG-C-Vicuna	Vicuna	13B	-	-	-	36.37	-	-	-
BLIP-2-13B	Vicuna	13B	-	-	-	-	46.4	20.3	10.3
InstructBLIP-13B	Vicuna	13B	-	-	-	-	-	41.2	24.8
EMU-I	LLaMA	13B	58.4	-	-	-	-	37.0	21.2
EMU-2	LLaMA	33B	67.0	-	-	-	**62.8**	49.0	31.4
Flamingo-80B	Chinchilla	70B	63.1	-	-	-	-	35.6	17.4
Brote-EX-XXL(ours)	FlanT5	11B	70.86	59.94	70.42	38.70	59.31	54.52	45.24
Brote-IM-XXL(ours)	FlanT5	11B	**71.71**	**60.31**	**80.71**	**38.94**	61.64	**57.29**	**45.94**

Table 11: Results for multi-image settings. The best results for models larger/smaller than 10B are separately **bolded** and the seconds are underlined. $^{\diamond}$: the InstructBLIP version. We evaluate results which are not officially announced using public checkpoints and mark them by *. SEED refers to SEED-Bench that contains both images and videos.

following settings, we also provide a table of results from other popular MLLMs. Table 11 and Table 12 record the results for multi-image and single-image settings respectively. The detailed results of the subtasks from DEMON-core are listed in Table 13.

In Table 11 and Table 13, our models demonstrate better performance over the others of different scales, including models of larger scales. We outperform strong baselines, such as InstructBLIP, MMICL and VPG-C, which include also consider prior-LLM instruction-image fusion. This support our finding that our browse-and-concentrate paradigm contributes to a more in-depth understanding of multimodal context with the assistant of these intermediate browsing insights.

However, for single-image tasks reported in Table 12, we notice a different trend on MME benchmark. For models with LLMs small than 10B, VPG-C-Vicuna and Otter show impressive performance on MME. For models with LLMs larger than 10B, MMICL-XXL (BLIP2) presents the best performance, followed by its variant MMICL-XXL (InstructBLIP). Our models only outperform InstructBLIP models. This is potentially caused by the limitaion of training data, where we exclude the visual instruction tuning dataset such as LLaVA (Liu et al., 2023d) during pre-training and fine-tuning, because the outputs can vary subjectively. On the contrary, our models continue to manifest progress for single-image QA tasks and the other MLLM benchmark MMBench.

G Details for Ablation Study on the Training Strategies

In this section, we provide the detailed results for ablation study of our proposed strategies as an accompany of Table 4. Table 14 lists the results for each tasks, and averaged scores for multi-image tasks (AVG-Multi), single-image tasks (AVG-Single) and overall tasks (AVG). In the settings without context-dropping strategies,, our model with pre-training presents superior multi-image comprehension, as evidenced by its performance on the A-OKVQA 4-shot and SEED-video settings, in comparison to its counterpart without pre-training. Nonetheless, without context-dropping strategies, both models exhibit a limitation in achieving a balanced performance across single-image and multi-image scenarios. To address this, we incorporate context-dropping strategies designed to encourage the models to effectively utilize the given condition context vector, as detailed in Section 3.4. We eventually adopt the "Drop-ALL" setting for training our Brote models.

Model	LLM Backbone	#Param LLM	VQAv2 0-shot	A-OKVQA 0-shot	ScienceQA -IMG	MME Perception	MME Cognition	MMBench
Models SMALLER than 10B								
KOSMOS-1	MAGNETO	1.3B	51.80	-	-	-	-	-
InstructBLIP-XL	FlanT5	3B	36.77	**54.57**	70.40	1093.70*	281.43*	69.68*
MMICL-XL	FlanT5	3B	69.13	52.12*	**72.58***	1184.54*	277.86*	73.11*
LLaVA†	LLaMA	7B	-	-	-	457.82	214.64	36.2
Otter†	MPT	7B	57.89*	41.92*	63.10	1292.26	306.43	48.3
VPG-C-Vicuna	Vicuna	7B	-	-	-	**1299.24**	**321.07**	-
Brote-EX-XL(ours)	FlanT5	3B	69.90	52.93	71.15	1203.87	301.79	73.27
Brote-IM-XL(ours)	FlanT5	3B	**70.24**	53.40	**72.58**	1181.95	266.79	**74.29**
Models LARGER than 10B								
InstructBLIP-XXL	FlanT5	11B	63.69	57.10	70.60	1212.82*	291.79*	70.34*
JiuTian†	FlanT5	11B	-	-	-	-	-	64.7
MMICL-XXL	FlanT5	11B	70.30	51.35*	74.92*	1313.88*	311.79*	76.58*
MMICL-XXL (BLIP2)†	FlanT5	11B	69.99	-	-	**1381.74**	**428.93**	65.24
Brote-EX-XXL(ours)	FlanT5	11B	71.58	56.47	77.69	1279.73	310.01	76.67
Brote-IM-XXL(ours)	FlanT5	11B	**73.02**	**57.83**	**78.38**	1284.13	300.00	**77.34**

Table 12: Zero-shot results for single-image settings. The best results for models larger/smaller than 10B are separately **bolded** and the seconds are underlined. †: results of these models are taken from Zhao et al. (2024). We evaluate results which are not officially announced using public checkpoints and mark them by *.

Model	#Param LLM	Multimodal Dialogue	Visual Storytelling	Visual Rel. Inference	Multimodal Cloze	Knowledge QA	Text-rich Images QA	Multi-image Reasoning	AVG
MiniGPT-4†	7B	13.69	17.07	7.95	16.60	30.27	26.40	43.50	22.21
Otter†	7B	15.37	15.57	11.39	16.00	41.67	27.73	43.85	24.51
BLIP-2-XXL†	11B	26.12	21.31	10.67	17.94	39.23	33.53	39.65	26.92
InstructBLIP-XL◇	3B	19.42	25.09	15.21	32.35	48.13	38.89	49.04	32.59
InstructBLIP-XXL†	11B	33.58	24.41	11.49	21.20	47.40	44.40	48.55	33.00
MMICL-XXL◇	11B	31.60	**28.76**	12.17	31.86	61.58	44.33	43.73	36.30
VPG-C-Vicuna†	13B	37.50	25.20	**25.90**	22.15	48.60	49.93	50.28	36.37
VPG-C-LLaMA2†	7B	**42.70**	24.76	25.50	22.95	51.00	44.93	48.68	37.22
MMICL-XL◇	3B	33.32	27.14	13.58	**34.17**	58.45	47.19	**53.10**	38.14
Brote-XL	3B	32.46	27.38	10.51	32.41	59.45	48.07	51.08	37.34
Brote-XXL	11B	34.95	28.23	11.11	29.51	**65.25**	**50.87**	52.65	**38.94**

Table 13: Evaluation on DEMON-Core benchmark. Models marked by †: results taken from Li et al. (2023d). Models marked by ◇: we evaluate the results with official checkpoints as stated in Appendix F.

Models	Pre -train	Fine -tune	Drop	A-OKVQA 0shot	A-OKVQA 4shot	SEED Image	SEED Video	NLVR2	SciQA -IMG	MSVD	AVG -Multi	AVG -Single	AVG
InstructBLIP	-	-	-	54.57	39.13	-	-	52.59	70.40	43.40	47.05	57.85	51.68
MMICL	-	-	-	51.53	53.32	58.81	35.40	62.40	63.21	37.07	47.05	57.85	51.68
Ours-sampled	✘	✓	✘	**53.15**	54.76	**60.38**	33.76	63.35	63.36	35.87	46.51	**58.96**	51.85
Ours-sampled	✓	✓	✘	49.94	56.16	57.57	36.85	60.58	60.58	34.17	47.12	56.03	50.94
Ours-sampled	✓	✓	IMG-N	50.39	54.79	60.16	37.31	64.62	61.87	35.51	48.06	57.47	52.09
Ours-sampled	✓	✓	IMG-B	48.53	55.22	59.21	37.57	65.00	**63.31**	34.36	48.06	57.02	51.90
Ours-sampled	✓	✓	TXT	50.14	54.51	59.14	36.90	61.56	62.92	**39.36**	48.08	57.40	52.08
Ours-sampled	✓	✓	All	48.22	**55.35**	59.86	**37.64**	**65.15**	63.16	37.35	**48.87**	57.08	**52.39**

Table 14: Ablation study of different training strategies on XL-sized (3B LLM) models trained with sampled subset, where "Ours" refers to *Our-sampled* described in §4.4. "AVG-Multi" is the averaged over A-OKVQA 4-shot, SEED image split, NLVR2 and MSVD, and "AVG-Single" is the averaged over the rest. "AVG" refers to the average accuracy of all the tasks in this table.