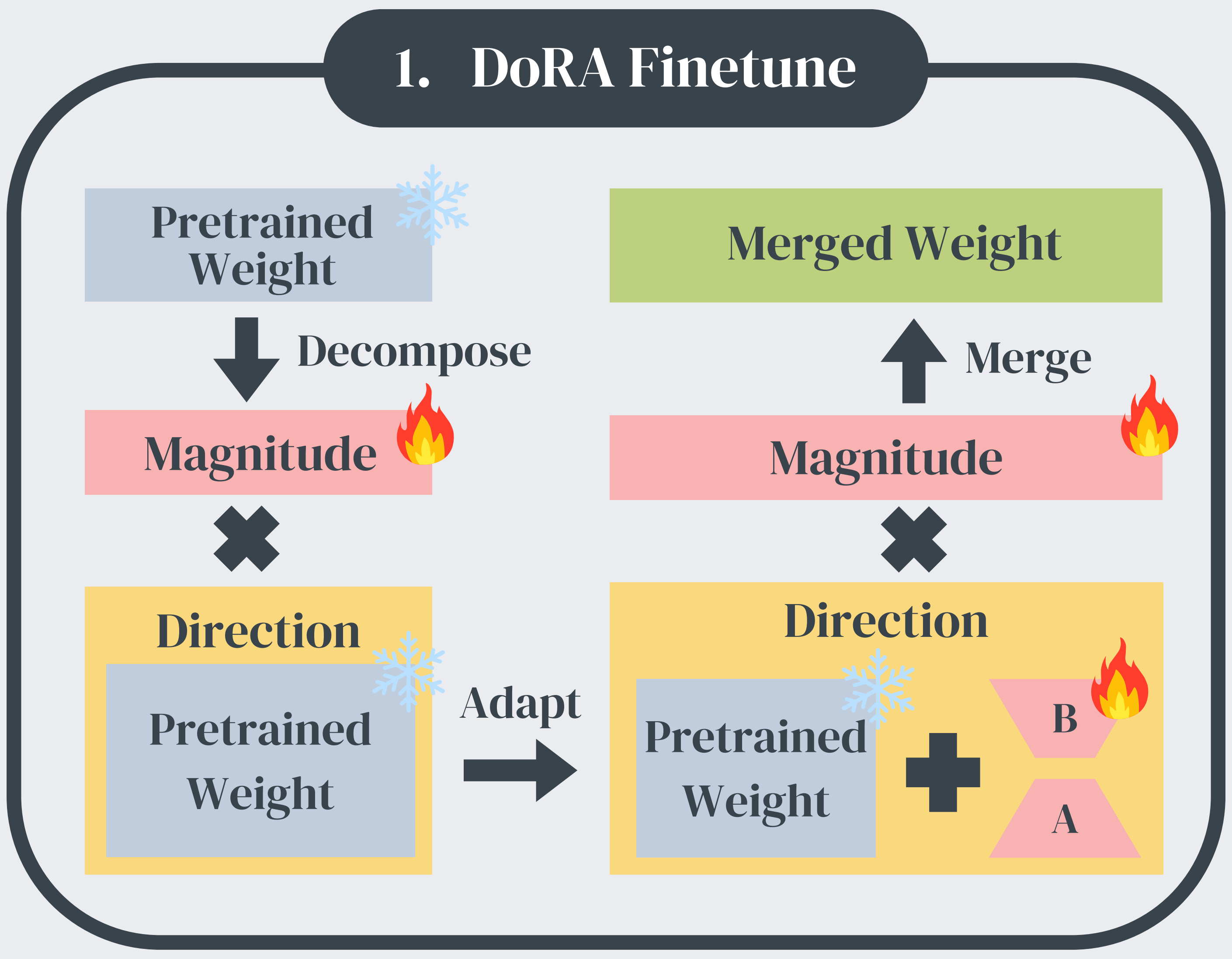
# Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

NTU DLCV 2024 FALL Team 10 XN

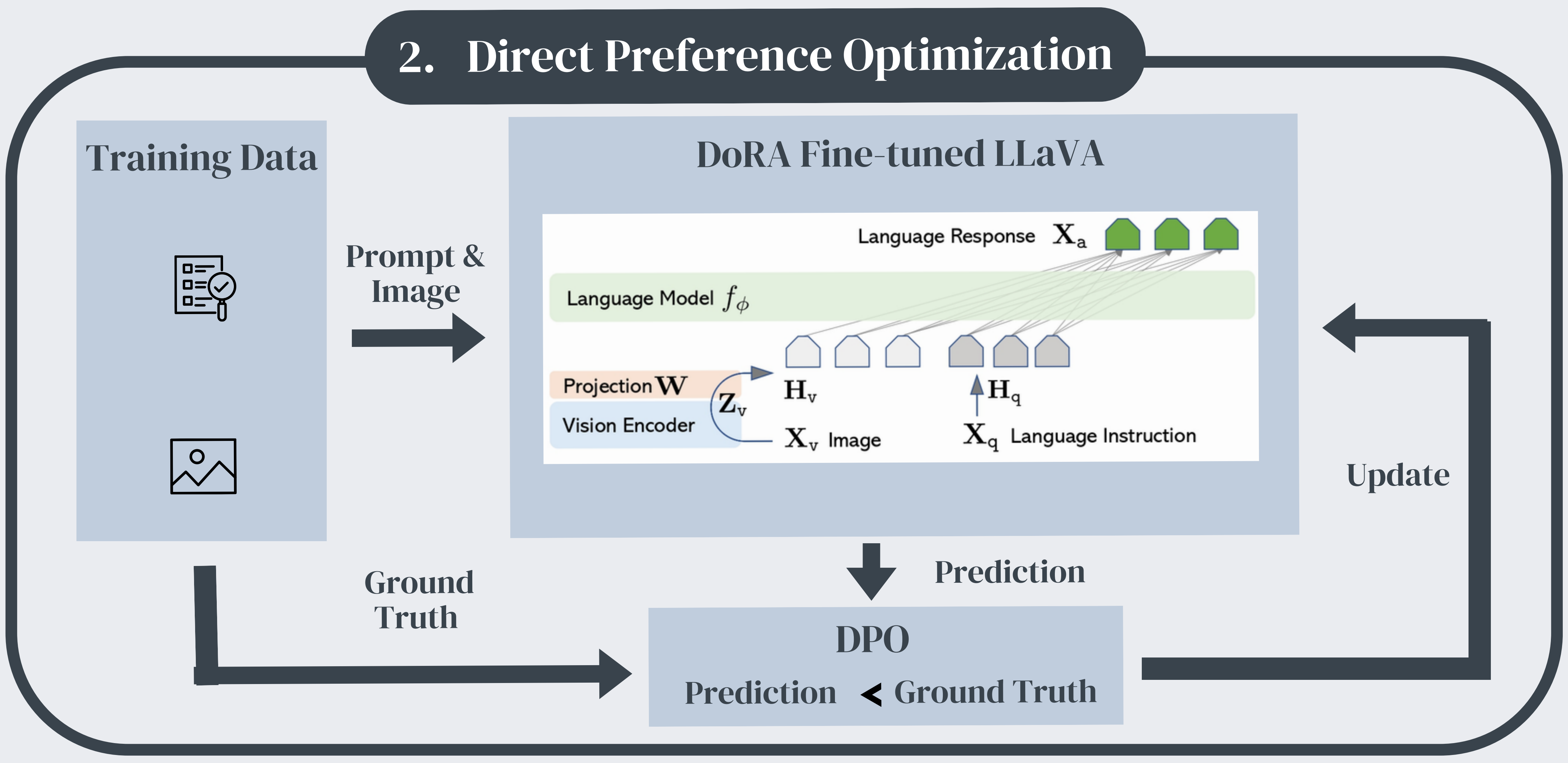
電機三 B11901039 林順文、電機三 B11901040 項達均、電機四 B10901098 蔡承恩、電機四 B10901183 賴奕銨

## Methodology

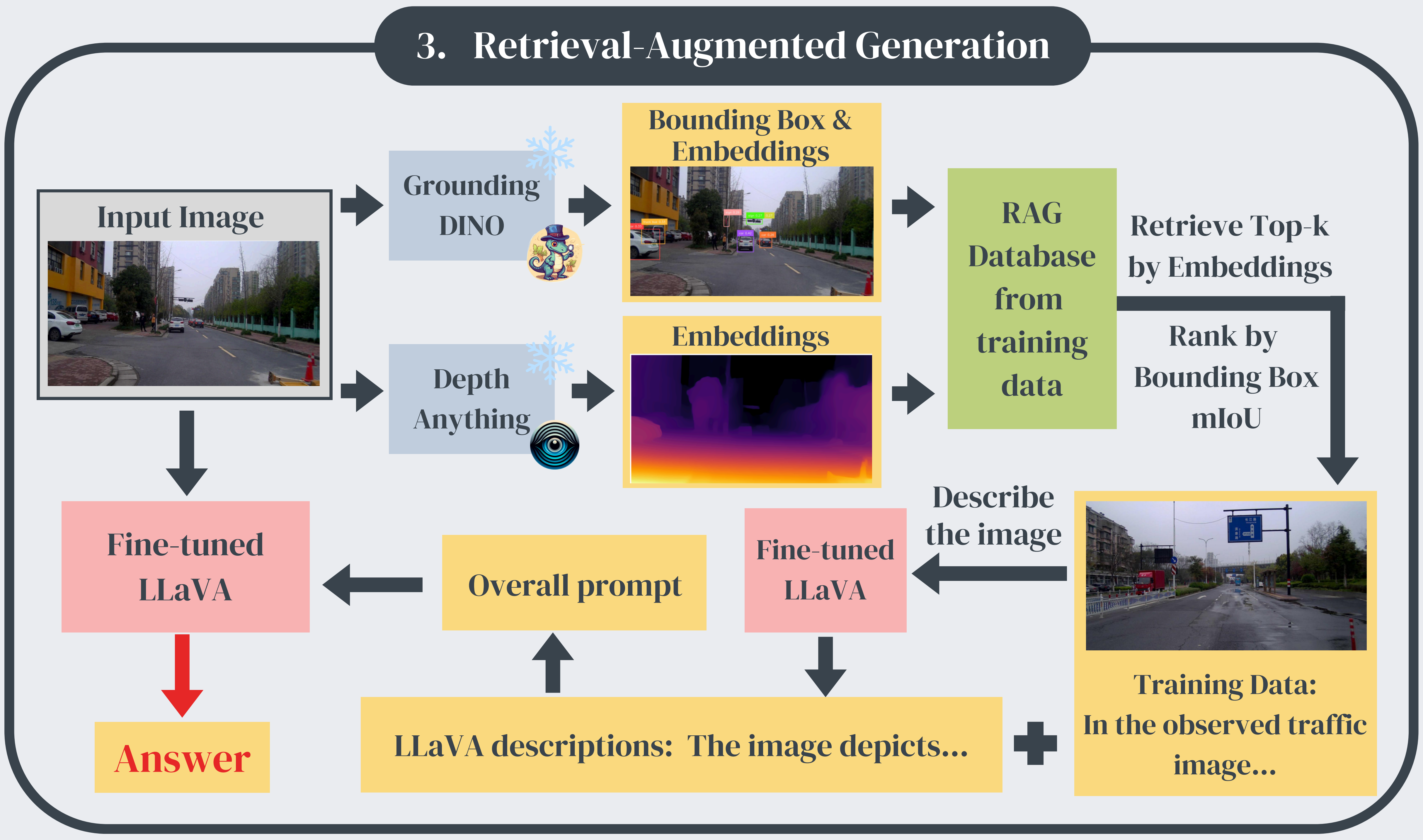
### 1. DoRA Finetune



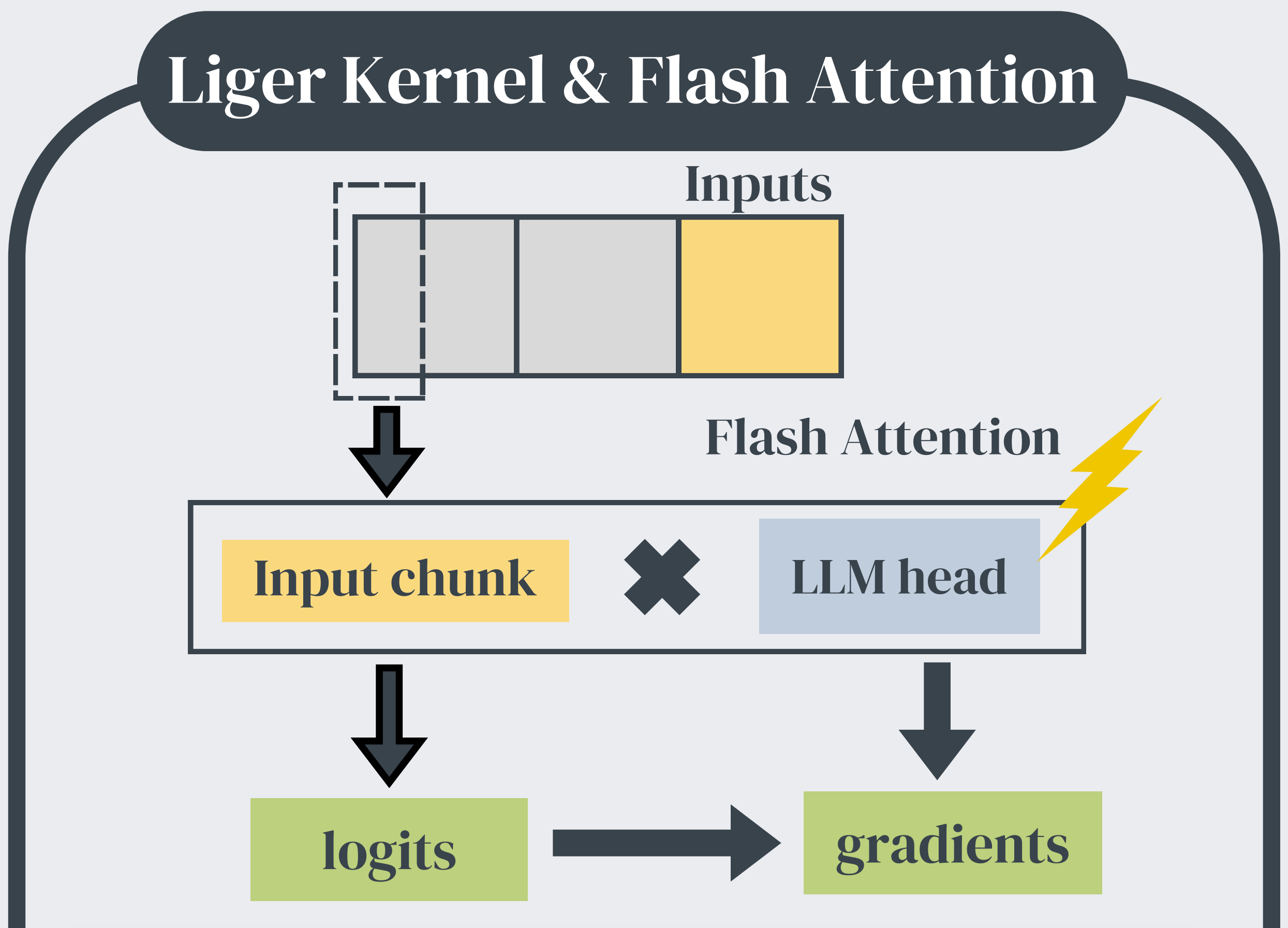
### 2. Direct Preference Optimization

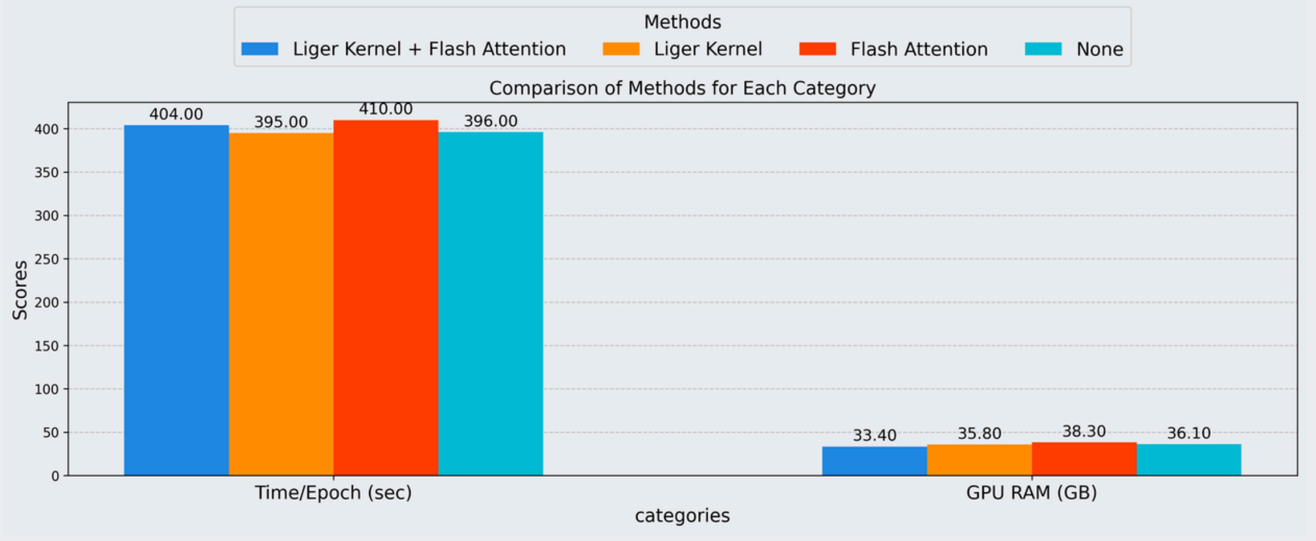


### 3. Retrieval-Augmented Generation


Input Image
Grounding DINO
Bounding Box & Embeddings
Depth Anything
Embeddings
RAG Database from training data
Retrieve Top-k by Embeddings
Rank by Bounding Box mIoU
Training Data: In the observed traffic image...
Describe the image
Fine-tuned LLaVA
Overall prompt
Fine-tuned LLaVA
LLaVA descriptions: The image depicts...
Answer

### Liger Kernel & Flash Attention

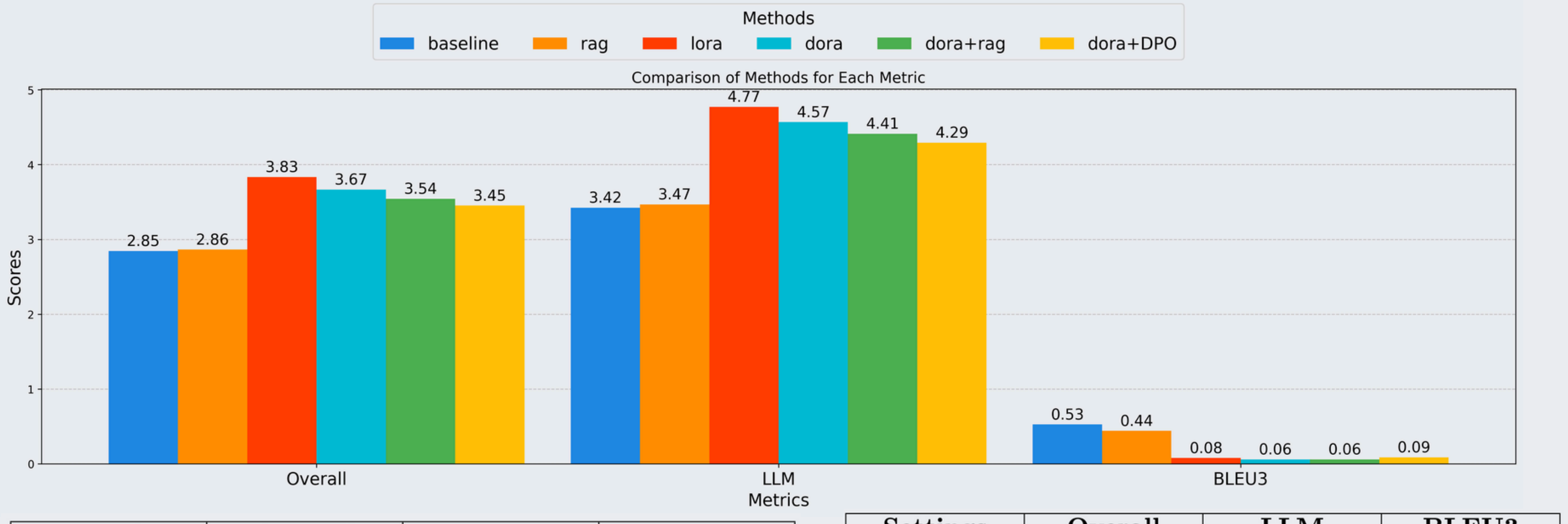




## Results

| Metric | baseline | RAG | LoRA | DoRA | DoRA+RAG | DoRA+DPO |
|---|---|---|---|---|---|---|
| Overall | 2.845 | 2.864 | 3.834 | 3.667 | 3.543 | 3.453 |
| LLM | 3.424 | 3.469 | 4.773 | 4.569 | 4.414 | 4.294 |
| BLEU3 | 0.526 | 0.442 | 0.078 | 0.059 | 0.060 | 0.087 |

Table 1: Model Performance Metrics



| max_token | Overall | LLM | BLEU3 |
|---|---|---|---|
| 250 | 3.661 | 4.560 | 0.066 |
| 300 | 3.667 | 4.569 | 0.059 |
| 500 | 3.051 | 3.781 | 0.131 |

Table 2: Different Max Token Settings

| Settings | Overall | LLM | BLEU3 |
|---|---|---|---|
| Rank = 8<br>alpha = 16 | 3.834 | 4.773 | 0.078 |
| Rank = 128<br>alpha = 256 | 2.833 | 3.520 | 0.085 |

Table 3: Different Rank and Alpha

## Key Features

- **DoRA-based Fine-Tuning:**
  - Adapts LLava for multimodal tasks with parameter-efficient fine-tuning.
  - Optimizes with directional updates.
- **Performance Optimization:**
  - **Liger Kernel:** Improves memory efficiency with fused linear cross-entropy.
  - **Flash Attention:** Speeds up memory read-write operations during attention computation.
- **Direct Preference Optimization:**
  - Refines DoRA fine-tuned LLava by training DPO with preference data comparing ground truth with LLava predictions.
- **Retrieval-Augmented Generation (RAG):**
  - Combines real-time data with learned knowledge to deliver actionable and context-aware suggestions.
  - **Depth Anything** for depth estimation and spatial understanding.
  - **Grounding DINO** for accurate object detection and scene comprehension.

## Analysis & Discussion

Our experiments showed that a max token length of 300 had the best results. The optimal token length of 300 provides sufficient context without being significantly longer than the ground truth. The superior performance of smaller rank and alpha values suggests that increasing r does not cover a more meaningful subspace, implying that a low-rank adaptation matrix is sufficient.

Even though the paper for DoRA showed that LoRA and DoRA performed comparatively in visual instruction tuning tasks, we found that LoRA outperforms DoRA in both LLM Score and BLEU. This is likely due to the stochastic nature of LLM grading, which results in inconsistent evaluations. We had observed varying LLM scores even when we submitted the same results.

Comparing the performance between DoRA and DoRA+DPO, DoRA+DPO performed slightly worse than DoRA, likely due to the variance between different submissions. Our training steps and data size are small for DPO training due to computational limitations, perhaps training with a larger dataset for more epochs will improve performance.