

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/187915/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Yu, Jiazuo, Zhuge, Yunzhi, Zhang, Lu, Huang, Zichen, Zhou, Wei, Wang, Dong, Lu, Huchuan, He, You and Chen, Long 2026. One aligned LLM to serve them all: A transfer recipe for training VLMs without visual-language re-alignment. International Journal of Computer Vision 134 (7) , 331. 10.1007/s11263-026-02930-z

Publishers page: <https://doi.org/10.1007/s11263-026-02930-z>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.





One Aligned LLM to Serve Them All: A Transfer Recipe for Training VLMs without Visual-Language Re-Alignment

Jiazuo Yu¹ · Yunzhi Zhuge¹ · Lu Zhang¹  · Zichen Huang¹ · Wei Zhou² · Dong Wang¹ · Huchuan Lu¹ · You He³ · Long Chen⁴

Received: 14 August 2025 / Accepted: 15 June 2026 / Published online: 1 July 2026

Abstract

Vision Language Models (VLMs) with different vision encoders, trained under the two-stage paradigm of pre-training and fine-tuning, have shown varying strengths in different Visual Question Answering (VQA) tasks. However, it remains unclear whether an LLM that has already been aligned during vision-language pre-training can be reused without re-alignment when combined with a different vision encoder for a specific task. To investigate this, we systematically explore VLMs with five different vision encoders using various combinations of LLMs and pre-training strategies. Specifically, we first demonstrate that the aligned LLM with a general-purpose vision encoder can effectively enhance downstream VQA performance with task-specific encoders. Secondly, we investigate several alignment strategies between the aligned LLM and new task-specific encoders. These include (i) feature distillation from the projector layer using both general and task-specific encoders, (ii) a two-stage training strategy with varying the proportion of pre-training data. We find the aligned LLM has acquired transferable vision-language alignment capabilities, such that when combined with new encoders, it no longer requires additional alignment strategy. Evaluated across 13 task metrics after transfer learning with 5 different vision encoders, this new training recipe reduces pre-training time by 2 to 9 hours while achieving comparable or even superior performance.

Keywords Vision encoders · Visual question answering · Vision-language models · Transfer learning

1 Introduction

With recent advances in artificial intelligence (AI), Large Language Models (LLMs) (Brown et al., 2020; Chiang et al., 2023; Raffel et al., 2020; Touvron et al., 2023) have demonstrated remarkable capacities in language understanding and reasoning, achieving unprecedented performance (Wang et al., 2025) across a wide range of natural language processing (NLP) tasks (Zhang et al., 2025). Their emergent abilities, such as in-context learning, compositional reasoning, and zero-shot generalization, have reshaped the landscape of machine learning and inspired a paradigm shift towards

Communicated by Hong Liu.

✉ Lu Zhang
zhangluu@dlut.edu.cn

Jiazuo Yu
yujiazuo@mail.dlut.edu.cn

Yunzhi Zhuge
zgyz@dlut.edu.cn

Zichen Huang
huangzichen@mail.dlut.edu.cn

Wei Zhou
zhouw26@cardiff.ac.uk

Dong Wang
wdice@dlut.edu.cn

Huchuan Lu
lhchuan@dlut.edu.cn

You He
heyu@tsinghua.edu.cn

Long Chen
longchen@ust.hk

¹ School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China

² School of Computer Science and Informatics, Cardiff University, Cardiff CF244AG, United Kingdom

³ Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

⁴ Hong Kong University of Science and Technology, Hong Kong, China

instruction-following and few-shot adaptation. Building on this success, researchers have extended the capabilities of LLMs beyond the text modality, giving rise to VLMs (Diao et al., 2023; Li et al., 2025; Pei et al., 2025; Truong et al., 2025; Wang et al., 2024; Zhang et al., 2025) that integrate visual, auditory, or spatial inputs with natural language for joint reasoning. Among these, visual question answering (VQA) has attracted significant attention due to its potential to handle real-world, perception-grounded tasks. A typical training framework for VQA comprises a visual encoder for extracting semantic-rich features from images, a frozen or partially tunable LLM for language reasoning, and a modality alignment module—often a lightweight projection network—that bridges visual representations with the LLM’s token space. This modular design leverages the strengths of each component while minimizing the cost of multi-modal adaptation. By adopting a two-stage pipeline that combines large-scale pretraining on noisy image-text pairs with instruction tuning on curated multi-modal datasets, vision-augmented LLMs have shown strong generalization across diverse downstream tasks, including object detection (He et al., 2024; Zang et al., 2025), visual grounding (Song et al., 2023; Yang et al., 2023), image captioning (Dai et al., 2024; Li et al., 2023), and image segmentation (Lai et al., 2024; Shi et al., 2025; Wang & Ke, 2024). These advances mark a crucial step toward general-purpose multi-modal intelligence.

With the growing diversity of vision encoders, recent studies (Zong et al., 2024) have demonstrated that encoder choice plays a crucial role in determining performance across various VQA tasks. We further validate this phenomenon in VLMs based on Qwen2-1.5B (2024) and LLaMA3.2-1B (Touvron et al., 2023). For instance, CLIP, as a general-purpose encoder, performs well on average across general VQA tasks, whereas a specialized encoder such as SAM (Kirillov et al., 2023) is better suited for referring expression segmentation and detection (see Fig. 1(a)). However, building a new encoder-specific VLM from scratch to improve performance on each specialized VQA task by following the conventional training paradigm faces several limitations (see Fig. 1(b)): (i) The training processes based on general-purpose and task-specific encoders are isolated from each other, which leads to a lack of transferability in the aligned language model to leverage previously learned visual knowledge. (ii) The training efficiency of the model is significantly hindered due to the huge overhead brought by the pre-training for re-alignment and fine-tuning stages. Therefore, some studies (Tong et al., 2024; Zong et al., 2024) have explored unified multimodal language understanding models that are based on multiple vision encoders and multi-task datasets. Among these, MoVA (Zong et al., 2024) proposes an LLM-based Mixture-of-Experts method, adaptively routing and fusing task-specific vision experts with a coarse-to-fine mechanism, achieving state-of-the-art (SOTA) performance

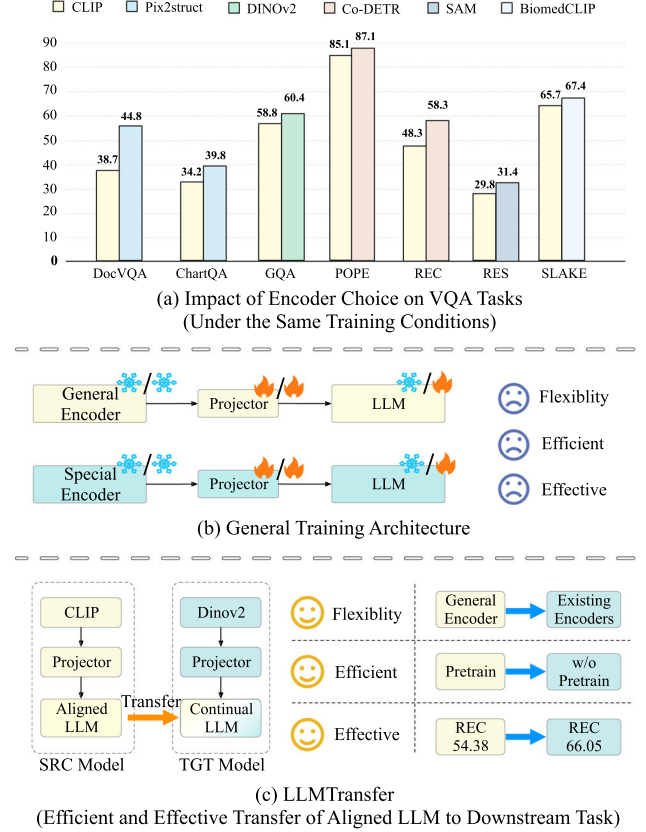


Fig. 1 (a) The results show that VLMs with special encoders can enhance performance across downstream VQA tasks. (b) The conventional two-stage training recipe fails to transfer the knowledge already learned by the aligned LLM when switching between different vision encoders, leading to redundant alignment and computational cost. (c) The continual transfer recipe (LLMTransfer) illustrates how an already aligned LLM can be reused for downstream tasks across different task-specific encoders. SRC Model and TGT Model refer to the source VLMs and the task-specific encoder with the transferred LLM, respectively. The transfer recipe can flexibly transfer aligned LLM across different vision encoders and improve the training efficiency and performance on downstream VQA tasks

across multiple tasks. In addition, Cambrain-1 (Tong et al., 2024) combines multiple vision encoders, including Self-Supervised Learning (SSL) (Oquab et al., 2023) models, which enhances VQA performance across various benchmarks, particularly in vision-centric tasks. However, these methods involve complex training procedures and strategies, as well as increased inference overhead. For example, MoVA introduces additional inference burden due to the use of dual LLMs and MoE parameters, while CamBrain-1 lacks the flexibility to transfer to new tasks such as medical image understanding. Therefore, it’s then natural to ask whether we can leverage the knowledge learned by the aligned LLM to directly promote the learning of new tasks on the task-special encoder, thereby saving parameter and data overhead while optimizing the performance of downstream tasks.

Hopefully, transferring a pre-trained model (which is the main body of trainable parts) from an existing VLM to a novel

LLM instead of training from scratch offers a promising solution. VPGTrans (Zhang et al., 2024) investigates the potential of transferring the same encoder across different LLMs, and demonstrates that the knowledge learned in a pre-trained source model can inspire the transfer of vision encoders across different LLMs, achieving competitive performance while saving both time and resources. However, this approach only explores the transfer of the same visual encoder and projection layer across different frozen language models, overlooking the impact of the growing number of vision encoders trained for different tasks on VLMs. Therefore, we systematically investigate the potential of transferring an aligned LLM with a general-purpose encoder (e.g., CLIP) to downstream tasks with specialized encoders, offering an efficient and effective approach for training the task-specific models with aligned LLM (see Fig. 1(c)). Specifically, we treat the VLM trained with the CLIP encoder as a source model and transfer the aligned LLM to other vision encoders, such as the self-supervised model DINOv2 (Oquab et al., 2023) and the fully supervised model SAM (Kirillov et al., 2023). Furthermore, we examine how varying the proportion of vision-language alignment during pretraining affects the performance of the aligned LLM when transferred to downstream tasks. Extensive experiments with more than 5 encoders and 13 metrics suggest several key findings:

- First, we find that effective aligned LLM transfer across different encoders can be achieved under the same training conditions, improving performance on downstream tasks by leveraging the vision-language alignment knowledge of aligned LLM and enhancing the capability of VLMs with new encoders. Experimental results show consistent improvements across multiple evaluation metrics compared to the original VLM (see Sect. 4.3&5.3).
- Second, we find that the aligned LLM inherently possesses the capability to perceive visual information, eliminating the need for additional pretraining alignment with new encoders. Without any extra data or re-alignment steps, it achieves performance comparable to that of re-aligned models, while reducing training time (see Sect. 4.4&5.3).
- Third, the aligned LLM does not require any additional pre-alignment procedures to optimize the projector layer. Instead, a new projector layer connecting the aligned LLM and a novel vision encoder can be effectively learned solely through multi-turn conversational fine-tuning during the downstream training stage (see Sect. 4.5&5.3).

Rather than proposing a new architectural module or transfer learning algorithm, the main contribution of this work is a systematic empirical study of an efficient training recipe for

vision-language models. We demonstrate that once a large language model has been sufficiently aligned with vision-language data using a general-purpose vision encoder, this alignment can be directly reused across diverse task-specific encoders without additional visual-language re-alignment. From this perspective, LLMTransfer is formulated as a practical transfer recipe validated through extensive experiments, showing that redundant pre-alignment can be safely removed while maintaining competitive or improved performance, leading to significant reductions in pre-training time and computational cost.

2 Related Works

Vision Encoders. In recent years, vision encoders have evolved significantly, with a growing diversity of models tailored for specific tasks. Early vision encoders, such as convolutional neural networks (CNNs), laid the foundation for visual feature extraction. However, with the advancement of transformer-based architectures, models like Vision Transformer (ViT) (Dosovitskiy et al., 2020; Vaswani et al., 2017) have emerged, demonstrating superior performance by leveraging self-attention mechanisms to capture long-range dependencies across image pixels. As the field progresses, vision encoders are increasingly being designed to address specific challenges in multimodal and specialized tasks. CLIP (Radford et al., 2021) introduced a vision encoder pre-trained on large-scale image-text pairs, enabling versatile cross-modal understanding for a wide range of tasks. On the other hand, BiomedCLIP (Zhang et al., 2023) tailors vision encoding for biomedical images, illustrating how encoders can be fine-tuned for domain-specific applications. Similarly, Pix2Struct (Lee et al., 2023) focuses on encoding images for structured data extraction, particularly useful for tasks involving graphs or complex layouts such as document and chart understanding. Recent innovations further enhance the versatility of vision encoders, adapting them to rapidly expanding domains. Vision encoders are now routinely designed with task-specific adaptations, enabling them to excel across diverse applications such as object detection, image captioning, visual grounding, and medical image interpretation. Early efforts such as Flamingo (Alayrac et al., 2022) lay the groundwork for few-shot vision-language learning by integrating frozen vision encoders with autoregressive language models, demonstrating strong cross-modal language reasoning with minimal supervision. Building on this foundation, more recent models like LLaVA (Liu et al., 2024) and BLIP series (Li et al., 2024, 2023, 2022) integrate vision encoders with language models, enabling rich vision-language alignment for various tasks like image-text generation and cross-modal reasoning. These advances highlight the increasing complexity and task specificity of vision

encoders, which are rapidly becoming specialized for an ever-expanding range of applications. In parallel, the growing number of vision encoders has spurred research into how these models can be adapted and transferred across different tasks. Therefore, the paper explores the performance of switching between different encoders in VLMs across various different tasks, aiming to build a general framework that efficiently and effectively transfers the aligned language model from trained models to new task-specific encoders for better adaptation to novel tasks.

Visual Question Answering Models. Visual Question Answering (VQA) models extend the capabilities of traditional Large Language Models (LLMs) by incorporating an additional visual modality (Dosovitskiy et al., 2020; Wang et al., 2025), thereby enabling multi-modal understanding and reasoning. With the rise of vision-language pretraining, recent advancements in VLMs have demonstrated remarkable progress in various tasks, including image-text match and comprehension (Li et al., 2024, 2023; Liu et al., 2024), referring expression segmentation (Lai et al., 2024; Wang & Ke, 2024), and chart&document understanding (Han et al., 2023). These models typically consist of a visual encoder, a frozen or trainable LLM, and a vision-language projection module that aligns visual features with the language domain. Despite their success, one of the main challenges in developing VLMs lies in their high computational cost. Training large-scale models requires significant GPU resources, making model development expensive. To address this, recent research explores transferring pretrained components across different models, enabling efficient adaptation to new architectures. VPGTrans (Zhang et al., 2024) investigates vision encoder transferability across different LLMs, while studies like MoVA (Zong et al., 2024) reveal that vision encoder choice significantly impacts VQA performance across diverse tasks. Building on these insights, we take a complementary perspective by examining the new direction—how pretrained LLMs can be effectively transferred across different visual encoders. We investigate a continual training strategy that enables scalable and efficient adaptation of VLMs without retraining from scratch, offering a practical solution for balancing performance and resource constraints.

Model Transfer. Recent advancements in model transfer research have extensively investigated diverse strategies for adapting pre-trained models to novel tasks. Some popular approaches (Lester et al., 2021; Zhou et al., 2022) focus on the optimization of soft task prompts to enhance the performance of pre-trained models on specific tasks. The transferability of these task prompts has been a subject of considerable research interest. For instance, (Su et al., 2021) conducts comprehensive experiments to demonstrate the transferability of prompts across various tasks and models, revealing that transferring prompts between similar tasks significantly

accelerates training and improves performance. Additionally, (Lester et al., 2022) introduces a method for recycling soft prompts across models using vocab-to-vocab transformations or linear-combination transformations. And (Deng et al., 2022) involves prompt tuning on discrete prompts, enabling their direct sharing across different models. Despite these advancements, existing prompt transfer methods often encounter the dimension change problem and suffer from a potential difference between the content prompts and task prompts. To address this issue, VPGTrans (Zhang et al., 2024) proposes a novel framework that facilitates the transfer of a shared vision encoder and projector across various Large Language Models (LLMs), effectively reducing the dimension change problem to a simple projector with negligible parameters. However, this approach fixes the vision encoder and overlooks the potential impact of task-specific encoders on different downstream tasks. In this study, we explore the transfer of task-specific encoders under consistent language model scales and training conditions, aiming to enhance the performance of VQA in downstream tasks. Furthermore, we introduce a continuous training recipe designed to efficiently migrate the visual encoder, thereby optimizing the adaptation process and improving overall model efficiency.

3 Preliminaries

To establish the foundation for our analysis, we review 1) a VLM model architecture, 2) training data, and 3) a training implementation.

3.1 Model Architecture

We utilize the general architecture used by many recent VLMs, such as LLaVA (Liu et al., 2024), whose architecture uses a pre-trained visual backbone to map an input image to a sequence of patch features that are then projected individually into the embedding space of an LLM. Formally, these VLMs take an image x and an instruction I as input, and these inputs are then fed into the following components: 1) a visual encoder, 2) a vision-language projection layer, and 3) a large language model.

The image x is first processed by a *vision encoder* E_θ , transforming it into visual embeddings $e_v = E_\theta(x)$. These embeddings e_v are then projected by a *vision-language projection layer* f_θ as vision-language embeddings $e_{vl} = f_\theta(e_v)$ to align with the language modality. Finally, the e_{vl} are combined with the embedded instruction e_i and fed into the *large language model* LM_θ for vision-language reasoning. The final output text o_t can be calculated by:

$$o_t = LM_\theta([f_\theta(E_\theta(x)); e_i]), \quad (1)$$

and all trainable parameters of θ are optimized via autoregressive loss:

$$\mathcal{L} = - \sum_t \log P(o_t \mid o_{<t}, e_{vl}, e_i; \theta), \quad (2)$$

where $P(o_t \mid o_{<t}, e_{vl}, e_i; \theta)$ represents the probability of generating the token o_t given the previously generated tokens $o_{<t}$, the vision-language embeddings e_{vl} and instruction embedding e_i . This objective ensures that the model learns to generate coherent and contextually relevant textual outputs in an autoregressive manner, aligning the vision and language modalities effectively.

3.2 Training Data

We adopt the publicly available datasets utilized by LLaVA (Liu et al., 2024), LLaVA-NeXT (Liu et al., 2024), and MoVA (Zong et al., 2024) to ensure comparability and reproducibility. Specifically, our pre-training stage leverages LLaVA-558k-Pretrain (Liu et al., 2024) dataset, a diverse collection of samples sourced from multiple captioning datasets, including Conceptual Captions (Sharma et al., 2018) and LAION (Schuhmann et al., 2021). This dataset provides a rich variety of image-text pairs, facilitating robust multi-modal representation learning.

For fine-tuning, the dataset (Zong et al., 2024) incorporates a total of 872K samples drawn from LLaVA-Next (Liu et al., 2024), SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), LLaVA-Med (Li et al., 2024), and RefCOCO (Yu et al., 2016). These datasets encompass a broad spectrum of vision-language tasks, including chart and document comprehension, medical reasoning, object segmentation, and detection, thereby enhancing the model’s adaptability across diverse domains. We maintain a consistent total dataset size across different encoder architectures and training configurations to ensure a rigorous and fair evaluation of our experimental recipes. This consistency eliminates confounding variables, allowing us to isolate and accurately assess the impact of architectural modifications and training strategies on model performance.

4 Exploring Different VLMs Training Strategies

4.1 Motivation

The growing diversity of task-specific vision encoders highlights their unique strengths in multi-modal large language models. For instance, CLIP (Radford et al., 2021) leverages large-scale image-text pairs for contrastive learning, and its pre-trained vision encoder has demonstrated general percep-

tual capabilities (Liu et al., 2024). SAM (Kirillov et al., 2023), trained with class-agnostic masks for universal, zero-shot, and interactive segmentation, has proven effective in Referring Expression Segmentation (RES) tasks (Lai et al., 2024), while Pix2struct (Lee et al., 2023) enhances fine-grained document and chart understanding. Meanwhile, MoVA (Zong et al., 2024) has shown that such specialized encoders exhibit distinct advantages in specific language reasoning tasks when trained on the same dataset. However, the approach of training encoder-specialized models for different domains results in independently learned parameters, preventing the effective transfer of vision-language-aligned LLM knowledge from fine-tuned models. Moreover, each VLM requires a two-stage training process from scratch, leading to substantial computational overhead.

To address these challenges, we conduct an exploratory study focusing on both the interactions among different VLMs and efficient transfer strategies from the following key perspectives: (i) whether VLMs trained with a general-purpose encoder can enhance the performance of specialized encoders in downstream tasks when transferred across different specialized encoders; (ii) how to more efficiently train a VLM to enable effective knowledge transfer across diverse encoder architectures while maintaining competitive performance; and (iii) whether the aligned LLM, without requiring any additional pre-alignment procedures for the projector, can directly adapt to a novel vision encoder by learning a new projector layer solely through multi-turn conversational finetuning during downstream training.

4.2 Build Base Model

As shown in Fig. 2(a)-(e), our method is built upon the Qwen2-1.5B (2024), denoted as LM_θ , which demonstrates flexible transferability of aligned LLM $VisionLM_\theta$ across more than five distinct vision encoders. The detailed exploration of our approach will be presented as follows.

We first build a two-stage training pipeline, as shown in Fig. 2(a), that establishes general vision-language alignment as a foundation for subsequent task processes. In the first stage, we leverage a general-purpose encoder (CLIP) to pre-train a projector layer f_θ , which serves as a crucial modality bridge between visual and textual representations. Building upon this foundation, the second stage involves joint fine-tuning of both the projector f_θ and pre-trained LLM LM_θ to optimize their interaction and enhance cross-modal understanding. Through this process, we obtain a vision-aligned language model, denoted as $VisionLM_\theta$, which interprets visual information through CLIP’s well-established visual representation space while maintaining strong linguistic reasoning abilities.

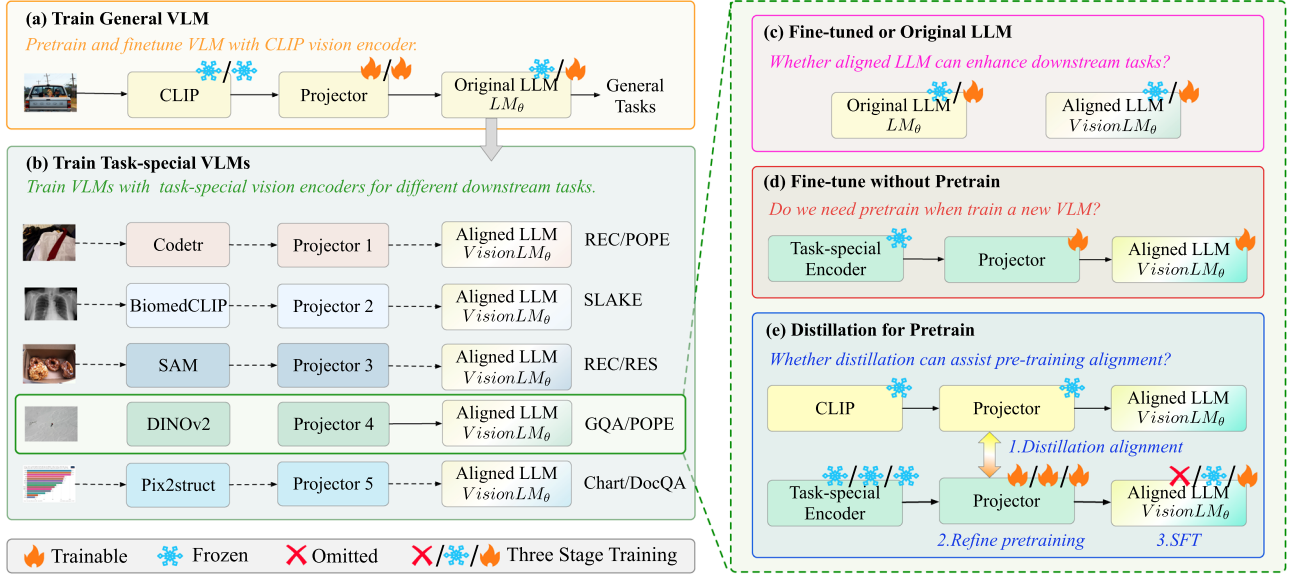


Fig. 2 Exploring different VLMs training strategies. **(a)** We train a general VLM with CLIP vision encoder and original Qwen2-1.5B/LLaMA3.2-1B LM_θ , resulting in a vision-aligned LLM denoted as $VisionLM_\theta$. **(b)** We train task-special models with task-special vision encoders and different LLMs to explore the efficient and effective training strategy. Among them, we first **(c)** use a two-stage training method based on the original LM_θ and the aligned LLM $VisionLM_\theta$ to verify whether continuous training of the aligned LLM can enhance

downstream tasks. After verifying that continuous training is more helpful for downstream tasks, we **(d)** explore whether pre-training is essential when fine-tuning a VLM with an aligned LLM by verifying the impact of different ratios of pre-training on the final performance. Finally, we **(e)** use a three-stage training strategy, including distillation for warming up, pre-training for refinement, and supervised full fine-tuning, to verify whether projection distillation can assist pre-training alignment and thus improve efficiency and performance

4.3 Continually Train Aligned LLM

To investigate the acquired multimodal knowledge with aligned visual representations in $VisionLM_\theta$ can be effectively transferred to optimize new tasks with task-specific encoders, we explore various training strategies to continually train task-specialized VLMs for different downstream tasks, as shown in Fig. 2(b). Firstly, we utilize five different vision encoders to conduct the two-stage training pipeline based on the fine-tuned $VisionLM_\theta$ and the original LLM LM_θ , respectively. Then, we evaluate the performance metrics of these downstream tasks corresponding to each encoder to verify whether the fine-tuned LLM can enhance downstream task performance, as shown in Fig. 2(c). Notably, in this process, we initialize a new projector for each task-specific encoder, which is necessary because different encoders may have varying feature dimensions. Such dimension mapping strategies have been widely adopted and proven effective in various VQA frameworks, including VPGTrans (Zhang et al., 2024), PathWeave Yu et al., and X-InstructBLIP (Panagopoulou et al., 2023), without compromising model performance.

4.4 Fine-Tune Without Pre-Train

Building on the above findings, we hypothesize that the language model of the already trained VLM has achieved visual-language alignment. Therefore, we reduce pre-training to varying degrees or completely remove the pre-training to explore how to efficiently perform continual fine-tuning, achieving performance comparable to traditional two-stage training. Finally, we find a generally transferable continual LLM strategy, LLMTransfer, which can efficiently transfer the vision-aligned language model $VisionLM_\theta$ to different encoders only with one-stage training (as shown in Fig. 2(d)). This pre-training free transfer strategy across different encoders significantly reduces computational and temporal overhead while maintaining competitive performance on downstream tasks. The effectiveness of the training pipeline supports our finding that the aligned LLM inherently possesses the capability to perceive visual information, thus eliminating the need for additional pretraining alignment with new visual encoders.

4.5 Distillation for Pre-Training

To further explore whether the aligned LLM based on CLIP can achieve seamless integration of visual and textual understanding, we record the task performance metrics of each

specialized encoder in its respective domain and attempt to further warm up the projector by assisting the alignment stage with distillation loss. As shown in Fig. 2(e), we adopt a three-stage training approach, which includes: dividing the original pre-training stage into distillation and pre-training for refinement, followed by full supervised fine-tuning. Among them, the distillation stage aims to leverage the pre-aligned vision and language model projector to help the new vision encoder quickly align with the data distribution of the language model. We explore two distinct distillation losses to train the new projector, which can be expressed by:

$$\begin{aligned}\mathcal{L}_{\text{MSE}} &= \sum_{i,j} \left(f_1^{(i,j)} - f_2^{(i,j)} \right)^2, \\ \mathcal{L}_{\text{KL}} &= \sum_{i,j} f_1^{(i,j)} \log \frac{f_1^{(i,j)}}{f_2^{(i,j)}},\end{aligned}\tag{3}$$

where $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{KL}}$ represent the Mean Squared Error (MSE) and the Kullback-Leibler (KL) divergence loss functions, respectively. f_1 refers to the target features, which come from a pre-aligned vision-language model. f_2 refers to the predicted features output by the new projector, which is trained to align with the target feature distribution for adapting the target LLM. Additionally, the extra pre-training phase is designed to refine the alignment stage. Extensive experimental results indicate that distillation does not effectively facilitate feature alignment. This further supports our finding that an aligned LLM can seamlessly adapt to a new vision encoder without requiring explicit pre-alignment of the projector layer, as effective feature alignment can be achieved solely through multi-turn conversational fine-tuning during downstream training.

5 Experiments

5.1 VQA Benchmarks

We empirically analyze the multimodal capacity and generalization ability of LLMs transferred across different vision encoders on a wide range of challenging VQA benchmarks, including DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), GQA (Hudson & Manning, 2019), POPE (Li et al., 2023), REC (Yu et al., 2016), RES (Yu et al., 2016), and SLAKE (Liu et al., 2021). Additionally, Table 1 summarizes the task categories and their corresponding inference templates, which span five major types: Chart&Document Understanding, Scene Understanding, Object Hallucination, Grounding and Localization, and Biomedical Understanding.

5.2 Implementation Details

As mentioned in Sect. 3.1, our training baseline consists of two stages. In the pretraining stage, we use the AdamW optimizer with an initial learning rate of 2×10^{-4} , a batch size of 256, and train the model for 1 epoch. We finetune the weights of the projection layer and LLMs with a batch size of 128 and an initial learning rate of 2×10^{-5} during supervised finetuning. For LoRA-based training, we set the rank to 64. All the models are trained using 4 A800 80G GPUs, with gradient accumulation steps set to 4. All the results are tested using 1 A800 80G GPUs. The detailed configurations of the task-specific vision encoders are shown in Table 2. We adopt the official checkpoint weights that are publicly available.

5.3 Quantitative Analysis

5.3.1 Codetr Encoder

We begin with a systematic comparison between a general CLIP-based VLM and Codetr-based VLMs, using consistent training data and protocols. As shown in Table 3, the Codetr-based VLM achieves noticeably better results in downstream metrics such as POPE and RES, confirming the benefit of using task-specialized vision encoders. To demonstrate the impact of LLM alignment, we compare two VLMs that share the same task-specific vision encoder (Codetr) but differ in the alignment status of their language model. Specifically, we evaluate the original LLM (LM_θ) against an aligned LLM ($VisionLM_\theta$) that has been pre-aligned with a different visual encoder. Experimental results show that $VisionLM_\theta$ consistently outperforms LM_θ , achieving a 0.6 improvement on the POPE benchmark and a 1.6 gain on $RES_{refcoco_val}$. These findings highlight the effectiveness and transferability of vision-language alignment, suggesting that once aligned, the LLM can generalize better across different vision encoders in downstream tasks.

We further investigate whether projector-layer distillation can serve as an effective replacement for pre-alignment. Experiments applying MSE and KL divergence-based distillation from the pre-aligned model to the new projector result in performance degradation compared to directly fine-tuning with the aligned LLM. These findings indicate that such distillation is insufficient for establishing strong cross-modal alignment. To assess the necessity of pre-training for the aligned LLM, we vary the proportion of pre-training data used (from 80% down to 0%). Surprisingly, the performance remains stable even with significantly reduced or entirely skipped pre-training. Notably, the configuration labeled LLMTransfer, in which no pre-training is conducted and only fine-tuning is applied with the aligned LLM. LLMTransfer recipe achieves competitive performance across most metrics, with consistent gains on REC and broadly com-

Table 1 Inference prompt template of different task metrics. The `<|im_start|>` and `<|im_end|>` placeholders represent structuring the prompt into distinct role-based segments. The `<expr>`

placeholder represents the expression in the REC and RES task. The `<image>` placeholder denotes an image is included in the input and the `<question>` is the question of Q&A in the inference datasets

Task	Dataset	Task prompt
Chart&Doc Understanding	ChartQA Masry et al. (2022)	<code>< im_start > system \n You are a helpful assistant.</code> <code>< im_end ></code> <code>\n < im_start > user \n <image> \n <question></code> <code>< im_end ></code> <code>\n < im_start > assistant \n</code>
	DocVQA Mathew et al. (2021)	<code>< im_start > system \n You are a helpful assistant.</code> <code>< im_end ></code> <code>\n < im_start > user \n <image> \n <question></code> <code>< im_end ></code> <code>\n < im_start > assistant \n</code>
Scene Understanding	GQA Hudson and Manning (2019)	<code>< im_start > system \n You are a helpful assistant.</code> <code>< im_end ></code> <code>\n < im_start > user \n <image> \n <question></code> <code>< im_end ></code> <code>\n < im_start > assistant \n</code>
Grounding and Localization	RES Yu et al. (2016)	<code>< im_start > system \n You are a helpful assistant.</code> <code>< im_end ></code> <code>\n < im_start > user \n <image> \n Please segment the re-</code> <code>gion this sentence describes: <expr> Give me the polygon</code> <code>mask coordinates. < im_end > \n < im_start > assis-</code> <code>tant \n</code>
	REC Yu et al. (2016)	<code>< im_start > system \n You are a helpful assistant.</code> <code>< im_end ></code> <code>\n < im_start > user \n <image> \n Provide the</code> <code>bounding</code> <code>box coordinates of the region that the given sentence</code> <code>describes. \n <expr> < im_end > \n < im_start ></code> <code>assistant \n</code>
Biomedical Understanding	Slake Liu et al. (2021)	<code>< im_start > system \n You are a helpful assistant.</code> <code>< im_end ></code> <code>\n < im_start > user \n <image> \n <question> \n</code> <code>Answer</code> <code>the question using a single word or phrase. < im_end ></code> <code>\n</code> <code>< im_start > assistant \n</code>
Object Hallucination	POPE Li et al. (2023)	<code>< im_start > system \n You are a helpful assistant.</code> <code>< im_end ></code> <code>\n < im_start > user \n <image> \n <question></code> <code>< im_end > \n</code> <code>< im_start > assistant \n</code>

Table 2 Vision expert model configurations of different vision encoders. Methods with * use a convolution layer to compress the output feature

Model	Params	Resolution	Width	Depth	Output shape
DINOv2-giant Oquab et al. (2023)	1.1B	518×518	1536	40	1536×37×37
Codetr-large* Zong et al. (2023)	304M	1280×1280	1024	24	256×80×80
SAM-huge* Kirillov et al. (2023)	632M	1024×1024	1280	32	256×64×64
Pix2Struct-large Lee et al. (2023)	513M	720×720	1536	18	1536×45×45
BiomedCLIP-base Zhang et al. (2023)	86M	224×224	768	12	768×16×16

Table 3 Comparison between the Codetr-based VLMs with different training recipes and the CLIP-based VLM with the general training recipe. Dis MSE and Dis KL denote projector-layer distillation via MSE loss and KL divergence, respectively. We label the best and second methods with **bold** and underline styles. The specific experimental settings are detailed in Sect. 5.2. *ref*+, *ref*-, and *refg* denote the RefCOCO, RefCOCO+, and RefCOCOg datasets, respectively

Model	Pretrain	FT	POPE	RES _{ref_val}	RES _{ref_testA}	REC _{ref+_val}	REC _{ref+_testA}
(a) <i>Trainable Qwen LLM with MLP Projector</i>							
CLIP+Original LLM	2h26m	15h28m	85.14	26.40/27.01	30.79/29.81	40.35	48.30
Codetr+Original LLM	9h16m	19h13m	87.10	29.50/29.98	33.81/33.85	46.93	52.86
Codetr+Aligned LLM	9h16m	19h13m	87.73	<u>31.15/31.50</u>	34.82/34.36	<u>51.52</u>	57.50
Codetr+Aligned LLM+Dis MSE	3h54m	19h13m	87.22	30.37/31.08	34.74/34.57	50.67	56.35
Codetr+Aligned LLM+Dis KL	2h26m	19h30m	87.22	30.26/31.13	35.04/34.40	50.34	55.75
Codetr+Aligned LLM (80% Pretrain Data)	7h15m	19h13m	<u>87.71</u>	30.81/31.22	<u>35.07/34.98</u>	51.38	<u>57.68</u>
Codetr+Aligned LLM (60% Pretrain Data)	6h36m	19h13m	87.38	31.37/31.69	34.37/34.46	51.40	57.34
Codetr+Aligned LLM (40% Pretrain Data)	4h6m	19h13m	87.55	31.11/31.46	35.12/34.80	51.24	56.96
Codetr+Aligned LLM (20% Pretrain Data)	2h2m	19h13m	87.45	30.17/31.33	34.99/34.72	51.26	57.11
Codetr+LLMTransfer	0h	19h13m	87.81	30.76/31.25	34.62/34.57	51.89	58.28
(b) <i>Trainable LLaMa LLM with MLP Projector</i>							
CLIP+Original LLM	1h47m	15h14m	80.30	20.05/21.24	22.53/23.29	21.13	28.79
Codetr+Original LLM	4h59m	17h34m	86.75	20.17/21.21	22.26/23.57	21.49	26.08
Codetr+LLMTransfer	0h	16h39m	87.30	24.03/24.51	27.91/27.33	34.40	38.27
(c) <i>Trainable Qwen LLM with QFormer Projector</i>							
CLIP+Original LLM	2h34m	20h47m	64.08	14.31/14.87	15.26/14.98	6.20	7.00
Codetr+Original LLM	5h20m	20h32m	78.93	16.46/17.66	17.26/17.95	9.30	10.51
Codetr+LLMTransfer	0h	21h41m	80.68	16.41/17.78	17.40/18.31	9.65	11.18
(d) <i>The performance of different FT time, with trainable or frozen Qwen LLM</i>							
Codetr+Original LLM (Frozen LLM)	9h16m	16h10m	81.39	12.01/14.06	13.46/15.68	26.69	30.74
Codetr+LLMTransfer (Frozen LLM)	0h	15h54m	85.77	22.35/23.00	25.91/25.37	27.39	31.94
Codetr+LLMTransfer (30% FT Data)	0h	5h41m	87.13	26.21/26.82	30.27/30.15	42.65	48.28
Codetr+LLMTransfer (60% FT Data)	0h	11h27m	87.66	28.40/28.78	32.45/32.14	48.16	53.85
(e) <i>Frozen Qwen LLM with trainable MLP projector and LoRA adapters</i>							
CLIP+Original LLM	2h26m	13h55m	81.09	18.72/21.97	18.49/22.01	36.58	37.83
Codetr+Original LLM	9h16m	18h03m	83.33	23.52/24.19	24.58/25.48	39.87	38.39
Codetr+LLMTransfer	0h	18h11m	84.41	24.24/26.26	27.56/27.34	39.55	38.65

parable performance on RES relative to the Codetr+Aligned LLM (40% pretraining data) recipe, where the performance difference remains within one point, while reducing the training cost by more than 4 hours. This underscores that once the LLM has been pre-aligned with a general visual encoder, it can be reused for new visual backbones without requiring further visual-language alignment procedures.

5.3.2 Dinov2 Encoder

Furthermore, we compare the CLIP-based general VLM with the Dinov2-based VLM and evaluate the GQA, POPE, and REC metrics under the same experimental setup. As shown in Table 4, the results demonstrate that employing Dinov2 as a task-specific visual encoder also significantly enhances downstream task performance, further validating the effectiveness of task-specific encoders in transfer learning. Additionally, by comparing the original LLM with the vision-aligned LLM, we found that the pre-aligned language model continues to exhibit superior performance in downstream tasks, with a 1.6% gain on GQA and over 9% improvement on REC.

To further investigate whether an aligned LLM still requires additional distillation strategies to warm up a new projector when paired with a different visual encoder, we conduct experiments with a distillation-based alignment strategy using the pre-trained projector of the VLM. The results show that while distillation can indeed shorten the training process and retain competitive performance on REC, it still incurs additional cost and leads to a performance drop of approximately 1.0 point on GQA.

In contrast, we investigate the impact of different pre-training ratios on the final fine-tuning outcomes when using the DINOv2 encoder. Experimental results reveal that omitting pre-training has minimal impact on the downstream performance of the DINOv2-based VLM, compared to the about 0.5–2 hours required for pre-training. This further substantiates that the aligned language model has already acquired visual grounding capabilities, enabling it to be directly transferred to a new vision encoder without additional pre-training, thereby accelerating adaptation to downstream tasks.

5.3.3 SAM Encoder

We evaluated the performance of VLMs with a general-purpose encoder and the segmentation-specialized SAM encoder on Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES) metrics. The results are presented in Table 5. For the direct integration with the raw language model, experiments revealed that using the SAM encoder generally improved these metrics. When we employed the aligned LLM to continually train, which is pre-

trained based on the CLIP encoder under the same training conditions, we observed an increase of approximately 3 percentage points in the RES metric and 6-8 percentage points in the REC metric. This demonstrates that continual LLMs can also enhance downstream task performance when adapted to grounding-related tasks.

Furthermore, as shown in Table 5, the experimental results using different distillation methods for pre-alignment indicate that this approach, which distills the feature distribution of new tasks based on the projection of old task features, fails to enhance the alignment of downstream tasks. Therefore, we conducted additional experiments with varying proportions of data for pre-training to further verify whether continual LLM still requires alignment when transferring to a new visual vision encoder. The results show that the LLMTransfer strategy not only reduces over 6 hours of pre-training time but also achieves comparable performance, indicating that the aligned LLM has already acquired strong vision-language alignment capabilities.

5.3.4 BiomedCLIP Encoder

We further examine whether our key findings generalize to the domain of medical image understanding by replacing the general visual encoder with the domain-specific BiomedCLIP. As shown in Table 6, directly using a raw LLM with BiomedCLIP yields improvement of about 2 points on the SLAKE_{open} benchmark. Furthermore, substituting the raw LLM with a vision-aligned LLM to a substantial gain of over 7.5 points. This result reinforces our first finding: an aligned LLM consistently enhances downstream VQA performance, even when paired with a previously unseen, domain-specific vision encoder.

To assess whether pre-training with the new encoder is necessary for adaptation, we compare different pre-training proportions and find that reducing or completely removing the pre-training phase does not degrade performance. In fact, skipping pretraining altogether and simply fine-tuning with the aligned LLM achieves the best results, supporting our second finding: an aligned LLM can be directly transferred across encoders without requiring additional vision-language pretraining. We also investigate whether distillation-based strategies (e.g., using MSE or KL divergence losses) can effectively warm up the new projector. However, these approaches yield inferior performance compared to directly fine-tuning the projector with the aligned LLM. This confirms our third finding: the aligned LLM provides sufficiently strong vision-language grounding such that the new projector can be trained effectively without any warm-up distillation. Together, these experiments demonstrate that our conclusions hold even in specialized domains like medical imaging, highlighting the robustness and transferability of vision-language alignment once established in the LLM.

Table 4 Comparison between the DINOv2-based VLMs with different training recipes and the CLIP-based VLM with the general training recipe. Dis MSE and Dis KL denote projector-layer distillation via MSE loss and KL divergence, respectively. We label the best and second methods with **bold** and underline styles. The specific experimental settings are detailed in Sect. 5.2. ref_+ , and ref_g denote the RefCOCO, RefCOCO+, and RefCOCOg datasets, respectively

Model	Pretrain	FT	GQA	$REC_{ref_+_{val}}$	$REC_{ref_+_{testA}}$	$REC_{ref_g_{test}}$
(a) <i>Trainable Qwen LLM with MLP Projector</i>						
CLIP+Original LLM	2h26m	15h28m	58.83	40.35	48.30	54.38
DINOv2+Original LLM	2h48m	14h42m	60.36	44.43	52.17	58.27
DINOv2+Aligned LLM	2h48m	15h11m	60.40	50.00	58.18	64.21
DINOv2+Aligned LLM+Dis MSE	1h03m	15h30m	59.41	50.83	58.97	64.46
DINOv2+Aligned LLM+Dis KL	1h01m	16h50m	59.38	<u>51.11</u>	58.77	64.48
DINOv2+Aligned LLM (80% Pretrain Data)	2h16m	14h27m	60.61	49.79	57.90	63.91
DINOv2+Aligned LLM (60% Pretrain Data)	1h40m	14h35m	<u>60.63</u>	50.41	58.65	64.40
DINOv2+Aligned LLM (40% Pretrain Data)	1h08m	15h01m	<u>60.63</u>	49.98	58.07	64.46
DINOv2+Aligned LLM (20% Pretrain Data)	0h33m	14h53m	60.39	50.80	<u>59.54</u>	64.88
DINOv2+LLMTransfer	0h	14h11m	60.79	51.91	60.19	64.92
(b) <i>Trainable LLaMa LLM with MLP Projector</i>						
CLIP+Original LLM	1h47m	15h14m	37.43	21.13	28.79	35.14
DINOv2+Original LLM	2h08m	17h07m	56.07	21.62	26.65	33.85
DINOv2+LLMTransfer	0h	17h09m	56.62	30.90	40.18	47.05
(c) <i>Trainable Qwen LLM with QFormer Projector</i>						
CLIP+Original LLM	2h34m	20h47m	39.76	6.20	7.00	11.20
DINOv2+Original LLM	2h41m	21h26m	39.79	6.16	6.60	11.46
DINOv2+LLMTransfer	0h	20h11m	51.97	10.59	13.12	19.68
(d) <i>The performance of different FT time, with trainable or frozen Qwen LLM</i>						
DINOv2+Original LLM (Frozen LLM)	2h48m	13h24m	38.19	3.49	3.94	3.18
DINOv2+LLMTransfer (Frozen LLM)	0h	13h19m	57.09	37.51	45.57	51.08
DINOv2+LLMTransfer (30% FT Data)	0h	4h53m	57.22	42.69	50.20	56.20
DINOv2+LLMTransfer (60% FT Data)	0h	9h44m	58.03	47.55	54.88	62.25
(e) <i>Frozen Qwen LLM with trainable MLP projector and LoRA adapters</i>						
CLIP+Original LLM	2h26m	13h55m	47.38	36.58	37.83	40.61
DINOv2+Original LLM	2h48m	14h12m	47.68	39.87	38.39	40.86
DINOv2+LLMTransfer	0h	13h56m	47.92	39.55	38.65	43.12

Table 5 Comparison between the SAM-based VLMs with different training recipes and the CLIP-based VLM with the general training recipe. Dis MSE and Dis KL denote projector-layer distillation via MSE loss and KL divergence, respectively. We label the best and second methods with **bold** and underline styles. “-” indicates that the model failed to converge for this metric, resulting in near-zero performance. The specific experimental settings are detailed in Sect. 5.2. *ref*+, *ref*+, and *refg* denote the RefCOCO, RefCOCO+, and RefCOCOg datasets, respectively

Model	Pretrain	FT	RES _{ref_val}	RES _{ref_testA}	REC _{ref+_val}	REC _{ref+_testA}	REC _{ref_testA}	REC _{refg_test}
(a) <i>Trainable Qwen LLM with MLP Projector</i>								
CLIP+Original LLM	2h26m	15h28m	26.40/27.01	30.79/29.81	40.35	48.30	54.38	41.88
SAM+Original LLM	6h47m	20h02m	27.40/27.65	31.60/31.37	39.93	46.23	59.71	44.74
SAM+Aligned LLM	6h47m	19h58h	30.17/30.27	34.78/34.45	47.42	53.22	66.38	<u>52.25</u>
SAM+Aligned LLM+Dis MSE	2h33m	19h12m	28.34/28.56	32.27/31.90	41.22	47.39	63.15	46.35
SAM+Aligned LLM+Dis KL	4h11m	18h54m	27.97/28.21	32.61/32.17	40.12	47.01	62.07	45.50
SAM+Aligned LLM (80% Pretrain Data)	5h26m	19h38m	30.21/30.37	34.42/34.30	47.35	53.86	66.51	51.77
SAM+Aligned LLM (60% Pretrain Data)	4h32m	19h52m	30.22/ 30.55	<u>34.62/34.41</u>	47.35	52.56	<u>67.45</u>	51.81
SAM+Aligned LLM (40% Pretrain Data)	4h25m	19h48m	30.27/30.61	34.59/34.32	47.03	53.32	67.13	51.98
SAM+Aligned LLM (20% Pretrain Data)	2h17m	20h05m	30.06/30.19	34.17/33.47	46.42	52.67	66.07	50.19
SAM+LLMTransfer	0h	19h56m	29.94/30.44	<u>34.62/34.30</u>	47.24	<u>53.84</u>	67.68	52.32
(b) <i>Trainable LLaMA LLM with MLP Projector</i>								
CLIP+Original LLM	1h47m	15h14m	20.05/21.24	22.53/23.29	21.13	28.79	35.14	23.04
SAM+Original LLM	5h39m	16h02m	23.94/24.11	28.52/27.98	30.95	35.98	49.50	33.33
SAM+LLMTransfer	0h	15h58m	24.08/25.21	27.66/27.18	31.06	36.03	48.58	34.88
(c) <i>Trainable Qwen LLM with QFormer Projector</i>								
CLIP+Original LLM	2h34m	20h47m	14.31/14.87	15.26/14.98	6.20	7.00	11.20	6.62
SAM+Original LLM	5h38m	21h34m	15.04/15.90	15.17/15.72	7.43	8.44	13.70	8.07
SAM+LLMTransfer	0h	22h11m	16.13/17.56	17.26/18.15	8.94	9.89	17.01	10.78
(d) <i>The performance of different FT time, with trainable or frozen Qwen LLM</i>								
SAM+Original LLM (Frozen LLM)	6h47m	15h05m	-	-	1.32	1.81	2.04	1.28
SAM+LLMTransfer (Frozen LLM)	0h	15h01m	18.89/19.70	20.73/21.14	18.83	22.64	34.11	23.09
SAM+LLMTransfer (30% FT Data)	0h	5h15m	25.41/29.05	29.35/29.05	36.13	42.30	56.29	41.17
SAM+LLMTransfer (60% FT Data)	0h	10h35m	28.20/28.68	32.30/31.95	43.18	49.10	63.00	47.94
(e) <i>Frozen Qwen LLM with trainable MLP projector and LoRA adapters</i>								
CLIP+Original LLM	2h26m	13h55m	18.72/21.97	18.49/22.01	36.58	37.83	40.61	37.86
SAM+Original LLM	6h47m	18h31m	20.42/22.31	23.42/25.32	37.87	39.39	40.86	39.65
SAM+LLMTransfer	0h	18h04m	22.65/23.86	28.16/30.55	35.55	41.65	43.12	40.39

Table 6 Comparison between the BiomedCLIP-based VLMs with different training precipes and the CLIP-based VLM with general training precipe. Dis MSE and Dis KL denote projector-layer distillation via

MSE loss and KL divergence, respectively. We label the best and second methods with **bold** and underline styles. The specific experimental settings are detailed in Sect. 5.2

Model	Pretrain	FT	SLAKE _{open}	SLAKE _{close}
(a) Trainable Qwen LLM with MLP Projector				
CLIP/ViT-B+Original LLM	2h06m	15h28m	65.68	30.14
BiomedCLIP+Original LLM	2h08m	14h43m	67.42	30.98
BiomedCLIP+Aligned LLM	2h08m	14h29m	<u>73.49</u>	31.70
BiomedCLIP+Dis MSE	1h01m	14h45m	72.02	30.46
BiomedCLIP+Dis KL	1h12m	14h39m	45.37	18.22
BiomedCLIP+Aligned LLM (80% Pretrain Data)	1h16m	14h44m	72.94	32.42
BiomedCLIP+Aligned LLM (60% Pretrain Data)	58m	14h26m	<u>73.49</u>	<u>32.66</u>
BiomedCLIP+Aligned LLM (40% Pretrain Data)	37m	14h52m	73.45	32.06
BiomedCLIP+Aligned LLM (20% Pretrain Data)	19m	14h22m	73.01	30.62
BiomedCLIP+LLMTransfer	0h	14h47m	74.35	32.89
(b) Trainable LLaMA LLM with MLP Projector				
CLIP/ViT-B+Original LLM	1h33m	14h56m	38.23	22.49
BiomedCLIP+Original LLM	1h06m	14h08m	63.05	27.87
BiomedCLIP+LLMTransfer	0h	14h23m	62.14	28.35
(c) Trainable Qwen LLM with QFormer Projector				
CLIP/ViT-B+Original LLM	2h15m	20h01m	43.57	23.92
BiomedCLIP+Original LLM	2h10m	16h57m	53.39	21.89
BiomedCLIP+LLMTransfer	0h	16h42m	60.71	29.78
(d) The performance of different FT time, with trainable or frozen Qwen LLM				
BiomedCLIP+Original LLM (Frozen LLM)	2h08m	11h44m	20.93	22.67
BiomedCLIP+LLMTransfer (Frozen LLM)	0h	10h34m	65.81	32.18
BiomedCLIP+LLMTransfer (30% FT Data)	0h	3h58m	67.40	30.50
BiomedCLIP+LLMTransfer (60% FT Data)	0h	8h7m	71.54	32.30
(e) Frozen Qwen LLM with trainable MLP projector and LoRA adapters				
CLIP/ViT-B+Original LLM	2h06m	13h22m	62.22	28.07
BiomedCLIP+Original LLM	2h08m	14h15m	63.45	29.31
BiomedCLIP+LLMTransfer	0h	13h47m	65.81	32.18

5.3.5 Pix2struct Encoder

In the domain of chart and document understanding, the advantages of replacing the encoder are evident, as shown in rows 1 and 2 of Table 7. When using aligned LLM, the performance on ChartQA and DocVQA improves by approximately 2 and 0.2 percentage points, indicating that the continual fine-tuning approach for LLM remains effective in the chart and document domain. Notably, similar to medical image understanding, the performance of the proposed LLM-Transfer without pre-training (the last row of Table 7) further optimizes the metrics on DocVQA and ChartQA by 0.3 to 2.5 points while saving pre-training time about 8h. This further demonstrates that LLM has already acquired instruction fine-tuning alignment capabilities during continual transfer across different vision encoders, enabling effective and efficient transfer learning without the need for pre-training.

The ablation study on the Pix2struct-based VLMs is presented in Table 7, showing the results of using different distillation strategies and pretraining datasets. In DocVQA (Mathew et al., 2021), the MSE distillation method achieved competitive results, but it still needs over 3 hours of pretraining overhead. In contrast, our proposed LLMTransfer not only reduces training time but also improves final performance, achieving the best trade-off between efficiency and effectiveness. The ablation results demonstrate that our approach provides a more efficient solution by minimizing pretraining time while achieving superior performance across tasks.

5.4 Generalization Analysis

Beyond demonstrating the effectiveness of LLMTransfer under a single default configuration, we further conduct a

Table 7 Comparison between the Pix2struct-based VLMs with different training recipes and the CLIP-based VLM with the general training recipe. Dis MSE and Dis KL denote projector-layer distillation via MSE loss and KL divergence, respectively. We label the best and second methods with **bold** and underline styles. The specific experimental settings are detailed in Sect. 5.2

Model	Pretrain	FT	DocVQA	ChartQA
<i>(a) Trainable Qwen LLM with MLP Projector</i>				
CLIP+Qwen1.5B	2h26m	15h28m	38.68	34.24
Pix2struct+Original LLM	8h09m	20h16m	43.83	39.76
Pix2struct+Aligned LLM	7h49m	20h16m	44.00	41.32
Pix2struct+Dis MSE	3h09m	20h21m	44.02	<u>43.56</u>
Pix2struct+Dis KL	3h01m	20h37m	10.00	07.52
Pix2struct+Aligned LLM (80% Pretrain Data)	5h41m	20h26m	43.99	40.88
Pix2struct+Aligned LLM (60% Pretrain Data)	4h05m	20h34m	<u>44.32</u>	41.60
Pix2struct+Aligned LLM (40% Pretrain Data)	2h45m	20h35m	44.20	42.52
Pix2struct+Aligned LLM (20% Pretrain Data)	1h35m	20h22m	44.27	43.09
Pix2struct+ LLMTransfer	0h	20h16m	44.34	43.80
<i>(b) Trainable LLaMA LLM with MLP Projector</i>				
CLIP+Original LLM	1h47m	15h14m	19.10	24.08
Pix2struct+Original LLM	6h09m	17h55m	31.37	32.56
Pix2struct+ LLMTransfer	0h	17h59m	30.92	32.76
<i>(c) Trainable Qwen LLM with QFormer Projector</i>				
CLIP+Original LLM	2h34m	20h47m	11.20	12.84
Pix2struct+Original LLM	6h46m	23h06m	11.35	12.80
Pix2struct+ LLMTransfer	0h	23h27m	14.21	13.52
<i>(d) The performance of different FT time, with trainable or frozen Qwen LLM</i>				
Pix2struct+Original LLM (Frozen LLM)	8h09m	17h28m	26.02	18.48
Pix2struct+ LLMTransfer (Frozen LLM)	0h	17h31m	31.33	24.56
Pix2struct+ LLMTransfer (30% FT Data)	0h	6h17m	14.84	12.44
Pix2struct+ LLMTransfer (60% FT Data)	0h	12h18m	38.92	36.88
<i>(e) Frozen Qwen LLM with trainable MLP projector and LoRA adapters</i>				
CLIP+Original LLM	2h26m	13h55m	25.48	16.28
Pix2struct+Original LLM	8h09m	19h30m	26.56	16.52
Pix2struct+ LLMTransfer	0h	19h09m	26.19	16.80

systematic generalization analysis to examine whether the observed transferability is robust across multiple dimensions. In particular, we investigate whether the proposed transfer recipe holds consistently when varying (i) the LLM backbone, (ii) the projector architecture used to interface heterogeneous vision encoders with the LLM, (iii) the downstream fine-tuning budget, and (iv) the LLM training strategy and parameter-efficiency regime. By explicitly ablating LLMTransfer under these diverse conditions, we aim to clarify the scope and limitations of its applicability, and to provide a more comprehensive empirical understanding of when and why reusing an aligned LLM can effectively support encoder-level transfer without additional vision-language re-alignment. We detail these analyses along the following four aspects, with the corresponding results shown in Tables 3–7.

5.4.1 Generalization Across LLM Backbones

To examine whether the proposed transfer recipe generalizes beyond a single LLM backbone, we additionally replace Qwen2-1.5B (2024) with LLaMA3.2-1B (Touvron et al., 2023) LLM while keeping the same training recipe and evaluation protocol. The results mostly support the same conclusion across all evaluated vision encoders: once an LLM is vision-language aligned in a source VLM, it can be effectively reused with new task-specific encoders without requiring additional re-alignment pretraining. Specifically, as reported in Table 3(b), LLMTransfer with LLaMA improves Codetr(Zong et al., 2023) performance on POPE from 86.75 to 87.30 while eliminating alignment pretraining. On DINOv2 (Oquab et al., 2023), Table 4(b) shows that GQA increases from 56.07 to 56.62, accompanied by consistent gains on REC-related metrics. For SAM-based (Kirillov et al., 2023) grounding, the results in Table 5(b) demonstrate comparable or improved REC/RES performance across most

splits. In the medical VQA setting, Table 6(b) reports a clear improvement on SLAKE-close for BiomedCLIP (Zhang et al., 2023), while Table 7(b) shows that Pix2Struct (Lee et al., 2023) achieves improved ChartQA accuracy, increasing from 32.56 to 32.76. Although minor task-dependent variations are observed on certain metrics, such as SLAKE-open or individual RES splits, the overall trend remains consistent across diverse grounding and VQA benchmarks, while completely eliminating the need for additional alignment pretraining. These results indicate that the effectiveness of LLMTransfer is not tied to a specific LLM family, and that transferable vision–language alignment can be preserved across different LLM backbones.

5.4.2 Generalization Across Projector Architectures

In addition to the default MLP projector, we adopt a cross-attention Q-Former (Li et al., 2023) as the vision–language interface while keeping the aligned LLM and all other training settings unchanged. As reported in the Q-Former settings in Tables 3(c)–7(c), LLMTransfer mostly maintains its effectiveness across all evaluated vision encoders without requiring additional vision–language re-alignment. Specifically, Codetr with LLMTransfer improves POPE from 78.93 to 80.68, as shown in Table 3(c). On DINOv2, Table 4(c) shows that LLMTransfer increases GQA from 39.79 to 51.97, together with consistent gains on REC-related metrics. For SAM-based grounding, Table 5(c) reports improved REC and RES performance across most splits, for example increasing $REC_{ref+testA}$ from 8.44 to 9.89. In the medical VQA setting, Table 6(c) shows that LLMTransfer improves $SLAKE_{close}$ from 21.89 to 29.78, while Table 7(c) demonstrates gains on document and chart understanding tasks, with ChartQA increasing from 12.80 to 13.52. Although a small number of metrics show slightly lower performance than the Original LLM baseline, the magnitude of such degradation remains tightly bounded within 1.1 points, while completely eliminating the need for additional pretraining time and data. These results indicate that the core conclusions of this work are not tied to a specific projector design. Instead, the aligned LLM can robustly accommodate newly initialized projector architectures, including cross-attention based interfaces, and continue to support effective encoder-level transfer without additional alignment pretraining.

5.4.3 Generalization Across Fine-Tuning Budgets

We further analyze how the effectiveness of LLMTransfer varies under different fine-tuning (FT) budgets, with the goal of understanding its robustness to reduced training time. We consider two complementary settings: reducing the proportion of FT data, and freezing the LLM while training only the projector. As summarized in Tables 3(d)–7(d), LLMTransfer

remains effective under moderate FT reduction. For example, on Codetr, using only 30% FT data already achieves a POPE score of 87.13, remaining close to the full-FT result, while 60% FT data largely preserves performance. Similar trends are observed on DINOv2 and SAM, where reducing FT to 30%–60% maintains competitive REC and RES performance across multiple splits. On BiomedCLIP, reducing FT to 60% retains most of the performance gains on SLAKE while substantially reducing training time. Although Pix2Struct exhibits a more noticeable performance drop under severe FT reduction, its performance is largely recovered when the FT budget is increased to 60%, indicating sensitivity to extreme but not moderate FT constraints.

In contrast, fully freezing the LLM consistently leads to performance degradation across all evaluated encoders, even when the overall training cost is comparable. As shown in Tables 3(d)–7(d), projector-only adaptation results in clear drops on both general metrics, such as POPE on Codetr, and task-specific grounding metrics on DINOv2, SAM, and BiomedCLIP. Overall, these results reveal a clear performance–efficiency trade-off: while full fine-tuning yields the best performance, LLMTransfer remains robust under reduced FT budgets by reusing an aligned LLM and allowing limited but essential LLM adaptation, whereas fully freezing the LLM imposes a performance bottleneck across diverse encoders.

5.4.4 Generalization Across LLM Training Strategies

To verify whether the gains of LLMTransfer originate from prior vision–language alignment rather than downstream full fine-tuning, we conduct ablations under different LLM training strategies. Specifically, we compare unfrozen LLM fine-tuning with frozen-LLM settings, where only the visual projector or lightweight adapters are trained. The corresponding results are summarized in Tables 3(a,d,e)–7(a,d,e).

First, fully freezing the LLM leads to performance degradation for all methods, confirming that downstream LLM fine-tuning is generally beneficial. However, even under frozen-LLM settings, LLMTransfer preserves a substantially larger fraction of its unfrozen performance than the Original LLM. For example, on DINOv2-based GQA in Table 4(a,d), frozen LLMTransfer retains 46.09 out of 60.79, whereas the frozen Original LLM retains only 38.19 out of 58.83. On SAM-based grounding tasks in Table 5(a,d), freezing the Original LLM often results in near-zero RES scores, while frozen LLMTransfer still achieves meaningful RES performance. More importantly, under identical frozen-LLM conditions, LLMTransfer consistently and substantially outperforms the corresponding frozen Original LLM across all evaluated encoders and tasks. On BiomedCLIP, as shown in Table 6(d), freezing the Original LLM causes $SLAKE_{open}/SLAKE_{close}$ to collapse to 20.93/22.67, whereas LLMTrans-

fer maintains strong performance at 64.61/31.06. Similarly, on Pix2Struct in Table 7(d), frozen LLMTransfer improves DocVQA and ChartQA from 26.02/18.48 to 31.33/24.56. For grounding tasks on DINOv2 in Table 4(d), frozen LLMTransfer improves $REC_{ref+val}$ from 3.49 to 37.51 and REC_{ref_test} from 3.18 to 36.77.

This advantage further generalizes to parameter-efficient adaptation using LoRA. Under identical frozen-LLM + LoRA settings, LLMTransfer mostly achieves competitive or improved performance compared to the Original LLM across multiple encoders. For instance, on Codetr in Table 3(e), PoPE improves from 83.33 to 84.41, and on DINOv2 in Table 4(e), REC_{ref_test} increases from 37.65 to 39.39. More pronounced gains are observed on SAM-based grounding in Table 5(e), where LLMTransfer improves REC_{ref_testA} from 40.86 to 43.12 under LoRA. Overall, these results demonstrate that the effectiveness of LLMTransfer is not solely attributable to downstream full fine-tuning. Instead, it is largely driven by transferable vision-language alignment already encoded in the aligned LLM, which remains effective even when the LLM is frozen or adapted using lightweight parameter-efficient methods. Allowing limited LLM fine-tuning further improves performance, suggesting that prior alignment and downstream adaptation are complementary rather than mutually exclusive.

5.5 Mechanistic Analysis via CKA

While the previous sections establish the empirical effectiveness and generalization of LLMTransfer across diverse settings, we further investigate why such transferability emerges from a mechanistic perspective. To this end, we conduct a Centered Kernel Alignment (Kornblith et al., 2019) (CKA) analysis to probe feature-space similarity among projected visual representations from heterogeneous vision encoders within the LLM embedding space. In our framework, CLIP serves as the base vision encoder used to derive the aligned LLM and therefore provides a natural reference for comparison. We first perform this analysis on the standard in-domain image setting used in our main evaluation, and the results are summarized in Fig. 3. For Qwen, continual alignment consistently increases feature-space similarity between CLIP and other encoders: BiomedCLIP improves from 0.391 to 0.420, Codetr from 0.324 to 0.349, DINOv2 from 0.432 to 0.452, and SAM from 0.574 to 0.616, whereas Pix2Struct remains low and nearly unchanged, increasing only from 0.079 to 0.081. A similar trend is observed for LLaMA, where alignment raises CLIP-to-encoder similarity for BiomedCLIP from 0.335 to 0.385, Codetr from 0.262 to 0.340, DINOv2 from 0.384 to 0.419, and SAM from 0.516 to 0.587. These in-domain results provide mechanistic evidence that the continual aligned LLM does not merely improve downstream performance, but systematically reshapes the

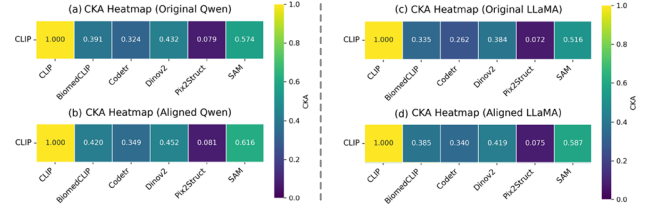


Fig. 3 Centered Kernel Alignment (CKA) analysis of projected visual features between our base encoder CLIP and downstream heterogeneous vision encoders. Figures (a)&(b) show the CKA similarity under the original and aligned Qwen, while (c)&(d) correspond to the original and aligned LLaMA, respectively

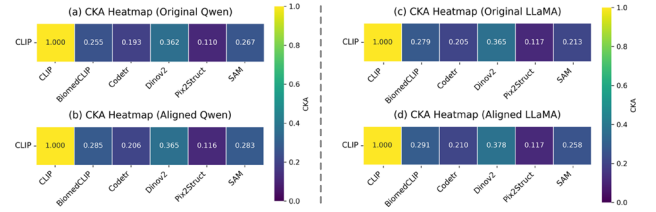


Fig. 4 Out-of-distribution Centered Kernel Alignment (CKA) analysis of projected visual features between our base encoder CLIP and downstream heterogeneous vision encoders on ImageNet-Sketch

LLM embedding space to better accommodate heterogeneous visual representations, pulling diverse encoder features closer to the CLIP-aligned manifold.

To further examine whether this alignment effect remains valid beyond the in-domain setting, we additionally conduct an out-of-distribution (OOD) CKA analysis on ImageNet-Sketch (Wang et al., 2019), as shown in Fig. 4. ImageNet-Sketch differs substantially from standard natural images in visual appearance, and therefore provides a useful testbed for evaluating whether the observed feature-space alignment is robust under sketch-style domain shift. Under the Qwen backbone, the aligned LLM improves or maintains the CKA similarity between CLIP and all task-specific encoders, increasing BiomedCLIP from 0.255 to 0.285, Codetr from 0.193 to 0.206, DINOv2 from 0.362 to 0.365, Pix2Struct from 0.110 to 0.116, and SAM from 0.267 to 0.283. A similar trend is observed under the LLaMA backbone, where the CKA similarity increases from 0.279 to 0.291 for BiomedCLIP, from 0.205 to 0.210 for Codetr, from 0.365 to 0.378 for DINOv2, remains stable at 0.117 for Pix2Struct, and increases from 0.213 to 0.258 for SAM.

These OOD results further suggest that the feature-space alignment induced by LLMTransfer is not merely fitted to standard natural-image inputs. Instead, the transferred aligned LLM provides a reusable vision-language embedding space that can still accommodate heterogeneous encoder features under sketch-style domain shift.

Table 8 Hallucination evaluation under different methods and LLM backbones. The PoPE denotes PoPE Precision Li et al. (2023), which evaluates object-existence hallucination with binary questions. CHAIR_s

Method	Qwen+MLP Projector			Qwen+QFormer			LLaMA+MLP Projector		
	PoPE ↑	CHAIR _s ↓	CHAIR _i ↓	PoPE ↑	CHAIR _s ↓	CHAIR _i ↓	PoPE ↑	CHAIR _s ↓	CHAIR _i ↓
CLIP+Original LLM	94.41	29.74	10.90	68.93	44.40	23.00	51.21	24.15	8.98
Codetr+Original LLM	94.52	26.58	10.01	89.97	41.23	18.95	71.16	23.83	8.53
Codetr+LLMTransfer	95.38	25.47	9.64	92.32	39.48	18.70	82.28	21.90	7.69

5.6 Hallucination Analysis

Beyond open-world task coverage, an important aspect of trustworthiness in visual recognition is reducing visual hallucination, i.e., producing confident but incorrect claims about non-existent objects. To assess this factor, we report PoPE Precision (Li et al., 2023) as a direct proxy for hallucination, defined as $\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$, where FP corresponds to false positive “Yes” predictions on negative samples. Higher precision indicates fewer hallucinated responses. Since PoPE mainly evaluates object-existence hallucination through binary questions, we further introduce CHAIR_s and CHAIR_i (Rohrbach et al., 2018) to assess hallucination in open-ended caption generation. CHAIR_s measures the proportion of generated captions containing at least one hallucinated object, while CHAIR_i measures the fraction of hallucinated object instances among all mentioned objects. Lower CHAIR values indicate fewer hallucinated object mentions.

As shown in Table 8, LLMTransfer consistently achieves higher PoPE Precision than CLIP-based models and Codetr with the Original LLM across different LLM backbones and projector architectures. Meanwhile, LLMTransfer also obtains lower CHAIR_s and CHAIR_i scores, showing that the reduction of hallucination is not limited to binary object-existence judgment but also holds in free-form caption generation. For example, under the Qwen with an MLP projector setting, PoPE Precision improves from 94.52% to 95.38%, while CHAIR_s and CHAIR_i decrease from 26.58 to 25.47 and from 10.01 to 9.64, respectively. Under a LLaMA configuration, it increases substantially from 71.16% to 82.28%, with CHAIR_s and CHAIR_i further decreasing from 23.83 to 21.90 and from 8.53 to 7.69. By transferring an LLM that has already been aligned with visual representations, LLMTransfer helps the model interpret projected visual tokens in a more visually grounded manner, thereby reducing the tendency to generate plausible but unsupported object concepts. Therefore, the aligned prior not only improves binary object-existence judgment, but also regularizes free-form language generation toward image-consistent descriptions.

In addition, as shown in Fig. 6, these quantitative improvements are further supported by qualitative examples. In

and CHAIR_i Rohrbach et al. (2018) evaluate object hallucination in open-ended caption generation

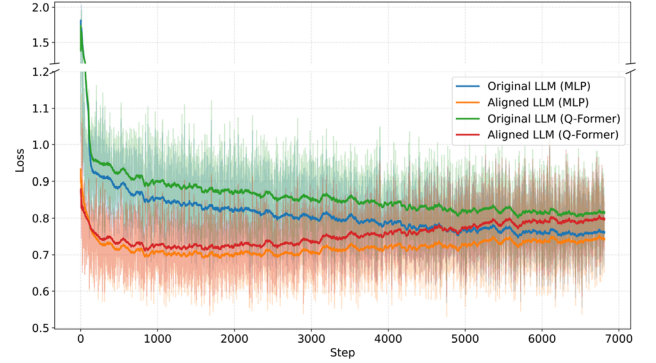


Fig. 5 Training loss curves of Codetr-based VLMs with different projectors. The aligned LLM improves convergence, while MLP converges faster and more stably than Q-Former

these cases, LLMTransfer avoids hallucinated affirmative responses to object-existence queries, such as incorrectly predicting a potted plant in a forest scene or a dining table in a pizza-related image. These results indicate that transferring an already vision-aligned LLM effectively suppresses hallucinated affirmative responses, thereby improving faithfulness and trustworthiness in open-world visual recognition scenarios.

5.7 Projector Convergence Dynamics

To further examine the optimization efficiency of different projectors, we compare the training loss curves of Codetr-based VLMs under four settings: Original LLM with MLP, Aligned LLM with MLP, Original LLM with Q-Former, and Aligned LLM with Q-Former. As shown in Fig. 5, the aligned LLM consistently leads to faster convergence than the original LLM, indicating that the transferred vision-language alignment provides a better initialization for downstream adaptation. Meanwhile, the MLP projector exhibits a more stable optimization trajectory and reaches a lower final loss than Q-Former, especially in the middle-to-late stage of fine-tuning.

The less stable and less favorable convergence behavior of Q-Former mainly comes from its higher optimization complexity. Different from MLP, which directly maps visual features into the LLM embedding space, Q-Former intro-

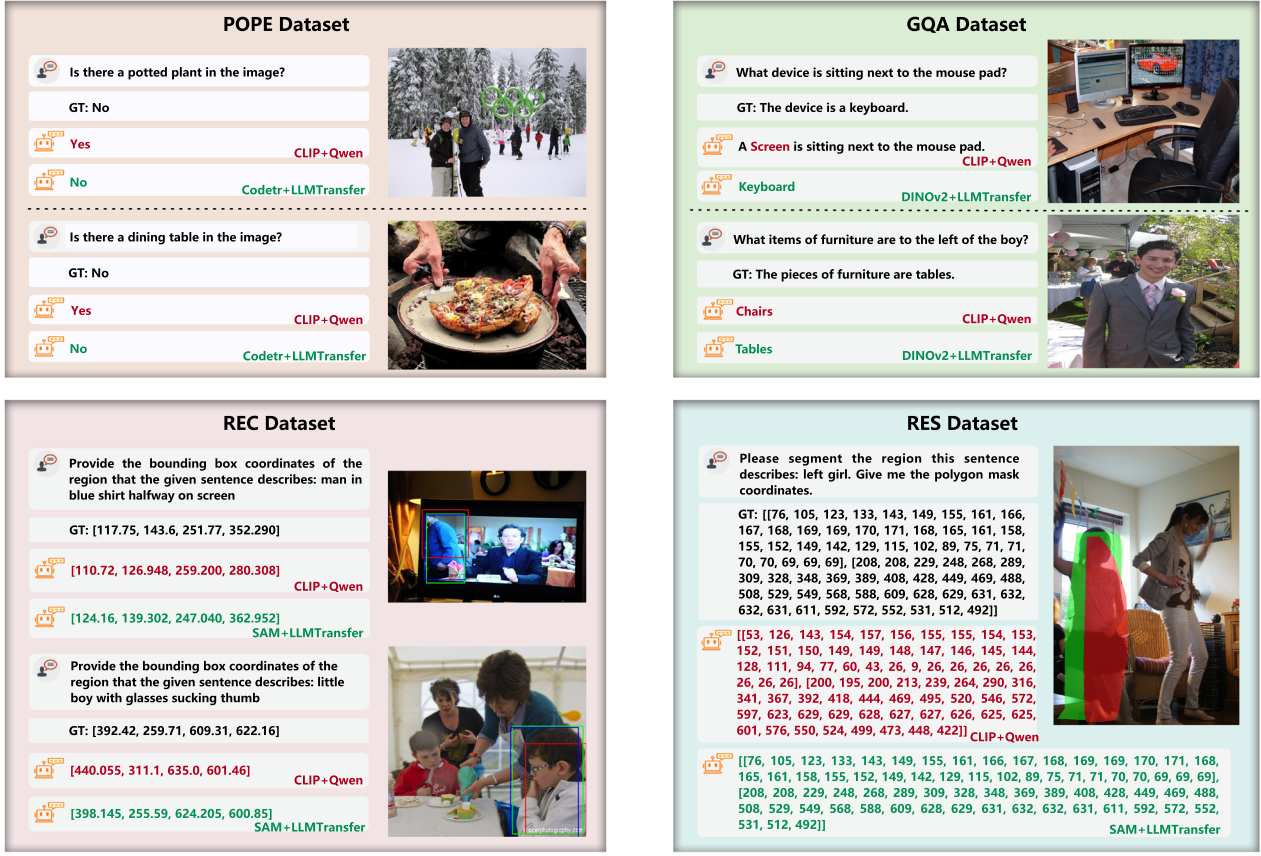


Fig. 6 Visualization of LLMTransfer strategy and the two-stage CLIP-based method on the POPE (Li et al., 2023), GQA (Hudson & Manning, 2019), REC and RES (Yu et al., 2016) datasets. GT represents the Ground Truth

duces learnable query tokens and cross-attention layers to resample visual features. This adds more randomly initialized parameters and creates an additional information bottleneck. For different encoders such as Codetr, whose features contain dense spatial and region-level cues, query-based compression may make fine-grained visual alignment harder during early training. Therefore, although LLMTransfer still saves the pre-training/re-alignment stage, the practical efficiency gain can be partially affected by the projector design. These results support using MLP as the efficient default projector in LLMTransfer, while Q-Former further verifies that aligned LLM transfer is not limited to a single projector architecture.

5.8 Visualization Results

To comprehensively evaluate the effectiveness of the LLM-Transfer strategy, we present visual comparisons across a range of encoders and datasets. As illustrated in Figs. 6 and 7, the LLMTransfer strategy consistently outperforms the two-stage CLIP-based VQA baseline in diverse multimodal tasks, including general visual question answering, spatial reasoning, medical image interpretation, document understanding, and fine-grained region localization.

5.8.1 General Visual Understanding

As shown in Fig. 6, on the POPE dataset (Li et al., 2023), the Codetr-based VLMs trained via LLMTransfer demonstrate stronger hallucination resistance. For example, in a case where the input image contains a plant that is not potted, our method accurately captures this distinction, whereas the CLIP-based model incorrectly labels it as a potted plant, leading to a hallucinated answer. Such robustness indicates that the transferred LLM can effectively adapt to the unique feature representations of Codetr, thereby preserving semantic fidelity in complex visual scenes. This improvement underscores the importance of retraining the aligned LLM under the Codetr encoder, validating the effectiveness of the continual cross-encoder LLM transfer recipe.

5.8.2 Spatial Reasoning

As shown in Fig. 6, on the GQA dataset (Hudson & Manning, 2019), the DINOv2-based VLMs trained with the LLMTransfer strategy excel in perceiving object relations in complex visual scenes. Our model correctly identifies the spatial relation between a keyboard and a mouse pad, avoid-

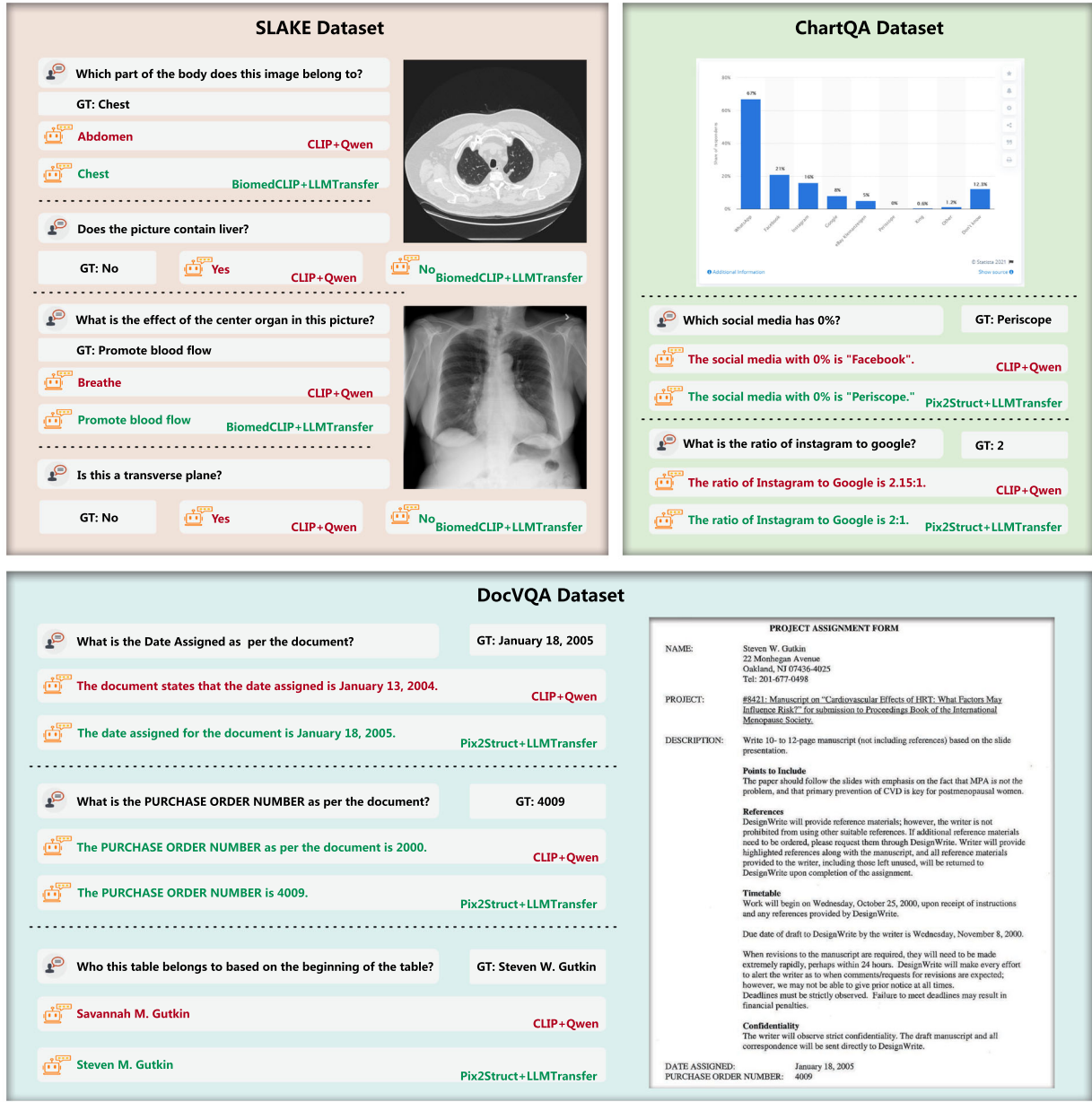


Fig. 7 Visualization of LLMTransfer strategy and the two stage CLIP-based method on the SLAKE (Liu et al., 2021), ChartQA (Masry et al., 2022) and DocVQA (Mathew et al., 2021) datasets. GT represents the Ground Truth

ing common misclassifications like mistaking a table for a chair when only partially visible. In contrast, the CLIP-based approach struggles under such occlusion or ambiguity. This indicates that the continual LLM exhibits a more refined understanding of scene layout and context under the DINOv2 encoder.

5.8.3 Fine-Grained Localization

In both referring expression comprehension (REC) and segmentation (RES) tasks, the SAM-based VLMs achieves

higher precision, as shown in Fig. 6. For instance, the bounding boxes of LLMTransfer (green) show greater alignment with ground truth (blue) compared to those from the CLIP-based model, which often include irrelevant regions. Furthermore, when localizing complex textual queries (e.g., "little boy with glasses sucking thumb"), the LLMTransfer strategy captures the full target, whereas the baseline omits critical parts. In segmentation, our green mask encompasses a larger, more accurate region than the red mask from the baseline. Although some missing arm areas remain due to the mixed-task training, they do not significantly affect the

overall evaluation, as our focus is on assessing LLM transfer effectiveness rather than achieving state-of-the-art segmentation performance.

5.8.4 Medical Image Understanding

The LLMTransfer strategy also proves effective in domain-specific scenarios. As shown in Fig. 7, on the SLAKE dataset, the BiomedCLIP-based VLMs trained with LLMTransfer accurately identifies anatomical structures such as the chest and liver, whereas the CLIP-based baseline confuses the chest with the abdomen and misses the liver entirely. Moreover, it successfully determines imaging planes (transverse vs. coronal) and identifies the heart's location and function, while the baseline fails on all fronts. These results highlight the effectiveness of LLMTransfer in improving both interpretability and diagnostic reasoning in medical tasks.

5.8.5 Document and Chart Comprehension

Finally, on the DocVQA (Mathew et al., 2021) and ChartQA (Masry et al., 2022) benchmarks, the Pix2Struct-based VQA performance demonstrates superior document and table understanding. As shown in Fig. 7, it accurately interprets dates, personal names, and chart scales, while the CLIP-based model misreads or omits key details, such as the correct day, year, or first name in a full name. In chart-related tasks, the LLMTransfer correctly infers the vertical axis scale, leading to precise classifications and value comparisons, whereas the baseline fails to extract this information. These results affirm the broad adaptability and robustness of the transfer recipe in structured visual information understanding.

In summary, across diverse tasks and modalities, the general transfer recipe enables the continual alignment of LLMs with various encoders—such as Codetr, DINOv2, SAM, BiomedCLIP, and Pix2Struct—without requiring additional pre-training. The consistent improvements over CLIP-based models in downstream tasks validate the effectiveness and generalizability of our transfer strategy.

6 Discussion

Limitation. Despite its effectiveness and efficiency, our approach has limitations: the transfer recipe assumes different VLMs share compatible feature dimensions and relies on lightweight projectors, which may limit scalability to highly heterogeneous vision encoders. It also depends on a well-aligned initial LLM, and performance may drop if the alignment is suboptimal.

Conclusion. In this work, we explore a simple yet effective transfer recipe for training VLMs without the need for re-aligning vision-language representations. By systemati-

cally investigating the transferability of an LLM that has been pre-aligned with a general-purpose vision encoder (e.g., CLIP), we demonstrate that such an aligned LLM can be directly reused with task-specific vision encoders without additional alignment procedures. Our extensive experiments across five vision encoders and 13 evaluation metrics reveal three key insights: (1) the aligned LLM retains transferable vision-language understanding that boosts downstream performance even when paired with new encoders; (2) additional alignment or distillation is unnecessary, leading to significant training efficiency gains; and (3) the new projector layer can be effectively optimized through downstream instruction tuning alone. These findings suggest that a single aligned LLM can serve as a universal foundation across diverse VLMs, offering a scalable and resource-efficient paradigm for future multi-modal learning.

Future work. The future directions involve investigating transfer strategies for dynamic or temporally structured encoders (e.g., in video or 3D), cross-modality settings (e.g., in audio or depth), extending the proposed recipe to more LLM families such as Mistral (Jiang et al., 2023) and Gemma (Team et al., 2024), as well as evaluating the framework under online and continual learning conditions.

Funding This work was supported by the National Natural Science Foundation of China under Grant Nos. 62572093 and 62531022, the Liaoning Provincial Science and Technology Joint Program Project under Grant No. 2024011188-JH2/1026, and Fundamental Research Funds for the Central Universities (DUT24RC(3)025).

Data Availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declarations

Competing interests The authors declare no competing interests.

References

- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35, 23716–23736.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *Seelmsys.org (accessed 14 April 2023)*, 2(3), 6. <https://vicuna>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale arXiv preprint. [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
- Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-

- language models with instruction tuning. *Advances in Neural Information Processing Systems* **36** (2024)
- Deng, M., Wang, J., Hsieh, C.-P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E.P., Hu, Z. (2022) Rlprompt: Optimizing discrete text prompts with reinforcement learning. arXiv preprint [arXiv:2205.12548](https://arxiv.org/abs/2205.12548)
- Diao, H., Zhang, Y., Liu, W., Ruan, X., & Lu, H. (2023). Plug-and-play regulators for image-text matching. *IEEE Transactions on Image Processing*, 32, 2322–2334.
- Hudson, D. A., Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6700–6709)
- He, J., Wang, Y., Wang, L., Lu, H., He, J.-Y., Lan, J.-P., Luo, B., Xie, X. (2024) Multi-modal instruction tuned llms with fine-grained visual perception. arXiv preprint [arXiv:2403.02969](https://arxiv.org/abs/2403.02969)
- Han, Y., Zhang, C., Chen, X., Yang, X., Wang, Z., Yu, G., Fu, B., Zhang, H. (2023) Chartllama: A multimodal llm for chart understanding and generation. arXiv preprint [arXiv:2311.16483](https://arxiv.org/abs/2311.16483)
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7b arXiv preprint. [arXiv:2403.08295](https://arxiv.org/abs/2403.08295)
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., & Lo, W.-Y. (2023). *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015–4026) Segment anything.
- Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. *International Conference on Machine Learning* (pp. 3519–3529) PMLR.
- Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning arXiv preprint [arXiv:2104.08691](https://arxiv.org/abs/2104.08691).
- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models arXiv preprint. [arXiv:2305.10355](https://arxiv.org/abs/2305.10355).
- Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1), 1–10.
- Lee, K., Joshi, M., Turc, I. R., Hu, H., Liu, F., Eisenschlos, J. M., Khandelwal, U., Shaw, P., Chang, M.-W., & Toutanova, K. (2023). Pix2struct: Screenshot parsing as pretraining for visual language understanding. *International Conference on Machine Learning* (pp. 18893–18912) PMLR.
- Li, D., Li, J., Hoi, S. (2024) Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems* **36**.
- Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J. (2024) LLaVA-NeXT: Improved reasoning, OCR, and world knowledge <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Li, J., Li, D., Savarese, S., Hoi, S. (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International Conference on Machine Learning*, pp. 19730–19742 PMLR.
- Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems* (p. 36).
- Li, J., Li, D., Xiong, C., Hoi, S. (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International Conference on Machine Learning*, pp. 12888–12900 PMLR.
- Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J. (2024) Lisa: Reasoning segmentation via large language model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589.
- Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J. (2024) Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36**.
- Li, W., Yuan, Y., Liu, J., Tang, D., Wang, S., Qin, J., Zhu, J., Zhang, L. (2025) Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 1–19.
- Lester, B., Yurtsever, J., Shakeri, S., & Constant, N. (2022). Reducing retraining by recycling parameter-efficient prompts arXiv preprint. [arXiv:2208.05577](https://arxiv.org/abs/2208.05577)
- Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., & Wu, X.-M. (2021). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (pp. 1650–1654). IEEE.
- Mathew, M., Karatzas, D., & Jawahar, C. (2021). Docvqa: A dataset for vqa on document images. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2200–2209)
- Masry, A., Long, D. X., Tan, J. Q., Joty, S., & Hoque, E. (2022). Chartqa: A benchmark for question answering about charts with visual and logical reasoning arXiv preprint. [arXiv:2203.10244](https://arxiv.org/abs/2203.10244).
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., & El-Nouby, A. (2023). Dinov2: Learning robust visual features without supervision arXiv preprint. [arXiv:2304.07193](https://arxiv.org/abs/2304.07193).
- Pei, B., Huang, Y., Chen, G., Xu, J., Wang, Y., Wang, L., Lu, T., Qiao, Y., Wu, F. (2025). Guiding audio-visual question answering with collective question reasoning. *International Journal of Computer Vision*, 1–18.
- Panagopoulou, A., Xue, L., Yu, N., Li, J., Li, D., Joty, S., Xu, R., Savarese, S., Xiong, C., Niebles, J.C. (2023). X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. arXiv preprint [arXiv:2311.18799](https://arxiv.org/abs/2311.18799)
- Qwen2 technical report (2024)
- Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., Saenko, K. (2018). Object hallucination in image captioning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4035–4045)
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning* (pp. 8748–8763) PMLR.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1–67.
- Shi, H., Dao, S. D., & Cai, J. (2025). Llmformer: Large language model for open-vocabulary semantic segmentation. *International Journal of Computer Vision*, 133(2), 742–759.
- Sharma, P., Ding, N., Goodman, S., Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565.
- Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs arXiv preprint. [arXiv:2111.02114](https://arxiv.org/abs/2111.02114)
- Su, Y., Wang, X., Qin, Y., Chan, C.-M., Lin, Y., Wang, H., Wen, K., Liu, Z., Li, P., Li, J., et al. (2021) On transferability of prompt tuning for natural language processing. arXiv preprint [arXiv:2111.06719](https://arxiv.org/abs/2111.06719)
- Song, C. H., Wu, J., Washington, C., Sadler, B. M., Chao, W.-L., & Su, Y. (2023). Llm-planner: Few-shot grounded planning for embodied agents with large language models. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2998–3009).

- Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. (2024). Cambrian-1: A fully open, vision-centric exploration of multimodal llms. arXiv preprint [arXiv:2406.16860](https://arxiv.org/abs/2406.16860).
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., & Azhar, F. (2023). Llama: Open and efficient foundation language models arXiv preprint. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971).
- Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., & Love, J. (2024). Gemma: Open models based on gemini research and technology arXiv preprint. [arXiv:2403.08295](https://arxiv.org/abs/2403.08295)
- Truong, T.-D., Nguyen, H.-Q., Nguyen, X.-B., Dowling, A., Li, X., Luu, K. (2025). Insect-foundation: A foundation model and large multimodal dataset for vision-language insect understanding. *International Journal of Computer Vision*, 1–26.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* **30**.
- Wang, H., Ge, S., Lipton, Z., Xing, E. P. (2019). Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32.
- Wang, J., Ke, L. (2024). Llm-seg: Bridging image segmentation and large language model reasoning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1765–1774).
- Wang, Z., Rao, Y., Sun, S., Liu, X., Wei, Y., Yu, X., Liu, Z., Wang, Y., Liu, H., Zhou, J., et al. (2025). Vision generalist model: A survey: Z. wang et al. *International Journal of Computer Vision*, 1–29.
- Wang, Y., Yu, Z., Wang, J., Heng, Q., Chen, H., Ye, W., Xie, R., Xie, X., & Zhang, S. (2024). Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1), 224–237.
- Wang, Z., Zhao, L., Zhang, J., Song, R., Song, H., Meng, J., Wang, S. (2025). Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, 1–23.
- Yang, J., Chen, X., Qian, S., Madaan, N., Iyengar, M., Fouhey, D. F., Chai, J. (2023). Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent arXiv preprint. [arXiv:2309.12311](https://arxiv.org/abs/2309.12311)
- Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L. (2016). Modeling context in referring expressions. In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pp. 69–85 Springer.
- Yu, J., Xiong, H., Zhang, L., Diao, H., Zhuge, Y., Lanqing, H., Wang, D., Lu, H., He, Y., & Chen, L. Lms can evolve continually on modality for x-modal reasoning. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Zhang, A., Fei, H., Yao, Y., Ji, W., Li, L., Liu, Z., Chua, T.-S. (2024). Vpgrans: Transfer visual prompt generator across llms. *Advances in Neural Information Processing Systems* **36**.
- Zang, Y., Li, W., Han, J., Zhou, K., & Loy, C. C. (2025). Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2), 825–843.
- Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., Liu, Y. (2024). Mova: Adapting mixture of vision experts to multimodal context. arXiv preprint [arXiv:2404.13046](https://arxiv.org/abs/2404.13046).
- Zhang, W., Shen, L., & Foo, C.-S. (2025). Source-free domain adaptation guided by vision and vision-language pre-training. *International Journal of Computer Vision*, 133(2), 844–866.
- Zong, Z., Song, G., & Liu, Y. (2023). Detrs with collaborative hybrid assignments training. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6748–6758).
- Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al. (2023). Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint [arXiv:2303.00915](https://arxiv.org/abs/2303.00915).
- Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9), 2337–2348.
- Zhang, P., Zhang, J., Cao, J., Li, H., & Jin, L. (2025). Smaller but better: Unifying layout generation with smaller large language models. *International Journal of Computer Vision*, 133(7), 3891–3917.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.