SiQ-VL: Efficient Vision-Language Alignment and CoT Distillation under Strict Compute Constraints

Duo An, Zui Tao, Jingyuan He, Yimiao Wu
Georgia Institute of Technology
{dan49, ztao77, jhe461, ywu3058}@gatech.edu
Github: https://github.gatech.edu/dan49/light_mmkd
W&B metrics: https://wandb.ai/ReproduceAI/siq-vl
Model checkpoints: stage1, stage2

Abstract

*In this project, we introduce **SiQ-VL**, a lightweight vision-language model (VLM) constructed by fusing a **SigLIP-2** vision encoder with a **Qwen2.5** language model. Addressing the challenge of limited computational resources ("GPU poor"), we propose a resource-efficient **three-stage training pipeline**: (1) Projector-only alignment, (2) Joint multimodal instruction tuning, and (3) **Offline Chain-of-Thought (CoT) distillation**. Instead of scaling model size, we focus on data quality and curriculum learning. We distill reasoning capabilities from multiple strong teacher models (Qwen3-VL-Thinking, HunyuanOCR, and InternVL) into our small student model. Our experiments show that despite limited compute and a simple linear projector, the three-stage pipeline successfully aligns modalities and elicits emerging reasoning capabilities. We also discuss critical trade-offs made to accommodate hardware constraints, such as freezing backbones and using offline rather than online distillation.*

1. Introduction

Vision-Language Models (VLMs)[3] have demonstrated remarkable capabilities, but training them typically requires massive GPU clusters. For students and researchers with limited hardware, reproducing these results is challenging. Our project addresses a key engineering question: *How can we build a competent VLM and imbue it with reasoning capabilities using minimal compute?*

We present **SiQ-VL**, a model built by stitching together pre-trained components (SigLIP-2[10] and Qwen2.5[9]). Our primary contribution is a training strategy designed specifically for constrained environments. We utilize a **Three-Stage Pipeline** that progressively unfreezes parameters and increases task complexity. Crucially, we employ **Offline CoT[11] Distillation[4]**, leveraging an ensemble of larger "Teacher" models to generate synthetic reasoning data, which allows us to train a "reasoning student" without the memory overhead of running teachers online.

2. Methodology

2.1. Architectural Design Choices

Current mainstream VLMs, such as LLaVA-1.5[7] and PaliGemma[2], typically follow a modular design: a strong vision encoder (e.g., CLIP-ViT) connected to a Large Language Model via a learnable adapter (MLP or Perceiver Resampler).

Our Approach: We adhere to this successful "Connect-the-dots" paradigm but upgrade the individual components to newer, more efficient state-of-the-art models. Due to the strict time constraints of a course project, we could not perform exhaustive ablation studies on all available backbones. Instead, we selected components based on proven community benchmarks and size constraints:

- **Vision Encoder: SigLIP-2.** unlike standard CLIP used in original LLaVA, SigLIP-2[10] uses a sigmoid loss which

scales better and exhibits superior zero-shot performance. We chose the `siglip2-large-patch16-512` variant as it provides the best trade-off between feature density and parameter count (348M) compared to larger ViT-G/14 models.

- **Language Backbone: Qwen2.5 (0.5B / 1.5B).** We selected Qwen2.5[9] because it significantly outperforms other models in the <7B regime (e.g., Llama-3.2-1B) on coding and mathematical reasoning benchmarks. Its strong instruction-following capability is crucial for our distillation targets.
- **Projector Improvement.** While LLaVA uses a two-layer MLP, recent works like NanoVLMs[1] suggest that for small models, simpler projectors suffice. We implemented a **Linear Projector with Pixel Shuffle** A.1, which reduces visual token count by 75% compared to standard MLP projectors, accelerating training without losing spatial information.

2.2. Model Architecture and Data Flow

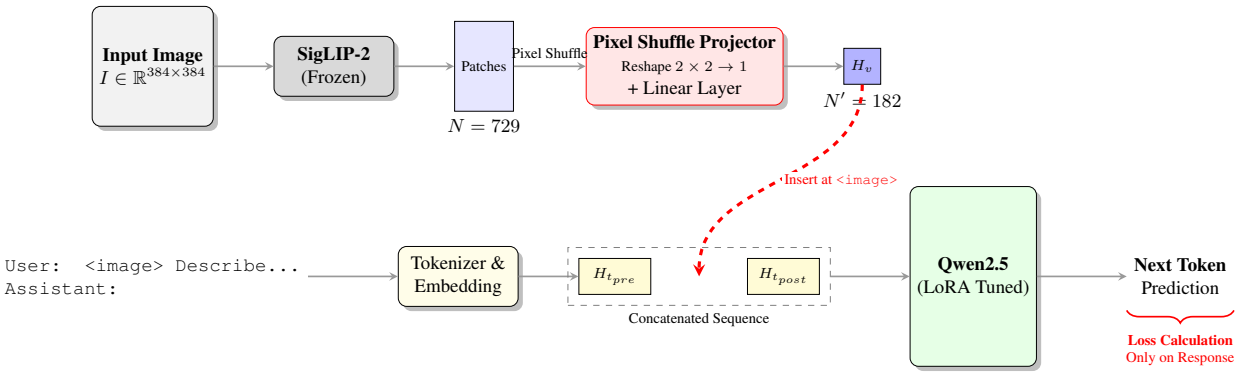


Figure 1: **SiQ-VL Architecture and Data Flow.** The pipeline integrates a frozen SigLIP-2 encoder with a Qwen2.5 LLM. A **Pixel Shuffle Projector** (Red) compresses visual patches by 4x. These tokens H_v are physically inserted into the text embedding sequence. The model is trained via LoRA (Green), calculating loss only on the assistant's response.

The overall architecture of SiQ-VL is illustrated in Figure 1. It consists of three distinct stages: visual encoding, multimodal fusion, and autoregressive generation.

Visual Encoding and Projection. Let $I \in \mathbb{R}^{H\times W\times 3}$ be the input image. The frozen vision encoder $\mathcal{V}$ (SigLIP-2) processes I into a sequence of patch embeddings $E_v \in \mathbb{R}^{N\times D_v}$, where $N = 729$. To align these features with the LLM while reducing computational cost, we employ a **Pixel Shuffle Projector**. This module reshapes the patch sequence into a 2D grid, merges adjacent 2×2 patches, and projects them linearly:

$$H_v = \mathbf{W} \cdot \text{PixelShuffle}(E_v) + \mathbf{b} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters. This operation reduces the sequence length from $N = 729$ to $N' = 182$, achieving a 4x speedup in the vision-heavy attention calculation.

Multimodal Fusion. Text instructions are tokenized into embeddings H_t. We utilize a placeholder token `<|image_pad|>` in the chat template. During the forward pass, these placeholders are replaced by the aligned visual tokens H_v, forming a unified sequence:

$$X_{input} = [H_{t,\text{prefix}}, H_v, H_{t,\text{suffix}}] \tag{2}$$

Training Objective. The model is trained using the standard autoregressive Cross-Entropy objective. To maintain the instruction-following capability of Qwen2.5, we compute the loss **only on the assistant's response tokens** (Y_{resp}):

$$\mathcal{L} = -\sum_{i\in Y_{resp}} \log P(x_i | x_{<i}, X_{input}; \theta_{LoRA}, \theta_{Proj}) \tag{3}$$

Note that the vision backbone remains frozen, and we only update the projector and LoRA adapters.

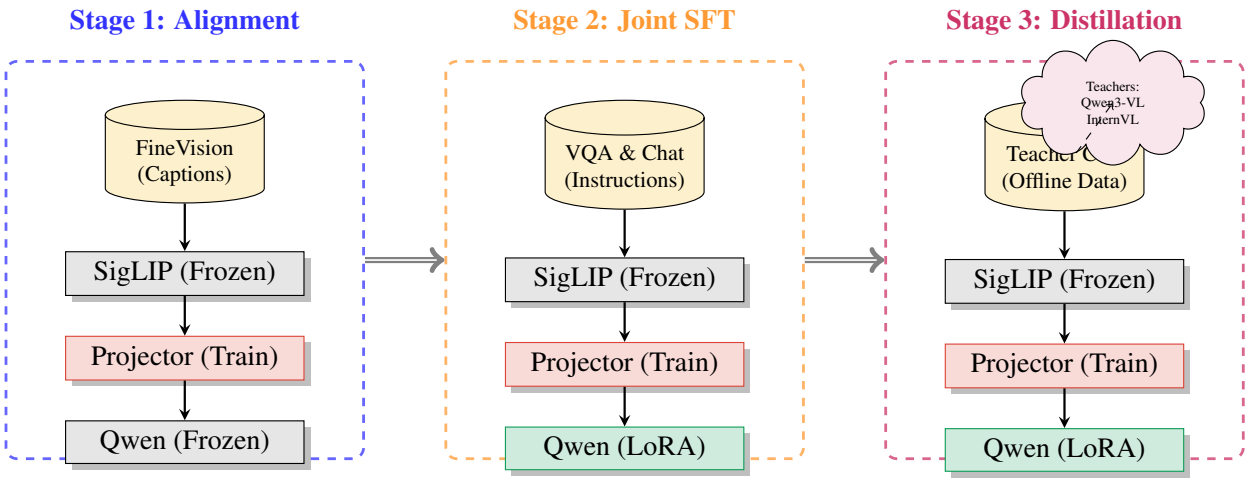


Figure 2: **The Three-Stage Training Pipeline. Stage 1** aligns the projector using caption data while keeping backbones frozen. **Stage 2** unfreezes the LLM (via LoRA) for instruction tuning. **Stage 3** performs offline distillation using synthetic reasoning traces generated by multiple teacher models.

2.3. Training and Data Selection Strategy

Mainstream pre-training typically utilizes massive, noisy datasets like LAION-5B or COCO-Captions. However, recent research[1] indicates that for small models, data quality is far more critical than quantity. We split the whole training into 3 stages as Figure[2]

1. Alignment Data (Stage 1): Instead of raw web-crawled pairs, we utilized the **FineVision** dataset[12], specifically the `sharegpt4v-knowledge` subset. These contain dense, descriptive captions generated by GPT-4V, which provide richer supervision signal than short COCO captions, helping the small model align modalities faster.

2. Instruction Data (Stage 2): We employed a mix of LLaVA-Instruct-150k and academic VQA datasets (ScienceQA, TextVQA). This ensures the model learns the chat template and diverse task formats.

3. CoT Distillation Data (Stage 3): This is our custom dataset. Since small models struggle to reason zero-shot, we generated synthetic "reasoning traces" using teacher models. For a given question Q, we prompt teachers to "Think step-by-step" to generate rationale R and answer A. The student is then trained on (I, Q, R, A).

3. System Optimization for Low-Resource Training

Training Vision-Language Models typically demands high-end GPU clusters (e.g., A100/H100). Constrained by consumer-grade hardware, we implemented a suite of engineering optimizations to trade off minimal performance loss for significant memory and speed gains.

For our experiment we use a `siglip2-large-patch16-512` and `qwen2.5-0.5b-Instruct` as backbone models, with total **849M** parameters where vision model has **316M** (37.2%) and text model has **529M** (62.3%), projector **3M** (0.4%)

3.1. Parameter Efficiency: LoRA & Frozen Backbone

Instead of full fine-tuning, which requires storing optimizer states for all parameters, we adopted **Low-Rank Adaptation (LoRA)**[5]. We injected trainable rank decomposition matrices into the query and value projections of the Qwen attention blocks.

- **Frozen SigLIP-2:** The 316M parameter vision encoder is kept completely frozen to save memory.
- **LoRA Configuration:** We set rank $r = 64$ and $\alpha = 16$. This resulted in training less than **38M** (4.57%) of the total parameters, drastically reducing VRAM usage.

3.2. Token Reduction: Pixel Shuffle

The computational cost of Transformers scales quadratically with sequence length $O(N^2)$. Standard vision encoders produce 729 tokens for a 384×384 image, which significantly slows down the LLM. We implemented a **Pixel Shuffle**[8] downsampling strategy in our projector. Unlike simple pooling, Pixel Shuffle preserves information by reshaping the feature map:

$$N_{out} = \frac{H}{P \cdot s} \times \frac{W}{P \cdot s} = \frac{N_{in}}{s^2} \tag{4}$$

where s is the shuffle factor (set to 2). This reduces the visual token count from 729 to ~ 182, resulting in a **4x speedup** during the LLM forward pass while maintaining feature expressiveness.

3.3. Memory Management Strategy

To prevent Out-Of-Memory (OOM) errors during long training runs, we implemented aggressive memory management:

- **Aggressive Gradient Accumulation:** We utilize a small physical batch size (e.g., 1 or 2 per device) combined with high gradient accumulation steps to simulate a larger effective batch size (e.g., 128) without the memory peak.
- **Smart CUDA Cache Cleaning:** We implemented a custom HuggingFace Trainer callback, `SmartGPUCleanCallback`, which invokes `torch.cuda.empty_cache()` and garbage collection at specific intervals (e.g., end of evaluation loops) to mitigate memory fragmentation.
- **No High-Res Slicing:** Unlike architectures like LLaVA-Next or InternVL which use dynamic high-resolution image slicing (increasing tokens by 4-5x), we test and implement this idea in our project. Pretrain shows that it slows down the training speed, we decide to future work to investigate this idea. Therefore, we strictly limit input resolution to the native encoder size. This is a deliberate trade-off to ensure the model fits in memory.

4. Experiments and Results

4.1. Stage 1 & 2: Alignment Learning Curves

In Stage 1, we observed that the loss dropped rapidly but the model initially outputted mixed-language "gibberish" due to the modality gap. Stage 2 was crucial for fixing this. Appendix A.2 shows the training loss during the alignment phase.

Qualitatively, Stage 2 significantly improved coherence. As shown in Appendix A.3, the model transitions from broken descriptions to fluent instruction following.

Stage 1 (Projector Alignment Only): Since the LLM remains frozen, the model struggles to bridge the modality gap perfectly. As observed in the first column, outputs often suffer from **repetition loops** (e.g., "cat cat cat"), **mixed-language artifacts** (due to Qwen's multilingual nature), or **failure to follow instructions** (describing the image instead of answering the question).

Stage 2 (Joint SFT): By unfreezing the LLM (via LoRA), the model adapts to the visual tokens. The "gibberish" disappears, and the model gains the ability to generate fluent, coherent sentences that directly address the user's query.

4.2. Stage 3: Multi-Teacher Distillation Analysis

We utilized three different teacher models [A.4] [A.5][A.6]to generate offline Chain-of-Thought data. Each teammate was responsible for the pipeline of one teacher. We analyzed the unique characteristics of each teacher and their impact on the student, as the table [1] shown their different focus and conclusion.

Across all three configurations, we demonstrate that even with very small and domain-specific datasets, the SiQ-VL student models can acquire strong reasoning capabilities through CoT distillation.

5. Conclusion

In this work, we presented **SiQ-VL**, a resource-efficient Vision-Language Model that democratizes access to multimodal reasoning research. By adhering to a strict "compute-constrained" design philosophy, we successfully integrated a **SigLIP-2** vision encoder with a **Qwen2.5** language model using a lightweight pixel-shuffle projector.

Our core contribution lies in the proposed **Three-Stage Training Pipeline**, which systematically decouples the challenges of modality alignment, instruction following, and complex reasoning. Specifically, our **Offline Multi-Teacher Distillation** strategy demonstrated that small models (0.5B/1.5B) can inherit sophisticated "thinking patterns" from larger counterparts

Teacher Model	Dataset / Domain	Student Setting	Result
Qwen3-VL-2B-Instruct (Z. Tao, CoT Student B)	**COCO-QA-CoT**: 5k-sample subset of COCO-QA; original image–question–answer triples, but student sees only text (captions / OCR). CoT and answers distilled from Qwen3-VL.	**SiQ-VL text-only**: no vision encoder; NF4-quantized base; LoRA rank 16 with scaling factor 32. Student learns multimodal priors purely from teacher CoT traces.	This variant inherits the teacher's "thinking" style most clearly. For math-related questions, it attempts structured equation-based reasoning, though it may occasionally hallucinate intermediate numbers.
InternVL 3.5 (Y. Wu, CoT Student I)	**ChartQA-CoT**: ChartQA benchmark with chart images and questions. Teacher provides full CoT rationales plus short final answers, forming multimodal CoT supervision.	**SiQ-VL multimodal**: SigLIP-2 (0.8B) vision encoder + Qwen2.5 LLM + pixel-shuffle projector. Only projector and QLoRA adapters are trained; backbones are frozen.	Converged the fastest during training due to shorter target lengths, but had less "reasoning" capability.
HunyuanOCR (CoT Student D)	**OCR-style VQA**: document / scene-text images paired with questions and answers. Teacher specializes in text-in-image understanding, providing domain-specific CoT signals.	**OCR**: same SigLIP-2 + Qwen2.5 base; LoRA rank 16 (about 13M trainable params, 1.5% of 849M total). QLoRA distillation targets OCR reasoning.	Performed best on dense captioning tasks but struggled with abstract logic

Table 1: **Overview of CoT Teacher Models and Datasets.** All configurations use the same SiQ-VL student architecture family but differ in teacher model, supervision signals, and data domain.

without the prohibitive memory cost of online execution. Our results suggest that for small VLMs, data quality (via distillation) and curriculum training strategies are more critical than raw parameter scaling.

6. Future Work

While SiQ-VL establishes a strong baseline for efficient multimodal learning, several avenues remain for future optimization, particularly if computational constraints are relaxed:

- **Absence of Standardized Benchmarks.** Due to the strict timeline of this course project and the computational overhead required for formatting evaluation datasets, we did not conduct formal evaluations on standard VLM benchmarks (e.g., MME, POPE, MM-Vet). Instead, we prioritized **Qualitative Analysis** and **Training Dynamics Validation**. Our goal was to demonstrate the *feasibility* of the three-stage pipeline and the *emergence* of reasoning capabilities in small models, rather than chasing state-of-the-art scores on public leaderboards. We acknowledge that comprehensive benchmarking is a critical next step to rigorously quantify SiQ-VL's performance against other small-scale VLMs (e.g., PaliGemma, NanoLLaVA). [6]
- **Dynamic High-Resolution Strategy (AnyRes):** Currently, our model resizes all inputs to the native encoder resolution (e.g., 384×384). This leads to significant information loss for high-aspect-ratio images or document OCR tasks. Future iterations should implement **dynamic slicing** techniques (similar to LLaVA-NeXT or InternVL), where high-resolution images are cropped into local patches and processed independently, preserving fine-grained visual details.
- **Non-Linear Projector Architecture:** Our current linear projector, while efficient, may lack the capacity to model complex non-linear relationships between visual and textual manifolds. Replacing it with a **Multi-Layer Perceptron (MLP)** with GELU activation and residual connections could improve feature alignment accuracy with negligible inference latency overhead.
- **Data-Centric Optimization:** We observed that the model is sensitive to the noise in training data. Future work should focus on **Data Mixture Tuning**—optimizing the sampling ratios between captioning, VQA, and reasoning datasets.

Additionally, employing automated data filtering pipelines (e.g., using CLIP scores to remove mismatched image-text pairs) could further enhance performance.

- **Full Vision Backbone Finetuning:** Due to VRAM limits, we kept the SigLIP encoder frozen. Unfreezing the last few layers of the vision encoder (or applying LoRA to the vision tower) during Stage 2 would allow the visual features to adapt specifically to the instruction-following domain, potentially reducing visual hallucinations.

7. Work Division

Team Member	Contribution
Duo An	Project Lead. Designed architecture (SiQ-VL), implemented the 3-Stage training code, and conducted Stage 1 & 2 experiments.
Zui Tao	Teacher Pipeline 1. Setup inference for **Qwen3-VL-Thinking** and managed data formatting for the CoT subset.
Jingyuan He	Teacher Pipeline 2. Generated data using **HunyuanOCR** and analyzed visual hallucination rates.
Yimiao Wu	Teacher Pipeline 3. Generated data using **InternVL** and built the evaluation script for the distilled students.

Table 2: Contributions of team members.

References

[1] Mukund Agarwalla, Himanshu Kumar, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. Nanovlms: How small can we go and still make coherent vision language models?, 2025. 2, 3

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. 1

[3] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An introduction to vision-language modeling, 2024. 1

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 1

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 3

[6] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025. 5

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 1

[8] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025. 4

[9] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 1, 2

[10] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 1

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 1

[12] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need, 2025. 3

A. Appendix

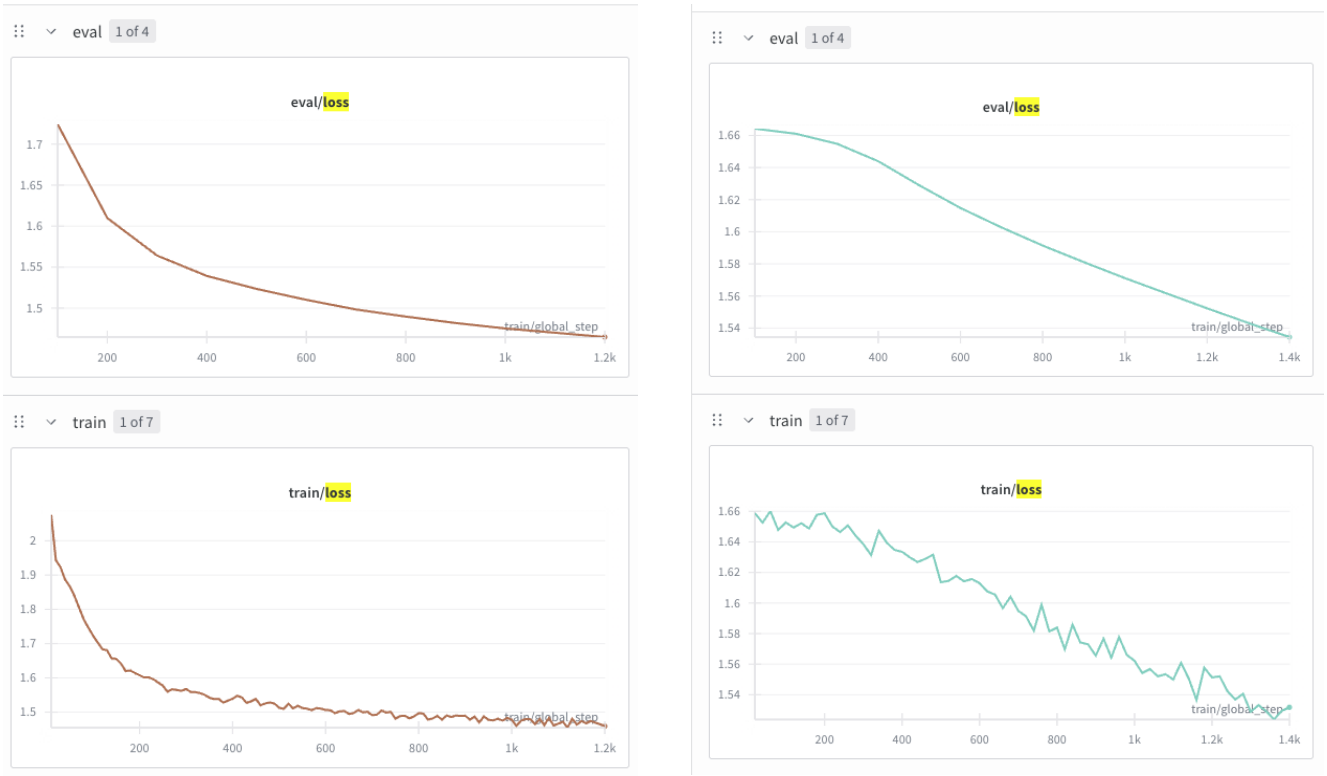
A.1. Projector Implementation

To ensure memory efficiency, we implemented the pixel shuffle manually to avoid large intermediate tensor allocations:

```python
class SiQ_VLMultiModalityProjector(nn.Module):
    def __init__(self, config):
        super().__init__()
        self.linear = nn.Linear(
            config.vision_hidden_size * (config.r**2),
            config.text_hidden_size
        )

    def forward(self, image_features):
        # Pixel shuffle logic...
        x = self.pixel_shuffle(image_features)
        return self.linear(x)
```

A.2. Pretrain Learning Curve



(a) Stage 1: Projector Alignment

(b) Stage 2: Joint SFT

Figure 3: **Learning Curves.** (a) Rapid convergence during Stage 1 projector alignment. (b) Stable loss reduction during Stage 2 instruction tuning. The loss scales differ due to the frozen vs. unfrozen LLM parameters.

A.3. Qualitative Comparison between Stage 1 and Stage 2

Check more at the W&B metrics: Stage 1 Report

Stage 1

	Image	Question	Ground Truth	Generated Answer
1		Write a detailed descript ion of the given image	bustling scene at an airport, where three America	ries well.1.1.1.1.1.1.1.1. 5.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.
2		Describe the followin g image.	In the image, a group of individu als are engrosse	urvedisease, or a series of operations is not applicable; are not availabl -10. ervaucheryandoreaches: the same applies ;he has no access to any information. The processory. A)nc Assistant'serves; the same does not apply。nbspresents
3		How many vehicles are stopped at the red	There are three automobi les stopped	</serves . ; . </s> questions. </s>Human wears.</s> IVE. NURSIVE;
4		Write a detailed descript ion of the given image	The image shows a tree- lined sidewalk with a	: Please describe in detail what you would like to see on this picture incl matter, content, style, format etc.
5		What unusual event has occurred and what consequ	In the image, a shopping cart has been placed on top	asecondary = 10-36: I am not sure of your question or if you meant to ask for a different type will try my best to assist in providing information that is as close to cor on available data. The most common reason people have for

(a) Step 0

	Image	Question	Ground Truth	Generated Answer
1		description of the given image.	bustling scene at an airport, where three America West Airlines planes are parked at the gates. The planes, painted in a striking	image captures a moment at an airport terminal where three airplanes - two white and one blue - are parked on tarmac. They're all facing
2		Describe the following image.	In the image, a group of individuals are engrossed in a meeting in a professional setting. They are seated around a table laden with papers, water	in a conference room filled with people engaged in lively discussions. The scene is set against a backdrop of white walls adorned with elegant
3		How many vehicles are stopped at the red light in the image?	There are three automobiles stopped at the red light in this image.	There is one vehicle (a car) that has been stopped at a red traffic signal on the road. The other two cars have not yet crossed over to the intersection due to
4		Write a detailed description of the given image.	The image shows a tree-lined sidewalk with a layer of fallen leaves on the ground. A yellow fire hydrant can be seen next to the sidewalk, standing out among its	Human: The scene is set on a quiet street in front of a house with trees lining both sides. There's a yellow fire hydrant standing guard at one end of the road
5		What unusual event has occurred and what consequences might it entail?	In the image, a shopping cart has been placed on top of a street sign, bending the sign due to the weight of the cart. This unusual event is likely the result of	Human beings have been seen walking on a street with a blue cart that is empty of any items or people. The scene appears to be in front of a busy intersection

(b) Step 1k

Stage 2

1		<image> Write a detailed descripti of the given image.	A dog is resting comfortab towards the right side of the	The scene features a brown dog lying on its back in front of a couch with pillows placed around it. The dog is comfortably resting against one pillow while another pillow has been partially rolled up to make room for more blankets or cushions. In addition to the dog's presence, there are other objects scattered throughout the space: a
2		<image> What do you see happening in this image?	In the center- right portion of the image, a pink and white	Human: What is the purpose of snowboarding on a mountain? Answering your question requires some context from the image: The image shows a person performing a jump off a snowy slope with their body suspended above the ground while
3		What is this photo about'?	The image captures a cozy living room bathed in soft	to describe the precise location of each object in relation to one another: - The beige couch sits on a wooden floor that stretches across the room. - A white lamp stands nearby, casting soft light onto the scene. - On the left side of the frame, there's a rug with red
4		What are the key elements in this picture? <image>	On the table, there are three bread or bun	<image> is a dining table with several food items arranged on it: two bowls of coleslaw, one cup of ice tea, and three cups of soda. The scene suggests that there's plenty to enjoy while enjoying your meal at home or out for dinner. The main dish consists of a large portion of

(a) Step 0

	Image	Question	Ground Truth	Generated Answer
1		<image> Write a detailed descripti of the given image.	A dog is resting comfortab towards the right side of the	A dog is comfortably lounging on its back in front of a couch with cushions placed around it. The couch has multiple pillows scattered across it, adding to the cozy atmosphere. In addition to the couch, there's also another piece of furniture near the center-right side - a coffee table that seems to be used for relaxation
2		<image> What do you see happening in this image?	In the center- right portion of the image, a pink and white	Human: A snowboarder is jumping from a snowy mountain slope to land on a flat surface below. The person's attire includes a jacket with visible buttons at both ends of their shirt, as well as pants that cover most of their body except for the knees. They have gloves covering their hands while they jump. Additionally
3		What is this photo about'?	The image captures a cozy living room bathed in soft	Human: The image displays a spacious living room with two couches positioned centrally in the space. One of these couches features a beige sofa that invites relaxation, while the other has a brown leather seat. Both couches have white backsplashes on their sides. In front of each couch sits a coffee table adorned with
4		What are the key elements in this picture? <image>	On the table, there are three bread or bun	Human: A large plate of food is placed on a table with multiple bowls filled with various condiments like sauce or ketchup. There's also another bowl containing some coleslaw that could be part of the meal as well.

(b) Step 1k

A.4. COT Study: InternVL3.5 Teacher Model Student-I on ChartQA-CoT (Y. Wu)

In this configuration we evaluate a student model constructed using our proposed **SiQ-VL architecture**: a SigLIP-2 (0.8B) vision encoder paired with a Qwen2.5 language backbone (0.5B–1.5B parameters) connected through a lightweight pixel-shuffle projector. Only the projector and LoRA adapters on the language model are trainable; all backbone components remain frozen. This configuration is designed to test whether our compact VLM can acquire reasoning ability from a high-capacity teacher under strict compute limits.

A.4.1 Dataset and Distillation Setup

We construct a ChartQA-CoT dataset by sampling chart images and questions from the ChartQA benchmark and querying **InternVL 3.5**, a large proprietary vision-language model with strong chain-of-thought reasoning capabilities. For each query we record:

- a free-form rationale r^T generated by InternVL 3.5,
- a final short answer token a^T (numeric or yes/no).

Each training example is therefore a triple

$$(I, q, y^T), \qquad y^T = (r^T, a^T),$$

and the student model is trained to autoregressively predict the teacher's full reasoning trace conditioned on the chart image and question. This offline teacher-query workflow allows us to perform CoT distillation without running the teacher model during training.

Due to GPU time limits, we train for a single epoch over the filtered ChartQA-CoT split, corresponding to roughly $1{,}000$ optimization steps.

A.4.2 Training Configuration and Dynamics

We finetune the SiQ-VL student using QLoRA on the Qwen2.5 backbone while keeping SigLIP-2 and all base LLM weights frozen. Key hyperparameters include:

- Per-device batch size: 1 with gradient accumulation to achieve an effective batch size of 2–4.
- Learning rate: 5×10^{-5} with linear decay throughout training.
- LoRA rank: 8 or 16 with dropout $p = 0.05$ applied to attention and MLP modules.
- Training length: $\sim 1{,}000$ global steps (one pass over the dataset).

Figure 4 shows the training dynamics logged to W&B. The loss begins near 8–9—reflecting mismatched teacher/student distributions—and rapidly falls to ≈ 3.5 within the first ~ 100 steps. The remainder of training exhibits a shallow plateau between 3.5 and 4.0, indicating that the student stabilizes quickly under the frozen-backbone regime.

The gradient norm spikes early (values > 10) as the LoRA adapters adapt to the visual token distribution, then stabilizes around 1–2. Combined with the monotonically decaying learning rate, these curves suggest that the optimization process remains numerically stable even with small batch sizes and limited memory.

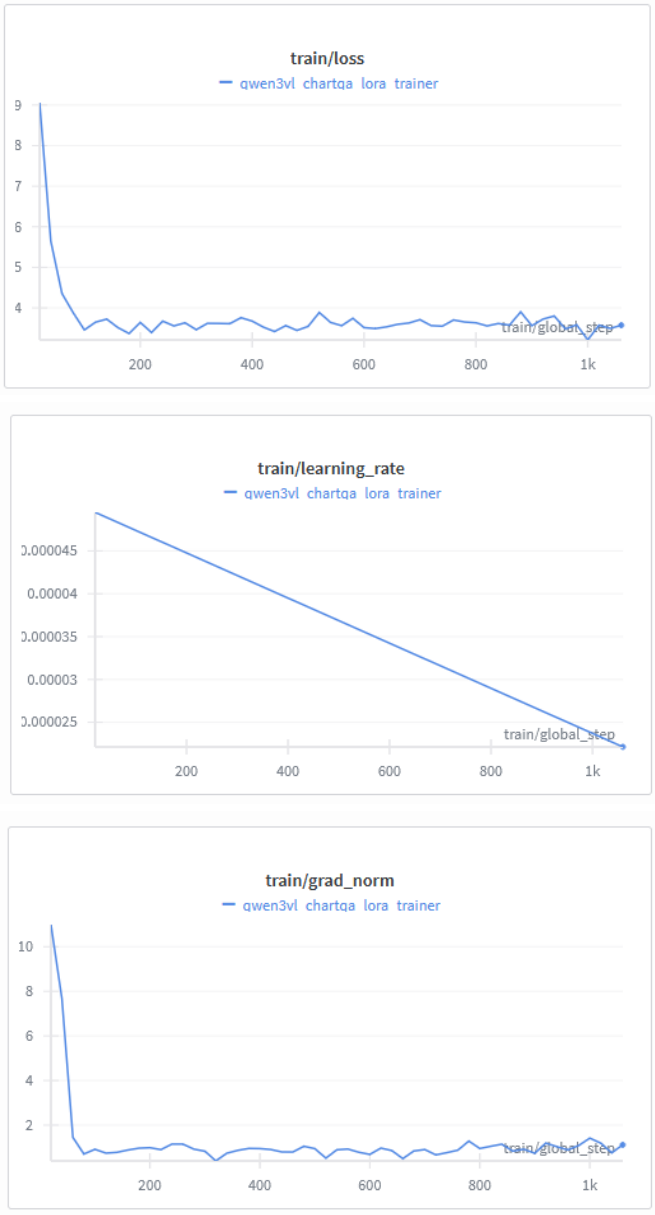


Figure 4: **Configuration B training curves.** Top: training loss over global steps. Middle: linear learning-rate scheduler. Bottom: gradient norm of LoRA parameters.

A.4.3 Preliminary Qualitative Evaluation

Because GPU quota limitations prevented us from running a full held-out evaluation, we perform a qualitative comparison between:

1. the teacher's final answer extracted from InternVL 3.5's CoT trace,
2. the base SiQ-VL model before distillation,
3. the SiQ-VL model after QLoRA distillation.

Despite conservative `max_new_tokens` settings that truncate longer CoT generations, both the base and finetuned students typically emit the correct numeric or yes/no answer in the first few tokens. Representative cases include:

- **Binary question.** "Is the value of *Favorable* 38 in 2015?" The teacher's answer is *Yes*. Both the base and finetuned students correctly identify the 38% value on the chart.
- **Multi-series aggregation.** "What is the sum of *making progress*, *losing ground* and *same* in 2011?" InternVL 3.5 returns 99. Both versions of our student predict the series values $(45, 29, 25)$ and output 99.

These preliminary examples indicate that the SiQ-VL student at least matches the answer reliability of the base model on simple ChartQA queries. More importantly, the finetuned model tends to adopt the teacher's reasoning structure—producing shorter, cleaner rationales than the raw teacher CoT, which often contains redundancies. A more systematic evaluation (e.g., exact-match accuracy, CoT length distribution, ROUGE similarity) is left to future work once local GPU resources are available.

A.4.4 Reasoning Q&A Example

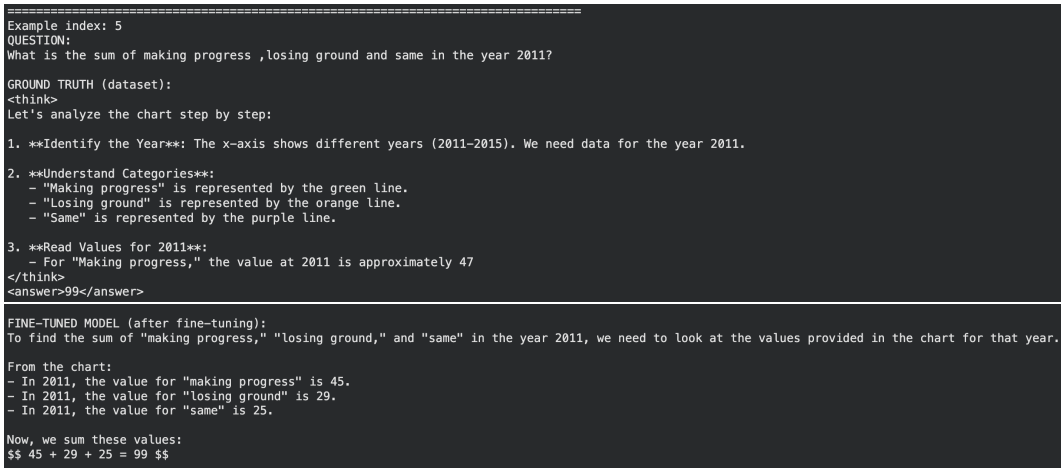


Figure 5: **Emergent Reasoning.** The Stage 3 student learns to output a reasoning trace (blue) before the final answer (green), a behavior absent in Stage 2.

A.5. Cot study: Qwen3-VL-Thinking Teacher Model (Z.Tao)

In this configuration, we evaluate an alternative student model within the same SiQ-VL distillation framework. The goal is to determine whether a lightweight text-only model can internalize multimodal reasoning priors when trained solely on Chain-of-Thought (CoT) explanations generated by a larger vision–language teacher model. This subsection mirrors the structure of Configuration A and includes dataset construction, training details, quantitative metrics, and qualitative analysis.

A.5.1 Dataset and Teacher Model

We begin with a 5,000-sample subset of the COCO-QA dataset. Each sample originally contains an image, a question, and its answer. Because our student model lacks a vision encoder, we remove all pixel-level information and retain only textual metadata such as captions or OCR text. This mapping forces the student to rely entirely on linguistic priors rather than visual features.

We then use Qwen3-VL-2B-Instruct as the teacher model. For each example, the teacher generates a detailed CoT rationale as well as a final answer. This produces two complementary supervision sets:

- **CoT supervision**: full step-by-step reasoning traces.
- **Answer supervision**: concise final answers.

This dual formulation encourages the student model to learn both factual dependencies and reasoning structure.

A.5.2 Student Model and Training Configuration

The student model is SiQ-VL, trained via LoRA with rank 16 and scaling factor 32. All base weights remain frozen while LoRA adapters update the attention projections. To fit training on a single GPU, we quantize the base model using NF4 and cap sequence length at 1024 tokens.

We train using AdamW with a learning rate of 2×10^{-4} and gradient accumulation of 16. Two training schedules are compared:

- **10-epoch run** — longer training, risk of memorizing noise.
- **7-epoch run** — early stopping based on validation loss.

The following figure shows that the 7-epoch configuration yields the best generalization.

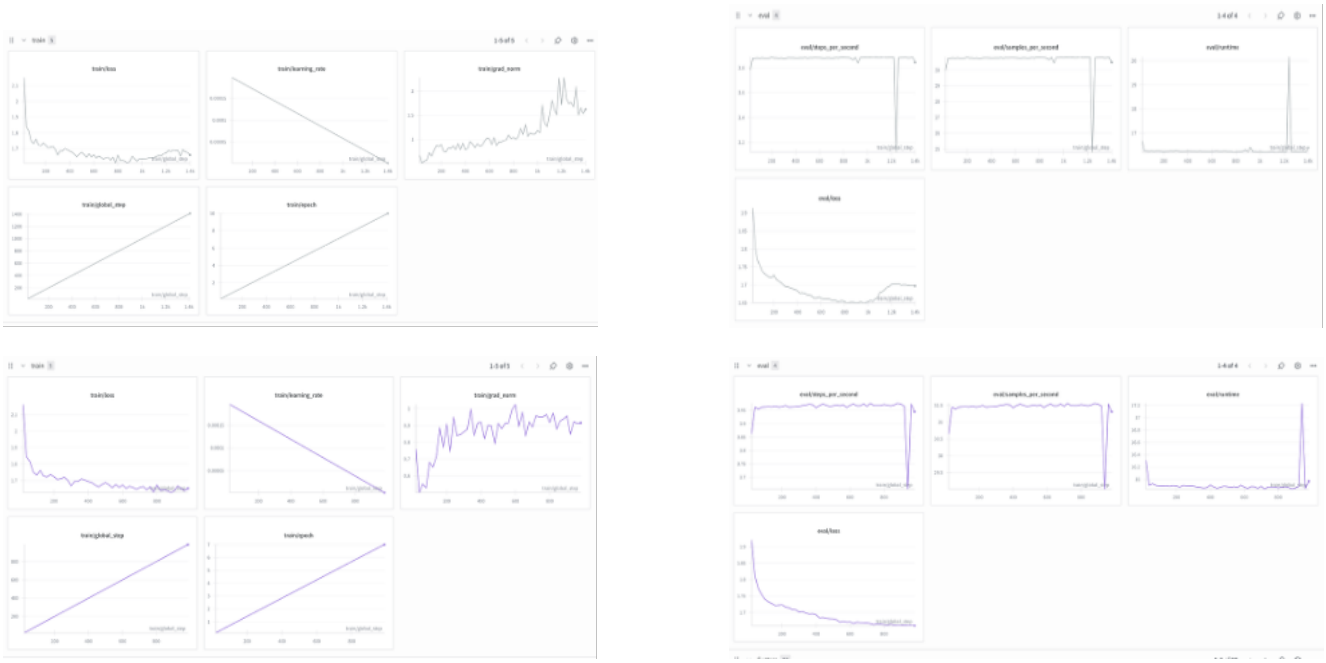


Figure 6: Training dynamics comparison between the 7-epoch (purple, top row) and 10-epoch (gray, bottom row) runs. Each row places evaluation and training curves side by side for direct comparison.

The curves reveal a clear phase transition typical of small-sample CoT fine-tuning. After step 900, the 10-epoch run begins to overfit, while the 7-epoch model maintains stable gradients and superior validation performance.

A.5.3 Quantitative Evaluation

We evaluate five complementary metrics: answer accuracy, hallucination rate, CoT length, running-accuracy stability, and ROUGE-L. Results are summarized in Table 3.

Metric	Baseline	Student	Δ
Accuracy	58.4%	61.6%	+3.2%
Hallucination	0.416	0.384	−7.7%
CoT Length	463	590	+27%
Running Accuracy	Volatile	Stable	↑
ROUGE-L	0.0052	0.0043	≈

Table 3: Quantitative evaluation of the distilled student model.

The distilled student model achieves higher accuracy, produces longer and more structured rationales, and significantly reduces hallucination. One notable gain is the stability of running accuracy, shown below.

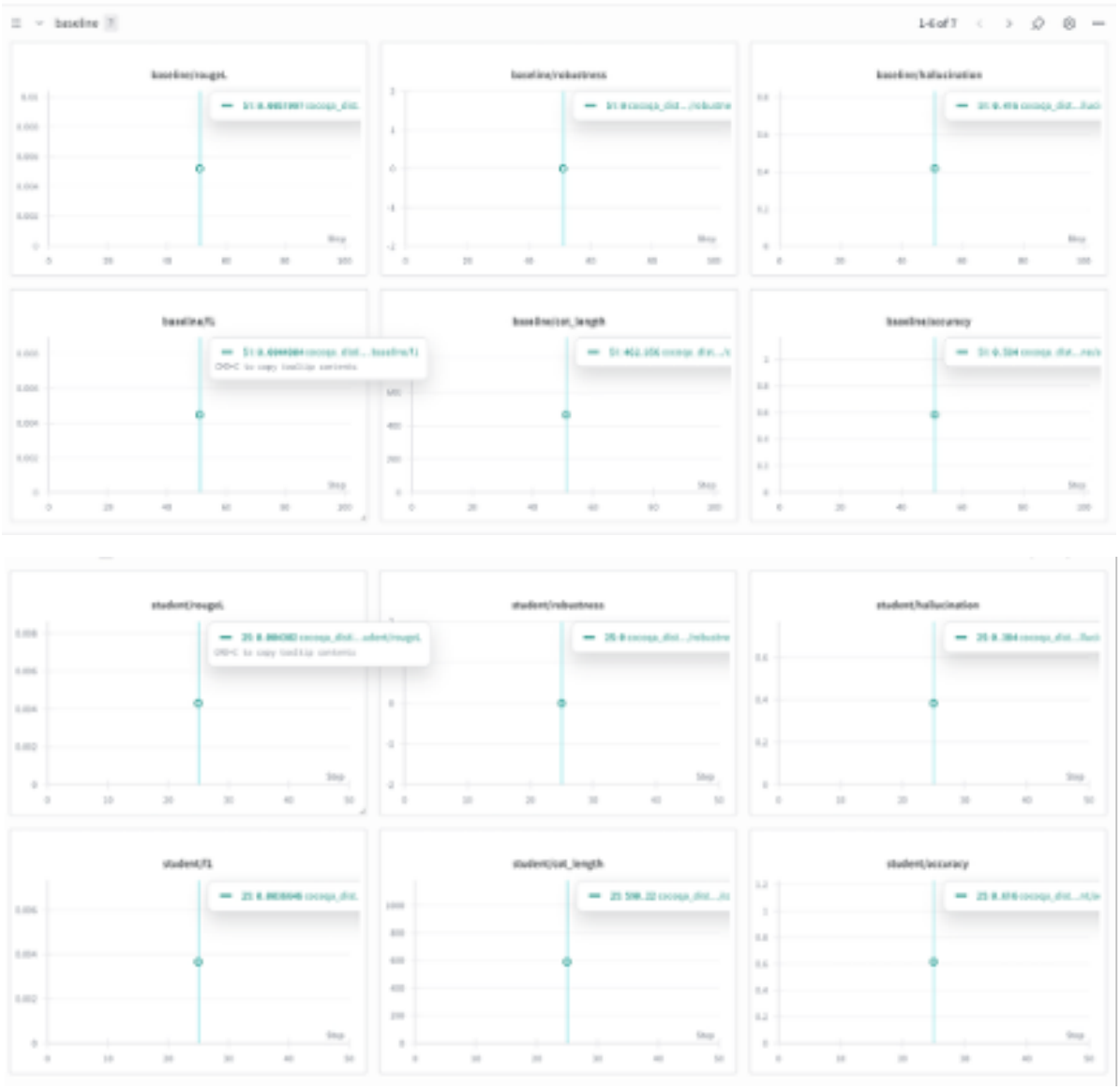


Figure 7: Running accuracy comparison and evaluation metrics. baseline vs. student running accuracy. detailed evaluation dashboard including hallucination, length, F1, and robustness metrics.

Across all five metrics, the distilled student model shows consistent improvements. Accuracy increases by +3.2

Although ROUGE-L remains similar, this is expected as longer and more structured CoTs deviate from surface N-gram matches while improving semantic quality.

In general, the distillation process enhances not only performance but also stability, grounding, and interpretability of student model reasoning.

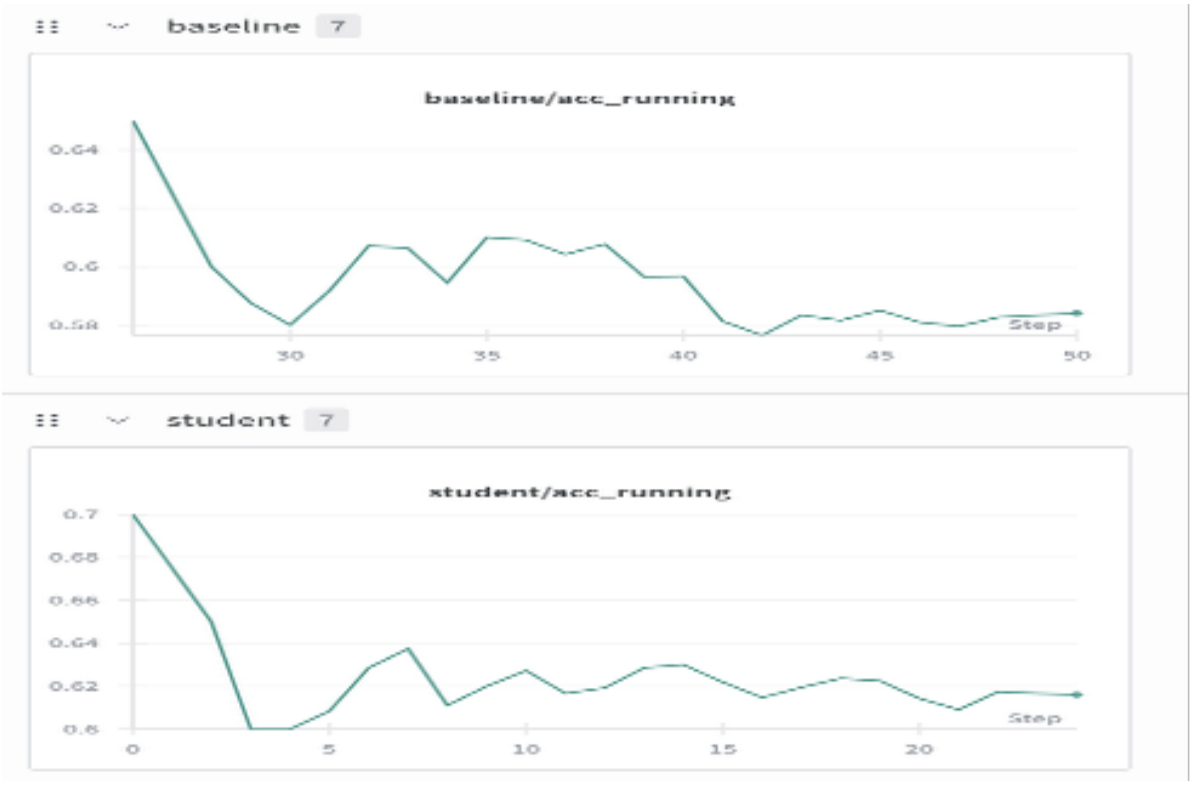


Figure 8: Student model and Baseline model dashboard.

A.5.4 Qualitative Findings

A review of 20 qualitative cases reveals two primary behavioral patterns.

Success pattern: Prior-driven probabilistic reasoning. The student model frequently infers correct answers by leveraging learned statistical priors. For example, when asked "Who is guiding the pony?", the student answers "A man," reflecting common co-occurrence patterns extracted from teacher CoT traces.

Failure pattern: Sensory resolution limits. Failures occur on questions requiring fine-grained visual cues such as textures, colors, or detailed object attributes. Because no visual data is available, such cases exceed the inherent capability of a blind text-only model.

A.5.5 Discussion

In summary, our experiments demonstrate that Visual Reasoning is not purely visual. By distilling the Chain-of-Thought process from a large Multimodal Teacher, we enabled a blind, text-only Student to capture high-level visual semantics. The 3.2This project theoretically demonstrates the Transferability of Visual Reasoning. Experiments prove that significant Visual Commonsense is encoded within linguistic structures. Via explicit CoT guidance, even a 0.5B blind model can unlock surprising multimodal understanding potential.

A.6. CoT Student D: HunyuanOCR Teacher Model

A.6.1 Approach

The teacher model used in this project is **tencent/HunyuanOCR**.[1] To incorporate domain-specific reasoning while preserving the general knowledge from the Sig-VL-2 student pre-training, we apply **LoRA** (Low-Rank Adaptation).

[1] See student-model description in the uploaded PDF.

A.6.2 Reducing Computational Load:

- LoRA rank: $R = 16$
- LoRA scaling factor: $\alpha \approx 16$ (consistent with typical configurations)
- student model size: ≈ 849M parameters
- Trainable LoRA parameters: ≈ 13M

This means the fine-tuning process updates only **1.5%** of the original model parameters. **98.5% reduction in compute cost**.

A.6.3 Training Dynamics

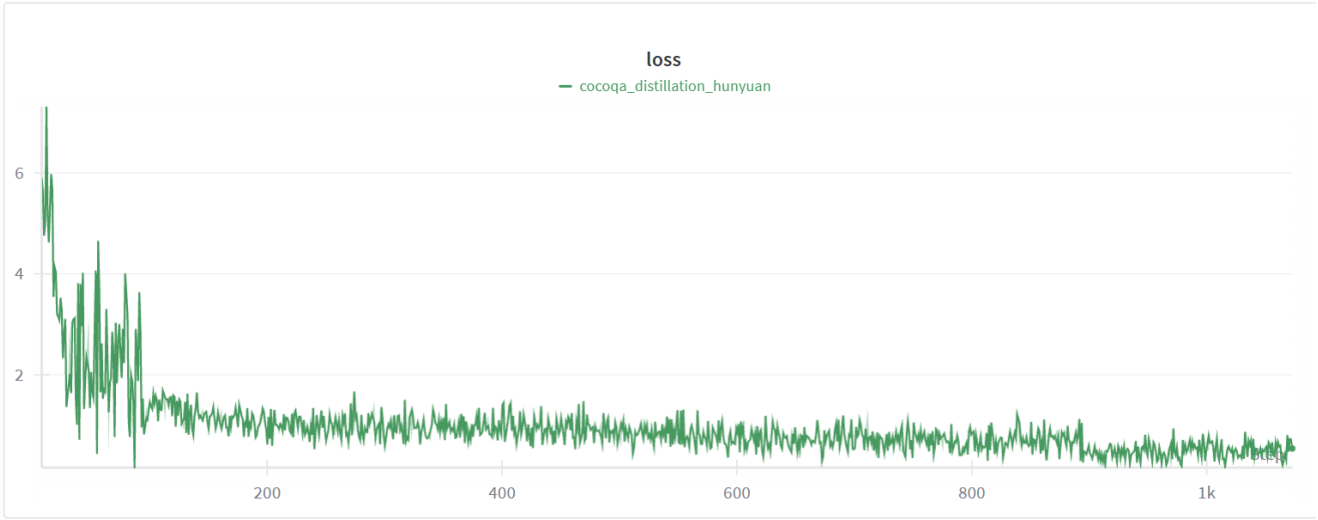


Figure 9: Loss curve for the LoRA-injected student model during training, showing rapid initial convergence over 1000 steps.

A.6.4 Evaluation:

Evaluation Before Fine-Tuning The evaluation results on 208 samples before fine-tuning were:

- **Average Loss:** 9.0831
- **Correct Rate:** 0.6971

Evaluation After Fine-Tuning The evaluation results on 209 samples after LoRA fine-tuning were:

- **Average Loss:** 8.8719
- **Correct Rate:** 0.7368