

---

03-논문 리뷰

# **Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization**

2025-07-16

연구실 스터디

---

---

03-논문 리뷰

# Enhancing the Reasoning Ability of **Multimodal Large Language Models** via Mixed Preference Optimization

2025-07-16

연구실 스터디

---

---

03-논문 리뷰

# Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization

2025-07-16

연구실 스터디

---

---

03-논문 리뷰

# Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization

2025-07-16

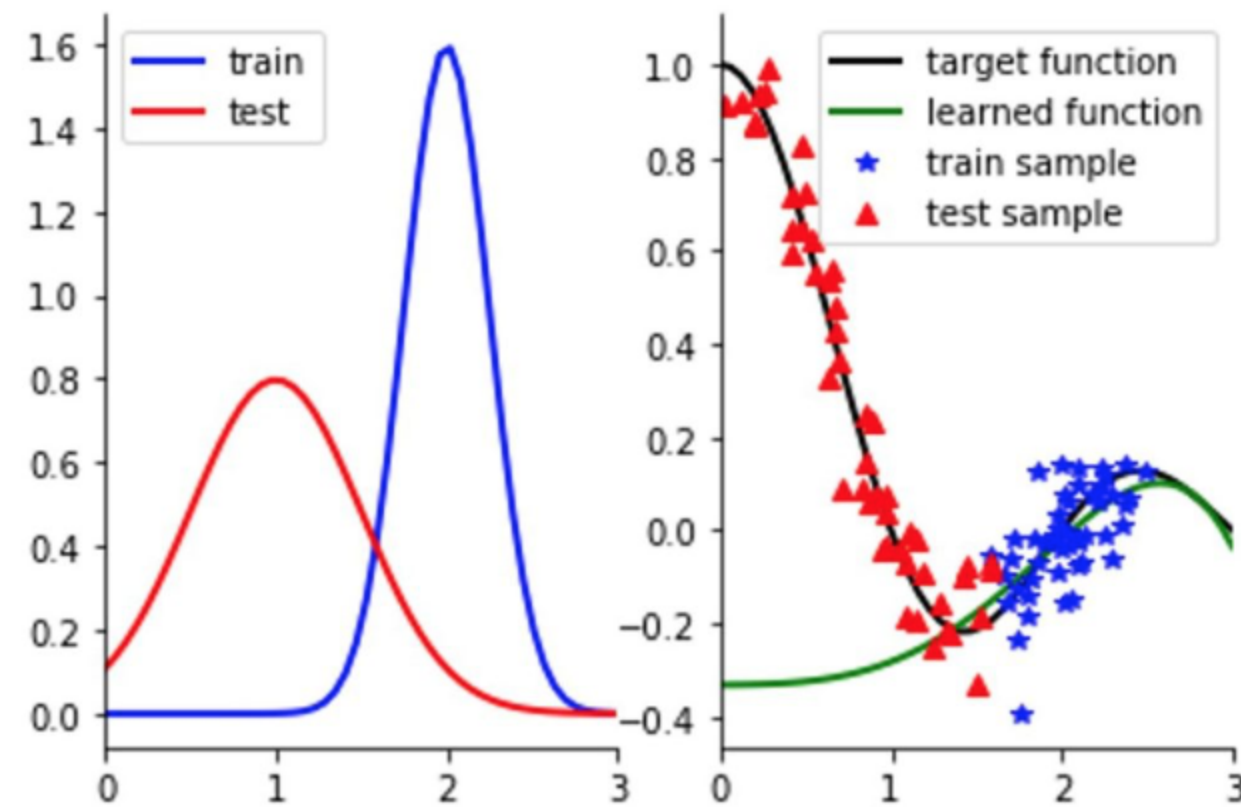
연구실 스터디

---

---

# Problem

## *Distribution Shift*

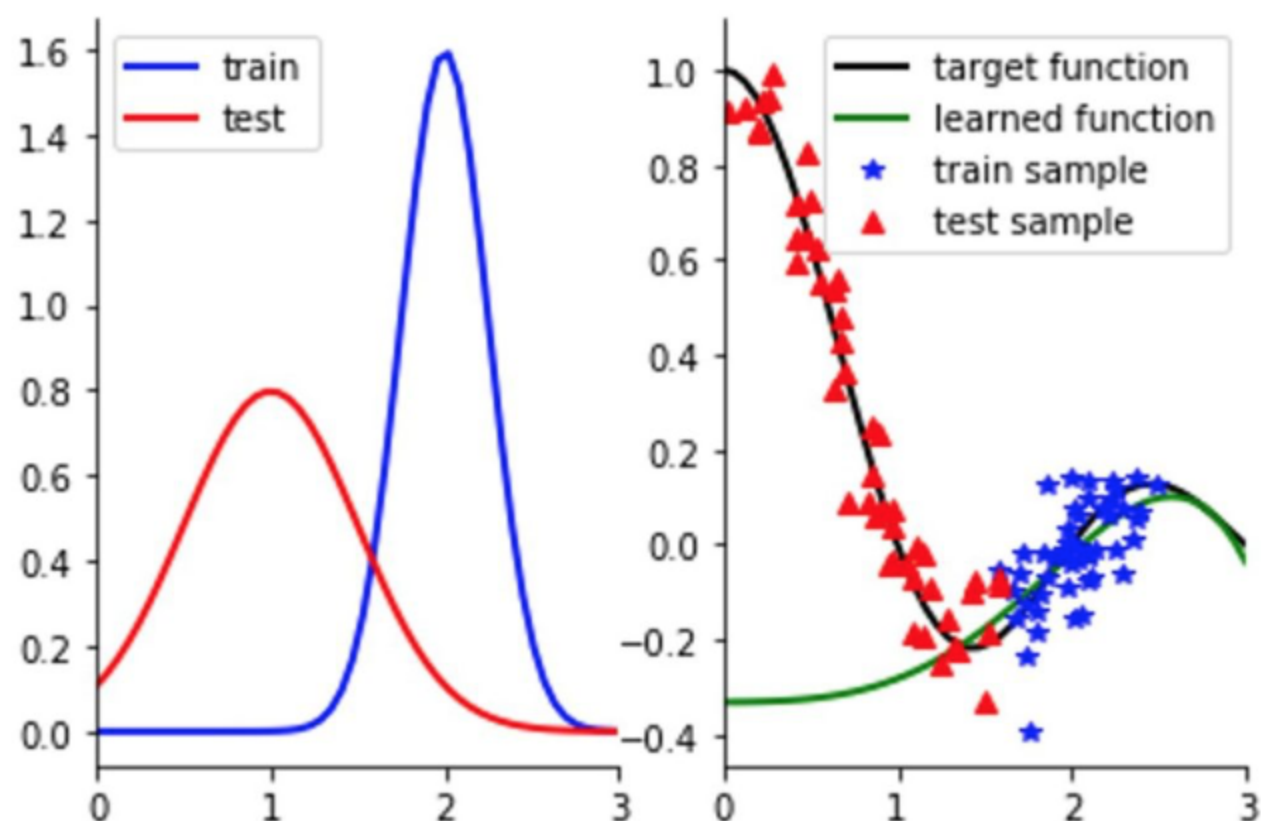


# Problem

## Distribution Shift



## SFT (Supervised Finetuning)



라벨된 정답 데이터로 학습(x, y) 쌍 필요

```
1 | {  
2 |   "prompt": "이순신 장군이 7년간 이룬 전과는 무엇인가요?",  
3 |   "response": "이순신 장군은 7년간 23전 23승의 전무후무한 기적을 이뤄냈습니다."  
4 | }
```

teacher forcing 방식

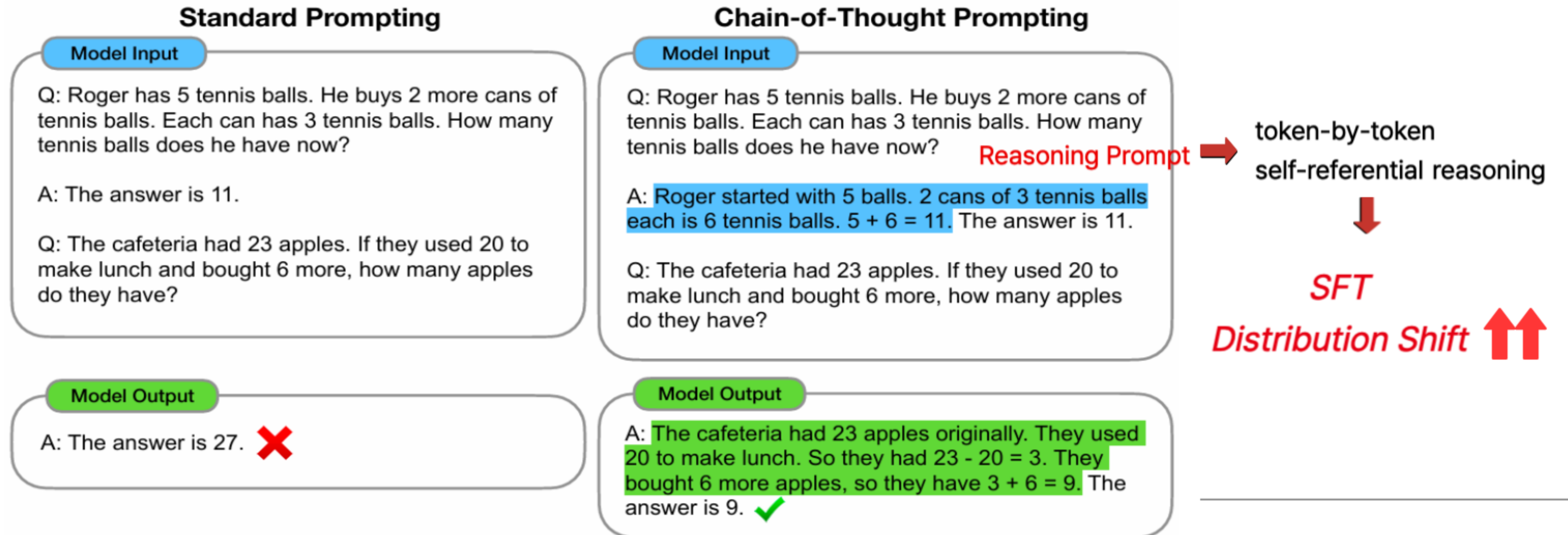
grouth truth response  $y^* = (y_1, y_2, \dots, y_T)$  매 시점  $t$ 마다  $y_t$ 를 예측  
입력  $\rightarrow$  이전 토큰은 항상 정답  $y_{<t}^*$ 로 주고 다음 토큰 예측

But inference에서는

이전 토큰은 모델이 스스로 예측 한 출력  $\hat{y}_{<t}$ 를 기반으로 다음 토큰 예측

# Problem

## Chain of Thoughts



# Solution

## Preference Optimization

DPO (Direct Preference Optimization) dataset

```
1 {  
2   "prompt": "이순신 장군이 7년간 이룬 전과는 무엇인가요?",  
3   "response_1": "이순신 장군은 7년간 23전 23승의 전무후무한 기적을 이뤄냈습니다.",  
4   "response_2": "이순신 장군은 7년간 15승 8패의 성적을 거두었습니다.",  
5   "preference": 0  
6 }
```

위 데이터에서 preference가 0인 경우, response\_1이 더 선호된다는 의미이다.

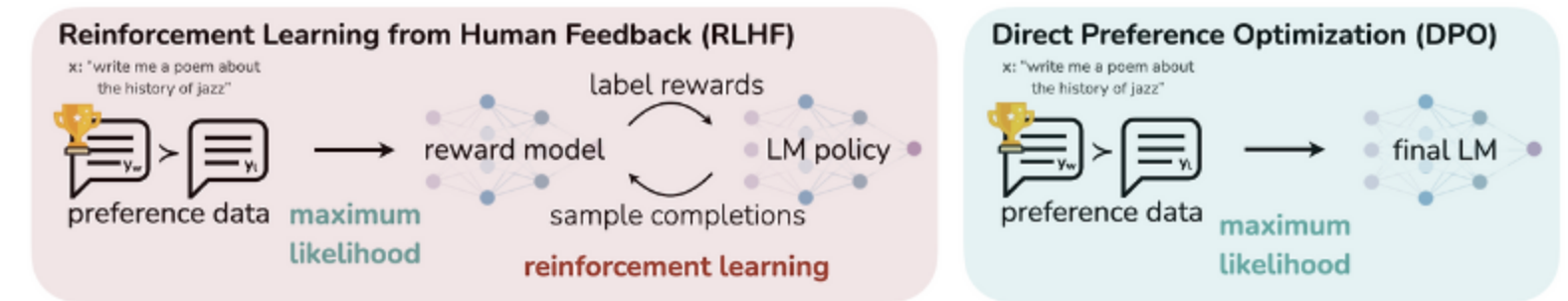
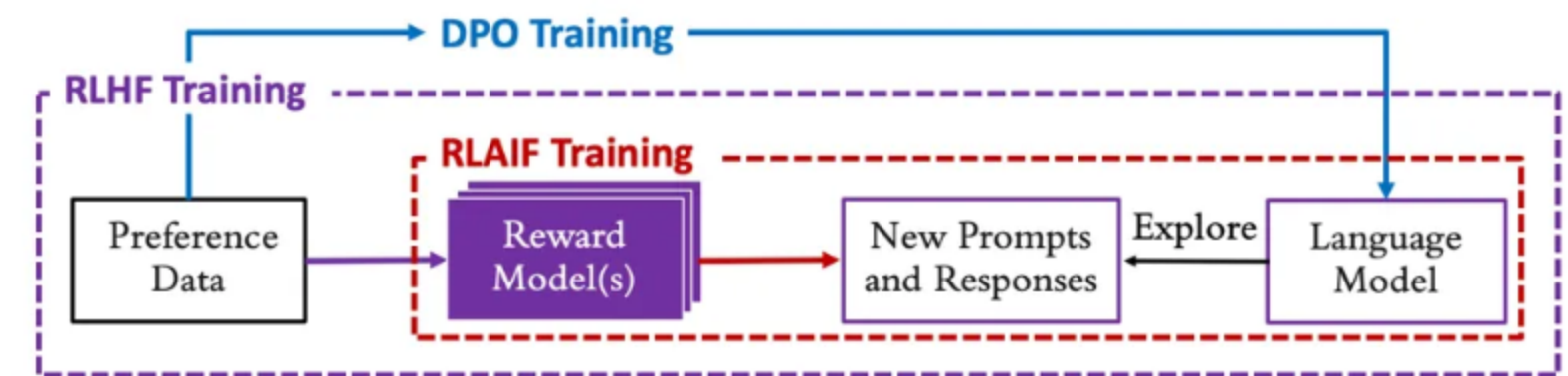


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes the learned reward. In contrast, DPO directly optimizes for the policy best satisfying the preferences with a simple classification objective, without an explicit reward function or RL.



RLHF의 복잡한 reward model + RL을 생략하고 preference pair만으로 alignment 학습을 가능하게 만든 방법

---

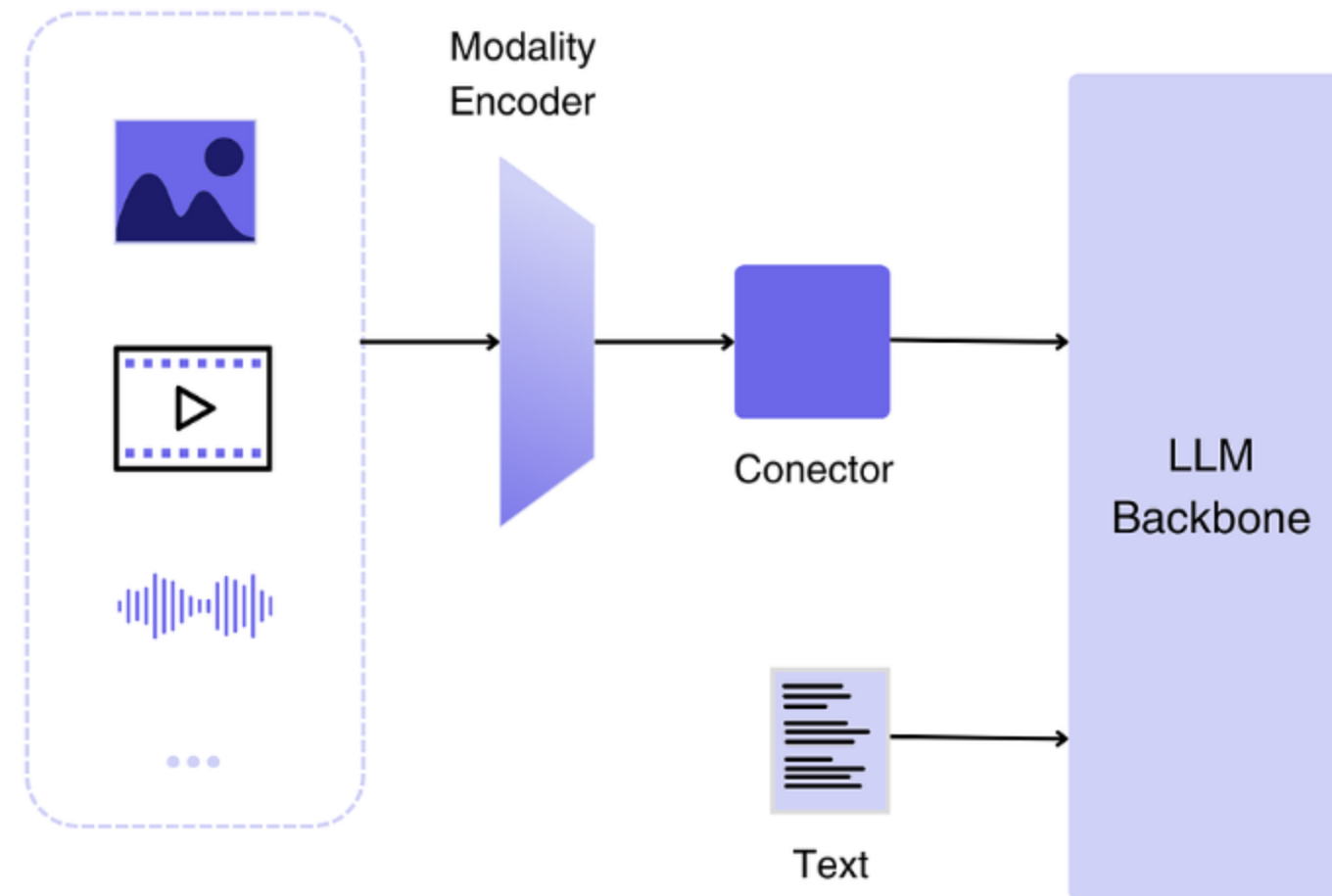
**Enhancing  
the Reasoning Ability of  
Multimodal Large Language Models  
via Mixed Preference Optimization**

---

---

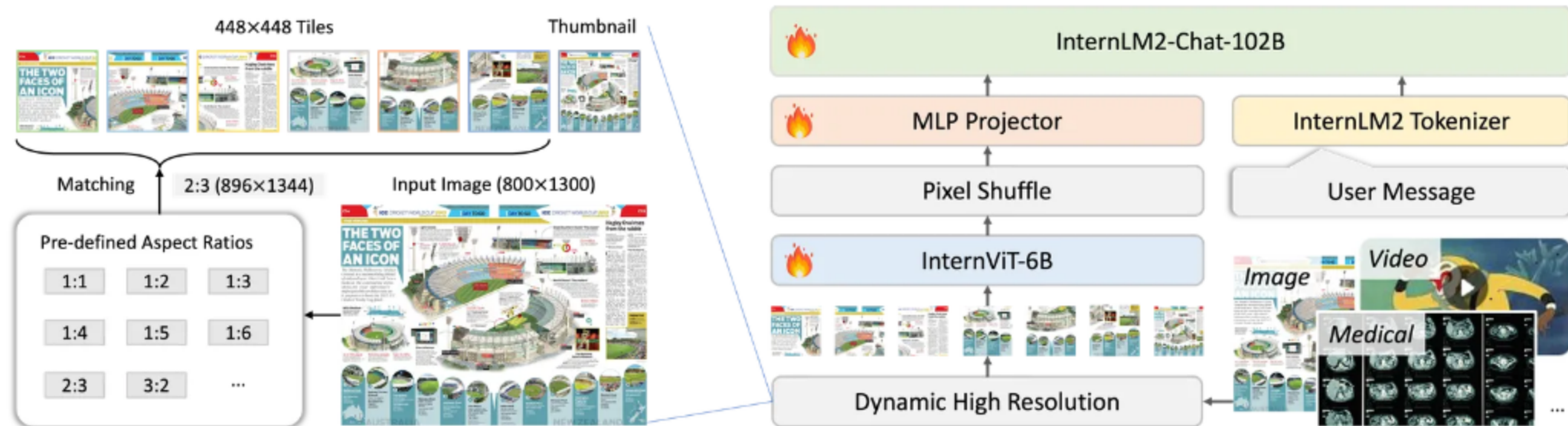
# MLLMs

## Multimodal Model Architecture



extend pre-trained LLMs + additional fusion layers for vision features →  
explorations into vision encoder-free architectures  
jointly processes visual and textual information  
without a separate encoder.

# MLLMs: InternVL2's Multimodal Architecture



cross-modal 연결 구조 : ViT + MLP + LLM

Vision Encoder : InternViT →

MLP Projector →

Language Model (LLM)- backbone

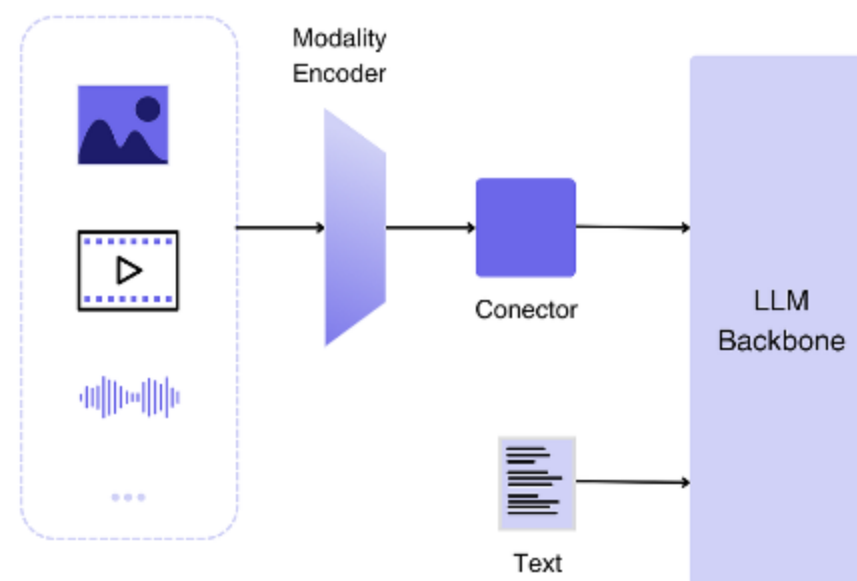


별도의 비전-언어 transformer 없이, MLP로 비전 → 언어 연결

Model Card

Name	InternVL2-2B	InternVL2-4B	InternVL2-8B	InternVL2-26B	InternVL2-40B	InternVL2-108B
Total	2.21B	4.15B	8.08B	25.51B	40.07B	108.70B
Model Size	ViT	304.01M	304.01M	304.01M	5.54B	5.54B
	MLP	12.60M	22.03M	33.57M	116.43M	143.17M
	LLM	2.21B	3.82B	7.74B	19.86B	34.39B
Resolution	dynamic resolution, max to 12 tiles of 448 × 448 in training, max to 40 tiles in testing (4K resolution).					

Multimodal Model Architecture



We extend the pre-training dataset used in InternVL 1.5 with data collected from diverse sources. These datasets span multiple tasks, including captioning, on Chinese images from cam data from uworld,

so data such as ology-VQA, PMC-recognize handwritten

---

# Experimental Design:

## Automated MMPR Data Pipeline and Mixed Preference Optimization

---

### 01 DATASET

automated preference data construction  
pipeline to create MMPR



**MMPR DATASET**

<https://huggingface.co/datasets/OpenGVLab/MMPR>

### 02 MODEL

Mixed Preference Optimization(MPO)



**InternVL2-8B-MPO**

<https://huggingface.co/OpenGVLab/InternVL2-8B-MPO>

---

---

## 01 DATASET

automated preference data construction  
pipeline to create MMRP



**MMRP DATASET**

---

# Scalable Multimodal Preference Dataset

## 01.Sampling the candidate response

## 02.Positive/Negative Labeling



$$y \sim M_0(y | x, I),$$

For instructions with clear ground truths

Prompt reasoning process

Final answer  
"Final Answer: \*\*\*"

$Y_p$ : chosen response  
ground truth와 일치하는 경우  
→ positive set  
 $Y_n$ : rejected response  
ground truth와 불일치  
명확한 final answer 없이 대답 끝남  
→ negative set

For instructions without clear ground truths

Response Without images

**Dropout NTP**

Prompt reasoning process  
*all positive set*

$$\tilde{y}_{\geq j} \sim M_0(\tilde{y}_{\geq j} | x, y_{< j})$$

$Y_p$ : chosen response  
 $y = [y_{< j}, y_{\geq j}]$   
 $Y_n$ : rejected response  
 $\tilde{y} = [y_{< j}, \tilde{y}_{\geq j}]$

what is clear ground truths?  
정답이 명확히 존재하는 instruction(지시문)들에 대해서는... 즉, 정답이 사람 기준으로 명확하게 평가될 수 있는 문제/과제에만 적용

# Scalable Multimodal Preference Dataset

## Instruction with clear ground truths



### Question:

Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: ..."

### Question:

What could be the reason for a person to have a brown teddy bear tied with a ribbon around its neck, a pack of colored pencils and a colorful patterned blanket?

- A. The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art
- B. The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art
- C. The person is a child who loves to play with teddy bears and draw with color pencils
- D. The person is a school teacher who uses these items to decorate their classroom for their students

### Chosen Response:

To determine the answer, let's analyze each option:

- A. The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art. This option assumes that the person collects multiple items as part of a fan interest in art. The items are not just limited to teddy bears, colored pencils, and blankets.
- B. The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art. This option suggests that the teddy bear was received as a gift and that the person uses colored pencils as they are a fan of art. This explanation is plausible.
- C. The person is a child who loves to play with teddy bears and draw with color pencils. **This option simply states that the person is a child, but it does not provide a clear connection between the other items like the vibrant blanket.**
- D. The person is a school teacher who uses these items to decorate their classroom for their students. This option seems less likely because the items (teddy bear, colored pencils, and blanket) might not be common classroom decoration items.

**Final answer: B**

### Rejected Response:

To determine the reason for a person to have a brown teddy bear tied with a ribbon around its neck, a pack of colored pencils, and a colorful patterned blanket, we need to consider the context and meaning behind these items.

- Option A: "The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art." This option suggests that the person has a specific interest in collecting items, which may not necessarily imply a connection to art.
- Option B: "The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art." This option suggests a personal connection to a gift while considering an individual interest in art.
- Option C: "The person is a child who loves to play with teddy bears and draw with color pencils." **This option assumes that the person is a child, which could be a plausible explanation for the items.**
- Option D: "The person is a school teacher who uses these items to decorate their classroom for their students." This option suggests a professional context rather than personal interest or collecting items.

Given the context of the items in the image and the most plausible explanation based on the presence of children's items and a child-friendly material like a teddy bear with a ribbon, the most likely explanation is:

**Final answer: C**

## Instruction without clear ground truths



### Question:

Where is the woman in the picture located?

### Chosen Response:

The woman in the picture is located in a public park. Here are the observations that support this conclusion:

1. Surroundings: The image shows a well-maintained area with grass, trees, and benches, which is typical of public parks. The presence of other people walking in the background also suggests a public space.
2. Lighting and Ambience: The bright daylight and open space are characteristic of outdoor environments like parks, rather than indoor locations like libraries.
3. Bench Design: *The style of the bench, which appears to be made of wood or metal with a decorative back, is common in public parks. Such benches are often placed along walking paths for public use.*
4. Vegetation: *The variety and arrangement of trees and shrubs look like planned landscaping, which is typical of parks rather than the natural, untamed growth found in a virgin forest. These elements collectively indicate that the setting is a public park.*

### Rejected Response:

The woman in the picture is located in a public park. Here are the observations that support this conclusion:

1. Surroundings: The image shows a well-maintained area with grass, trees, and benches, which is typical of public parks. The presence of other people walking in the background also suggests a public space.
  2. Lighting and Ambience: The bright daylight and open space are characteristic of outdoor environments like parks, rather than indoor locations like libraries.
  3. Bench: *If she were in a private garden or library, it would likely be more private and less likely to have communal spaces such as a bench.*
  4. Background Activity: ***The presence of other people in the background suggests a public space where people gather, such as a park.***
- Given these observations, the picture most likely shows a woman in a public park.***

# Multimodal Preference Dataset (instruction + image sample)

Task	Dataset
General VQA	VQAv2 [29], GQA [34], OKVQA [63], IconQA [59]
Science	AI2D [39], ScienceQA [60], M3CoT [16]
Chart	ChartQA [64], DVQA [37], MapQA [13]
Mathematics	GeoQA+ [12], CLEVR-Math [51], Geometry3K [58], GEOS [82], GeomVerse [38], Geo170K [27]
OCR	OCRvQA [68], InfoVQA [66], TextVQA [83], STVQA [8], SROIE [33]
Document	DocVQA [65]

Table 1. Datasets used to build our preference dataset. We collect images and instructions from various tasks to ensure the diversity of our dataset.

Open-ended sample: 정답이 한가지로 제한되지 않는 열린 질문  
 correctness-based pipeline: 모델 답변의 정확성(ground truth와 일치)을 기준으로 positive/negative label을 붙여 학습에 사용하는 방식  
 ( VQA와 document 제외)

### General VQA



**Question:** What are the toppings on the pizza?

**Chosen Response:**  
The pizza in the image has the following toppings:  
1. Tomatoes: There are halved cherry tomatoes scattered across the pizza, adding a fresh and juicy element.  
2. Artichokes: Pieces of artichoke hearts are visible, providing a tender and slightly tangy flavor.  
3. Cheese: A layer of melted cheese covers the pizza, likely mozzarella or a similar type providing a creamy and rich texture.  
4. Pesto or Herb Sauce: There appears to be a green sauce, possibly pesto, spread beneath the toppings, adding a herby and savory flavor.  
The pizza has a rustic appearance with a slightly charred crust, suggesting it was baked at high temperatures.


### Science



**Question:** Two magnets are placed as shown. Hint: Magnets that attract pull together. Magnets that repel A. attract B. repel

**Chosen Response:**  
In the image, there are two magnets. Each magnet has two poles labeled as "N" for North or "S" for South.  
Magnetic Interaction:  
- Like Poles Repel: If similar poles (North-North or South-South) face each other, they repel.  
- Unlike Poles Attract: If opposite poles (North-South) face each other, they attract.  
Analysis of the Image:  
1. Left Magnet: The North pole is on the left, and the South pole is on the right.  
2. Right Magnet: The North pole is on the left, and the South pole is on the right.  
Conclusion:  
- The North pole of the left magnet is facing the North pole of the right magnet.  
- The South pole of the left magnet is facing the South pole of the right magnet.  
Conclusion: Since both pairs of like poles (North-North and South-South) are facing each other, the magnets will repel each other.  
Therefore, the correct answer is B. repel

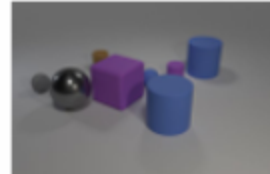
### Chart



**Question:** How many color does the graph shows?

**Chosen Response:**  
The graph uses five distinct colors, each representing a different country. Here's detailed breakdown:  
1. Blue - Represents Brazil, with an internet access percentage of 49.2%.  
2. Purple - Represents Latvia, with an internet access percentage of 70%.  
3. Green - Represents Austria, with an internet access percentage of 82.4%.  
4. Red - Represents France, with an internet access percentage of 82.4%.  
5. Orange - Represents Norway, with an internet access percentage of 96.0%.  
Each color is used to differentiate the countries and their corresponding internet access percentages for 2015.


### Mathematics



**Question:** Subtract all balls. How many objects are left?

**Chosen Response:**  
To determine how many objects are left after subtracting all the balls, let's examine the image and identify each object:  
1. Balls: There are three spherical objects (balls):  
- A small grey ball on the far left.  
- A shiny metallic ball next to the small grey ball.  
- A medium-sized blue ball near the center.  
2. Other Objects:  
- A purple cube, which is a square-shaped object.  
- A large blue cylinder in the foreground.  
- A small orange cylinder behind the purple cube.  
- A medium-sized blue cylinder in the background.  
Total Objects: Total objects initially present = 3 (balls) + 3 non-ball objects.  
Subtracting the Balls: Removing the 3 balls leaves us with the non-ball objects.  
Remaining Objects: 3 purple cube, 3 cylinders (1 large blue, 1 small orange, 1 medium blue).  
Therefore, after subtracting all the balls, there are 3 objects left.

### OCR



**Question:** What is the name of the event?

**Chosen Response:**  
The event is part of the "RBC Summer Kids Staff." It includes two programs:  
1. Camp Killeenjaque: This is a Vacation Bible School (VBS) program for children in Kindergarten through 5th grade. It is themed as "An Epic Expedition Through Proverbs."  
2. Cross Training: This is for junior-high students in 6th through 8th grade, organized by Ravick Report Church.  
The event is scheduled from June 22 to June 26, with activities running from 8:45 AM to 12:00 PM. Registration is open, and the cost is free.

### Document

**Question:** What is the name of entity mentioned in the given form?

**Chosen Response:**  
The form is a "Principal Investigator's Statement of Economic Interest" related to research funding. It details information about a research project led by Paul Salzman from the Biology department at the University of California, San Diego.  
The specific section of interest is Section A, which requires details about the non-governmental entity providing funding for the research. According to the form:  
- Name of Entity: *Proctor & Gamble*  
- Address of Entity: *Cincinnati, Ohio*  
- Principal Business of Entity: *Household products and foods*  
- Amount of Funding: *\$25,000.00 (if total)*  
This section is crucial as it discloses the financial backing from a non-governmental source, which is essential for transparency and compliance in research funding.

**Rejected Response:**  
The form is a "Principal Investigator's Statement of Economic Interest" related to research funding. It details information about a research project led by Paul Salzman from the Biology department at the University of California, San Diego.  
The specific section of interest is Section A, which requires details about the non-governmental entity providing funding for the project. The information in this section pertains to the funding organization, their role, and the type of funds provided. The organization mentioned in the form is the *Cystic Fibrosis Foundation R (CFP R)*.  
The CFP R provides grants to support innovative and promising research projects in the field of cystic fibrosis, a genetic disorder affecting the lungs and digestive system. The funding primarily supports research aimed at improving patient outcomes, disease understanding, and treatment. In this case, Dr. Salzman's project aims to explore the molecular mechanisms underlying how development during the fetal stages, potentially connecting them to cystic fibrosis and other conditions.

### Science



**Question:** Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: ..."

**Chosen Response:**  
To determine which term matches the picture, let's analyze the characteristics of each type of growth:  
1. Exponential Growth: This type of growth occurs when a population inhabits an environment with abundant resources and few limiting factors. In this case, the population grows increasingly rapidly, leading to a J-shaped curve. The growth is rapid and unbounded over time.  
2. Logistic Growth: This type of growth occurs when resources are scarce or a population faces considerable limiting factors, such as predators. In this scenario, the population rises steeply at first but then levels off as it reaches the carrying capacity, which is the environment's maximum sustainable population level. The growth curve resembles an S-shaped curve; it increases rapidly, reaches a limit, and then stabilizes.  
Based on the description and the shape of the graph, the term that matches the picture is:  
Final answer: B

**Rejected Response:**  
To determine which term matches the picture, let's analyze the characteristics of each type of growth:  
1. Exponential Growth: This type of growth occurs when a population inhabits an environment with abundant resources and few limiting factors. In this case, the population grows increasingly rapidly, leading to a J-shaped curve. The growth is rapid and unbounded over time.  
2. Logistic Growth: This type of growth occurs when resources are scarce or a population faces considerable limiting factors, such as predators. In this scenario, the population rises steeply at first but then levels off as it reaches the carrying capacity, which is the environment's maximum sustainable population level. The growth curve resembles an S-shaped curve; it increases rapidly, reaches a limit, and then stabilizes.  
The graph in the picture shows a population that grows rapidly at first and then levels off, resembling an S-shaped curve. This is characteristic of logistic growth.  
Final answer: A

### Chart




**Question:** Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: ..."

**Chosen Response:**  
To determine the value of the "total" category in the "times" row, we need to look at the color-coded legend and the corresponding values for each category in the "times" row:  
1. The "times" row has the following categories: large blue - index green - engine red - total purple.  
2. According to the legend, the value for "total" is represented by the purple color.  
3. By examining the "times" row, we can see that the purple square is in the "total" column.  
4. The purple square is aligned with the value on the right side of the image, which corresponds to 10.  
Therefore, the value of the "total" in the "times" row is 10.  
Final answer: 1000

**Rejected Response:**  
To determine the value of the total in times, we need to analyze the data given for each category in the "times" row:  
1. large: The value is 10<sup>3</sup> (one thousand).  
2. index: The value is 10<sup>4</sup> (ten thousand).  
3. engine: The value is 10<sup>5</sup> (one hundred thousand).  
4. total: The value is not directly given, but we can infer it by summing the values of large, index, and engine.  
Let's calculate the total: total = large + index + engine = 10<sup>3</sup> + 10<sup>4</sup> + 10<sup>5</sup> = 10<sup>3</sup> + 10<sup>3</sup> \* 10 + 10<sup>3</sup> \* 100 = 10<sup>3</sup> \* (1 + 10 + 100) = 10<sup>3</sup> \* 111 = 111,000.  
Therefore, the value of the "total" in the "times" row is 111,000.  
Final answer: 1000

### OCR



**Question:** Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: ..."

**Chosen Response:**  
To find the total amount on the receipt, let's break down the expenses listed on the receipt:  
1. Chicken Karaage (fried chicken): 12.00  
2. Ono Curry Fried Rice Set (Set): 9.00  
3. Kani Don (Fried Rice with Tuna): 9.00  
4. Kani Don (Fried Rice with Tuna): 9.00  
5. Kani Don (Fried Rice with Tuna): 9.00  
6. Kani Don (Fried Rice with Tuna): 9.00  
7. Kani Don (Fried Rice with Tuna): 9.00  
8. Kani Don (Fried Rice with Tuna): 9.00  
9. Kani Don (Fried Rice with Tuna): 9.00  
10. Kani Don (Fried Rice with Tuna): 9.00  
11. Kani Don (Fried Rice with Tuna): 9.00  
12. Kani Don (Fried Rice with Tuna): 9.00  
13. Kani Don (Fried Rice with Tuna): 9.00  
14. Kani Don (Fried Rice with Tuna): 9.00  
15. Kani Don (Fried Rice with Tuna): 9.00  
16. Kani Don (Fried Rice with Tuna): 9.00  
17. Kani Don (Fried Rice with Tuna): 9.00  
18. Kani Don (Fried Rice with Tuna): 9.00  
19. Kani Don (Fried Rice with Tuna): 9.00  
20. Kani Don (Fried Rice with Tuna): 9.00  
21. Kani Don (Fried Rice with Tuna): 9.00  
22. Kani Don (Fried Rice with Tuna): 9.00  
23. Kani Don (Fried Rice with Tuna): 9.00  
24. Kani Don (Fried Rice with Tuna): 9.00  
25. Kani Don (Fried Rice with Tuna): 9.00  
26. Kani Don (Fried Rice with Tuna): 9.00  
27. Kani Don (Fried Rice with Tuna): 9.00  
28. Kani Don (Fried Rice with Tuna): 9.00  
29. Kani Don (Fried Rice with Tuna): 9.00  
30. Kani Don (Fried Rice with Tuna): 9.00  
31. Kani Don (Fried Rice with Tuna): 9.00  
32. Kani Don (Fried Rice with Tuna): 9.00  
33. Kani Don (Fried Rice with Tuna): 9.00  
34. Kani Don (Fried Rice with Tuna): 9.00  
35. Kani Don (Fried Rice with Tuna): 9.00  
36. Kani Don (Fried Rice with Tuna): 9.00  
37. Kani Don (Fried Rice with Tuna): 9.00  
38. Kani Don (Fried Rice with Tuna): 9.00  
39. Kani Don (Fried Rice with Tuna): 9.00  
40. Kani Don (Fried Rice with Tuna): 9.00  
41. Kani Don (Fried Rice with Tuna): 9.00  
42. Kani Don (Fried Rice with Tuna): 9.00  
43. Kani Don (Fried Rice with Tuna): 9.00  
44. Kani Don (Fried Rice with Tuna): 9.00  
45. Kani Don (Fried Rice with Tuna): 9.00  
46. Kani Don (Fried Rice with Tuna): 9.00  
47. Kani Don (Fried Rice with Tuna): 9.00  
48. Kani Don (Fried Rice with Tuna): 9.00  
49. Kani Don (Fried Rice with Tuna): 9.00  
50. Kani Don (Fried Rice with Tuna): 9.00  
51. Kani Don (Fried Rice with Tuna): 9.00  
52. Kani Don (Fried Rice with Tuna): 9.00  
53. Kani Don (Fried Rice with Tuna): 9.00  
54. Kani Don (Fried Rice with Tuna): 9.00  
55. Kani Don (Fried Rice with Tuna): 9.00  
56. Kani Don (Fried Rice with Tuna): 9.00  
57. Kani Don (Fried Rice with Tuna): 9.00  
58. Kani Don (Fried Rice with Tuna): 9.00  
59. Kani Don (Fried Rice with Tuna): 9.00  
60. Kani Don (Fried Rice with Tuna): 9.00  
61. Kani Don (Fried Rice with Tuna): 9.00  
62. Kani Don (Fried Rice with Tuna): 9.00  
63. Kani Don (Fried Rice with Tuna): 9.00  
64. Kani Don (Fried Rice with Tuna): 9.00  
65. Kani Don (Fried Rice with Tuna): 9.00  
66. Kani Don (Fried Rice with Tuna): 9.00  
67. Kani Don (Fried Rice with Tuna): 9.00  
68. Kani Don (Fried Rice with Tuna): 9.00  
69. Kani Don (Fried Rice with Tuna): 9.00  
70. Kani Don (Fried Rice with Tuna): 9.00  
71. Kani Don (Fried Rice with Tuna): 9.00  
72. Kani Don (Fried Rice with Tuna): 9.00  
73. Kani Don (Fried Rice with Tuna): 9.00  
74. Kani Don (Fried Rice with Tuna): 9.00  
75. Kani Don (Fried Rice with Tuna): 9.00  
76. Kani Don (Fried Rice with Tuna): 9.00  
77. Kani Don (Fried Rice with Tuna): 9.00  
78. Kani Don (Fried Rice with Tuna): 9.00  
79. Kani Don (Fried Rice with Tuna): 9.00  
80. Kani Don (Fried Rice with Tuna): 9.00  
81. Kani Don (Fried Rice with Tuna): 9.00  
82. Kani Don (Fried Rice with Tuna): 9.00  
83. Kani Don (Fried Rice with Tuna): 9.00  
84. Kani Don (Fried Rice with Tuna): 9.00  
85. Kani Don (Fried Rice with Tuna): 9.00  
86. Kani Don (Fried Rice with Tuna): 9.00  
87. Kani Don (Fried Rice with Tuna): 9.00  
88. Kani Don (Fried Rice with Tuna): 9.00  
89. Kani Don (Fried Rice with Tuna): 9.00  
90. Kani Don (Fried Rice with Tuna): 9.00  
91. Kani Don (Fried Rice with Tuna): 9.00  
92. Kani Don (Fried Rice with Tuna): 9.00  
93. Kani Don (Fried Rice with Tuna): 9.00  
94. Kani Don (Fried Rice with Tuna): 9.00  
95. Kani Don (Fried Rice with Tuna): 9.00  
96. Kani Don (Fried Rice with Tuna): 9.00  
97. Kani Don (Fried Rice with Tuna): 9.00  
98. Kani Don (Fried Rice with Tuna): 9.00  
99. Kani Don (Fried Rice with Tuna): 9.00  
100. Kani Don (Fried Rice with Tuna): 9.00  
101. Kani Don (Fried Rice with Tuna): 9.00  
102. Kani Don (Fried Rice with Tuna): 9.00  
103. Kani Don (Fried Rice with Tuna): 9.00  
104. Kani Don (Fried Rice with Tuna): 9.00  
105. Kani Don (Fried Rice with Tuna): 9.00  
106. Kani Don (Fried Rice with Tuna): 9.00  
107. Kani Don (Fried Rice with Tuna): 9.00  
108. Kani Don (Fried Rice with Tuna): 9.00  
109. Kani Don (Fried Rice with Tuna): 9.00  
110. Kani Don (Fried Rice with Tuna): 9.00  
111. Kani Don (Fried Rice with Tuna): 9.00  
112. Kani Don (Fried Rice with Tuna): 9.00  
113. Kani Don (Fried Rice with Tuna): 9.00  
114. Kani Don (Fried Rice with Tuna): 9.00  
115. Kani Don (Fried Rice with Tuna): 9.00  
116. Kani Don (Fried Rice with Tuna): 9.00  
117. Kani Don (Fried Rice with Tuna): 9.00  
118. Kani Don (Fried Rice with Tuna): 9.00  
119. Kani Don (Fried Rice with Tuna): 9.00  
120. Kani Don (Fried Rice with Tuna): 9.00  
121. Kani Don (Fried Rice with Tuna): 9.00  
122. Kani Don (Fried Rice with Tuna): 9.00  
123. Kani Don (Fried Rice with Tuna): 9.00  
124. Kani Don (Fried Rice with Tuna): 9.00  
125. Kani Don (Fried Rice with Tuna): 9.00  
126. Kani Don (Fried Rice with Tuna): 9.00  
127. Kani Don (Fried Rice with Tuna): 9.00  
128. Kani Don (Fried Rice with Tuna): 9.00  
129. Kani Don (Fried Rice with Tuna): 9.00  
130. Kani Don (Fried Rice with Tuna): 9.00  
131. Kani Don (Fried Rice with Tuna): 9.00  
132. Kani Don (Fried Rice with Tuna): 9.00  
133. Kani Don (Fried Rice with Tuna): 9.00  
134. Kani Don (Fried Rice with Tuna): 9.00  
135. Kani Don (Fried Rice with Tuna): 9.00  
136. Kani Don (Fried Rice with Tuna): 9.00  
137. Kani Don (Fried Rice with Tuna): 9.00  
138. Kani Don (Fried Rice with Tuna): 9.00  
139. Kani Don (Fried Rice with Tuna): 9.00  
140. Kani Don (Fried Rice with Tuna): 9.00  
141. Kani Don (Fried Rice with Tuna): 9.00  
142. Kani Don (Fried Rice with Tuna): 9.00  
143. Kani Don (Fried Rice with Tuna): 9.00  
144. Kani Don (Fried Rice with Tuna): 9.00  
145. Kani Don (Fried Rice with Tuna): 9.00  
146. Kani Don (Fried Rice with Tuna): 9.00  
147. Kani Don (Fried Rice with Tuna): 9.00  
148. Kani Don (Fried Rice with Tuna): 9.00  
149. Kani Don (Fried Rice with Tuna): 9.00  
150. Kani Don (Fried Rice with Tuna): 9.00  
151. Kani Don (Fried Rice with Tuna): 9.00  
152. Kani Don (Fried Rice with Tuna): 9.00  
153. Kani Don (Fried Rice with Tuna): 9.00  
154. Kani Don (Fried Rice with Tuna): 9.00  
155. Kani Don (Fried Rice with Tuna): 9.00  
156. Kani Don (Fried Rice with Tuna): 9.00  
157. Kani Don (Fried Rice with Tuna): 9.00  
158. Kani Don (Fried Rice with Tuna): 9.00  
159. Kani Don (Fried Rice with Tuna): 9.00  
160. Kani Don (Fried Rice with Tuna): 9.00  
161. Kani Don (Fried Rice with Tuna): 9.00  
162. Kani Don (Fried Rice with Tuna): 9.00  
163. Kani Don (Fried Rice with Tuna): 9.00  
164. Kani Don (Fried Rice with Tuna): 9.00  
165. Kani Don (Fried Rice with Tuna): 9.00  
166. Kani Don (Fried Rice with Tuna): 9.00  
167. Kani Don (Fried Rice with Tuna): 9.00  
168. Kani Don (Fried Rice with Tuna): 9.00  
169. Kani Don (Fried Rice with Tuna): 9.00  
170. Kani Don (Fried Rice with Tuna): 9.00  
171. Kani Don (Fried Rice with Tuna): 9.00  
172. Kani Don (Fried Rice with Tuna): 9.00  
173. Kani Don (Fried Rice with Tuna): 9.00  
174. Kani Don (Fried Rice with Tuna): 9.00  
175. Kani Don (Fried Rice with Tuna): 9.00  
176. Kani Don (Fried Rice with Tuna): 9.00  
177. Kani Don (Fried Rice with Tuna): 9.00  
178. Kani Don (Fried Rice with Tuna): 9.00  
179. Kani Don (Fried Rice with Tuna): 9.00  
180. Kani Don (Fried Rice with Tuna): 9.00  
181. Kani Don (Fried Rice with Tuna): 9.00  
182. Kani Don (Fried Rice with Tuna): 9.00  
183. Kani Don (Fried Rice with Tuna): 9.00  
184. Kani Don (Fried Rice with Tuna): 9.00  
185. Kani Don (Fried Rice with Tuna): 9.00  
186. Kani Don (Fried Rice with Tuna): 9.00  
187. Kani Don (Fried Rice with Tuna): 9.00  
188. Kani Don (Fried Rice with Tuna): 9.00  
189. Kani Don (Fried Rice with Tuna): 9.00  
190. Kani Don (Fried Rice with Tuna): 9.00  
191. Kani Don (Fried Rice with Tuna): 9.00  
192. Kani Don (Fried Rice with Tuna): 9.00  
193. Kani Don (Fried Rice with Tuna): 9.00  
194. Kani Don (Fried Rice with Tuna): 9.00  
195. Kani Don (Fried Rice with Tuna): 9.00  
196. Kani Don (Fried Rice with Tuna): 9.00  
197. Kani Don (Fried Rice with Tuna): 9.00  
198. Kani Don (Fried Rice with Tuna): 9.00  
199. Kani Don (Fried Rice with Tuna): 9.00  
200. Kani Don (Fried Rice with Tuna): 9.00  
201. Kani Don (Fried Rice with Tuna): 9.00  
202. Kani Don (Fried Rice with Tuna): 9.00  
203. Kani Don (Fried Rice with Tuna): 9.00  
204. Kani Don (Fried Rice with Tuna): 9.00  
205. Kani Don (Fried Rice with Tuna): 9.00  
206. Kani Don (Fried Rice with Tuna): 9.00  
207. Kani Don (Fried Rice with Tuna): 9.00  
208. Kani Don (Fried Rice with Tuna): 9.00  
209. Kani Don (Fried Rice with Tuna): 9.00  
210. Kani Don (Fried Rice with Tuna): 9.00  
211. Kani Don (Fried Rice with Tuna): 9.00  
212. Kani Don (Fried Rice with Tuna): 9.00  
213. Kani Don (Fried Rice with Tuna): 9.00  
214. Kani Don (Fried Rice with Tuna): 9.00  
215. Kani Don (Fried Rice with Tuna): 9.00  
216. Kani Don (Fried Rice with Tuna): 9.00  
217. Kani Don (Fried Rice with Tuna): 9.00  
218. Kani Don (Fried Rice with Tuna): 9.00  
219. Kani Don (Fried Rice with Tuna): 9.00  
220. Kani Don (Fried Rice with Tuna): 9.00  
221. Kani Don (Fried Rice with Tuna): 9.00  
222. Kani Don (Fried Rice with Tuna): 9.00  
223. Kani Don (Fried Rice with Tuna): 9.00  
224. Kani Don (Fried Rice with Tuna): 9.00  
225. Kani Don (Fried Rice with Tuna): 9.00  
226. Kani Don (Fried Rice with Tuna): 9.00  
227. Kani Don (Fried Rice with Tuna): 9.00  
228. Kani Don (Fried Rice with Tuna): 9.00  
229. Kani Don (Fried Rice with Tuna): 9.00  
230. Kani Don (Fried Rice with Tuna): 9.00  
231. Kani Don (Fried Rice with Tuna): 9.00  
232. Kani Don (Fried Rice with Tuna): 9.00  
233. Kani Don (Fried Rice with Tuna): 9.00  
234. Kani Don (Fried Rice with Tuna): 9.00  
235. Kani Don (Fried Rice with Tuna): 9.00  
236. Kani Don (Fried Rice with Tuna): 9.00  
237. Kani Don (Fried Rice with Tuna): 9.00  
238. Kani Don (Fried Rice with Tuna): 9.00  
239. Kani Don (Fried Rice with Tuna): 9.00  
240. Kani Don (Fried Rice with Tuna): 9.00  
241. Kani Don (Fried Rice with Tuna): 9.00  
242. Kani Don (Fried Rice with Tuna): 9.00  
243. Kani Don (Fried Rice with Tuna): 9.00  
244. Kani Don (Fried Rice with Tuna): 9.00  
245. Kani Don (Fried Rice with Tuna): 9.00  
246. Kani Don (Fried Rice with Tuna): 9.00  
247. Kani Don (Fried Rice with Tuna): 9.00  
248. Kani Don (Fried Rice with Tuna): 9.00  
249. Kani Don (Fried Rice with Tuna): 9.00  
250. Kani Don (Fried Rice with Tuna): 9.00  
251. Kani Don (Fried Rice with Tuna): 9.00  
252. Kani Don (Fried Rice with Tuna): 9.00  
253. Kani Don (Fried Rice with Tuna): 9.00  
254. Kani Don (Fried Rice with Tuna): 9.00  
255. Kani Don (Fried Rice with Tuna): 9.00  
256. Kani Don (Fried Rice with Tuna): 9.00  
257. Kani Don (Fried Rice with Tuna): 9.00  
258. Kani Don (Fried Rice with Tuna): 9.00  
259. Kani Don (Fried Rice with Tuna): 9.00  
260. Kani Don (Fried Rice with Tuna): 9.00  
261. Kani Don (Fried Rice with Tuna): 9.00  
262. Kani Don (Fried Rice with Tuna): 9.00  
263. Kani Don (Fried Rice with Tuna): 9.00  
264. Kani Don (Fried Rice with Tuna): 9.00  
265. Kani Don (Fried Rice with Tuna): 9.00  
266. Kani Don (Fried Rice with Tuna): 9.00  
267. Kani Don (Fried Rice with Tuna): 9.00  
268. Kani Don (Fried Rice with Tuna): 9.00  
269. Kani Don (Fried Rice with Tuna): 9.00  
270. Kani Don (Fried Rice with Tuna): 9.00  
271. Kani Don (Fried Rice with Tuna): 9.00  
272. Kani Don (Fried Rice with Tuna): 9.00  
273. Kani Don (Fried Rice with Tuna): 9.00  
274. Kani Don (Fried Rice with Tuna): 9.00  
275. Kani Don (Fried Rice with Tuna): 9.00  
276. Kani Don (Fried Rice with Tuna): 9.00  
277. Kani Don (Fried Rice with Tuna): 9.00  
278. Kani Don (Fried Rice with Tuna): 9.00  
279. Kani Don (Fried Rice with Tuna): 9.00  
280. Kani Don (Fried Rice with Tuna): 9.00  
281. Kani Don (Fried Rice with Tuna): 9.00  
282. Kani Don (Fried Rice with Tuna): 9.00  
283. Kani Don (Fried Rice with Tuna): 9.00  
284. Kani Don (Fried Rice with Tuna): 9.00  
285. Kani Don (Fried Rice with Tuna): 9.00  
286. Kani Don (Fried Rice with Tuna): 9.00  
287. Kani Don (Fried Rice with Tuna): 9.00  
288. Kani Don (Fried Rice with Tuna): 9.00  
289. Kani Don (Fried Rice with Tuna): 9.00  
290. Kani Don (Fried Rice with Tuna): 9.00  
291. Kani Don (Fried Rice with Tuna): 9.00  
292. Kani Don (Fried Rice with Tuna): 9.00  
293. Kani Don (Fried Rice with Tuna): 9.00  
294. Kani Don (Fried Rice with Tuna): 9.00  
295. Kani Don (Fried Rice with Tuna): 9.00  
296. Kani Don (Fried Rice with Tuna): 9.00  
297. Kani Don (Fried Rice with Tuna): 9.00  
298. Kani Don (Fried Rice with Tuna): 9.00  
299. Kani Don (Fried Rice with Tuna): 9.00  
300. Kani Don (Fried Rice with Tuna): 9.00  
301. Kani Don (Fried Rice with Tuna): 9.00  
302. Kani Don (Fried Rice with Tuna): 9.00  
303. Kani Don (Fried Rice with Tuna): 9.00  
304. Kani Don (Fried Rice with Tuna): 9.00  
305. Kani Don (Fried Rice with Tuna): 9.00  
306. Kani Don (Fried Rice with Tuna): 9.00  
307. Kani Don (Fried Rice with Tuna): 9.00  
308. Kani Don (Fried Rice with Tuna): 9.00  
309. Kani Don (Fried Rice with Tuna): 9.00  
310. Kani Don (Fried Rice with Tuna): 9.00  
311. Kani Don (Fried Rice with Tuna): 9.00  
312. Kani Don (Fried Rice with Tuna): 9.00  
313. Kani Don (Fried Rice with Tuna): 9.00  
314. Kani Don (Fried Rice with Tuna): 9.00  
315. Kani Don (Fried Rice with Tuna): 9.00  
316. Kani Don (Fried Rice with Tuna): 9.00  
317. Kani Don (Fried Rice with Tuna): 9.00  
318. Kani Don (Fried Rice with Tuna): 9.00  
319. Kani Don (Fried Rice with Tuna): 9.00  
320. Kani Don (Fried Rice with Tuna): 9.00  
321. Kani Don (Fried Rice with Tuna): 9.00  
322. Kani Don (Fried Rice with Tuna): 9.00  
323. Kani Don (Fried Rice with Tuna): 9.00  
324. Kani Don (Fried Rice with Tuna): 9.00  
325. Kani Don (Fried Rice with Tuna): 9.00  
326. Kani Don (Fried Rice with Tuna): 9.00  
327. Kani Don (Fried Rice with Tuna): 9.00  
328. Kani Don (Fried Rice with Tuna): 9.00  
329. Kani Don (Fried Rice with Tuna): 9.00  
330. Kani Don (Fried Rice with Tuna): 9.00  
331. Kani Don (Fried Rice with Tuna): 9.00  
332. Kani Don (Fried Rice with Tuna): 9.00  
333. Kani Don (Fried Rice with Tuna): 9.00  
334. Kani Don (Fried Rice with Tuna): 9.00  
335. Kani Don (Fried Rice with Tuna): 9.00  
336. Kani Don (Fried Rice with Tuna): 9.00  
337. Kani Don (Fried Rice with Tuna): 9.00  
338. Kani Don (Fried Rice with Tuna): 9.00  
339. Kani Don (Fried Rice with Tuna): 9.00  
340. Kani Don (Fried Rice with Tuna): 9.00  
341. Kani Don (Fried Rice with Tuna): 9.00  
342. Kani Don (Fried Rice with Tuna): 9.00  
343. Kani Don (Fried Rice with Tuna): 9.00  
344. Kani Don (Fried Rice with Tuna): 9.00  
345. Kani Don (Fried Rice with Tuna): 9.00  
346. Kani Don (Fried Rice with Tuna): 9.00  
347. Kani Don (Fried Rice with Tuna): 9.00  
348. Kani Don (Fried Rice with Tuna): 9.00  
349. Kani Don (Fried Rice with Tuna): 9.00  
350. Kani Don (Fried Rice with Tuna): 9.00  
351. Kani Don (Fried Rice with Tuna): 9.00  
352. Kani Don (Fried Rice with Tuna): 9.00  
353. Kani Don (Fried Rice with Tuna): 9.00  
354. Kani Don (Fried Rice with Tuna): 9.00  
355. Kani Don (Fried Rice with Tuna): 9.00  
356. Kani Don (Fried Rice with Tuna): 9.00  
357. Kani Don (Fried Rice with Tuna): 9.00  
358. Kani Don (Fried Rice with Tuna): 9.00  
359. Kani Don (Fried Rice with Tuna): 9.00  
360. Kani Don (Fried Rice with Tuna): 9.00  
361. Kani Don (Fried Rice with Tuna): 9.00  
362. Kani Don (Fried Rice with Tuna): 9.00  
363. Kani Don (Fried Rice with Tuna): 9.00  
364. Kani Don (Fried Rice with Tuna): 9.00  
365. Kani Don (Fried Rice with Tuna): 9.00  
366. Kani Don (Fried Rice with Tuna): 9.00  
367. Kani Don (Fried Rice with Tuna): 9.00  
368. Kani Don (Fried Rice with Tuna): 9.00  
369. Kani Don (Fried Rice with Tuna): 9.00  
370. Kani Don (Fried Rice with Tuna): 9.00  
371. Kani Don (Fried Rice with Tuna): 9.00  
372. Kani Don (Fried Rice with Tuna): 9.00  
373. Kani Don (Fried Rice with Tuna): 9.00  
374. Kani Don (Fried Rice with Tuna): 9.00  
375. Kani Don (Fried Rice with Tuna): 9.00  
376. Kani Don (Fried Rice with Tuna): 9.00  
377. Kani Don (Fried Rice with Tuna): 9.00  
378. Kani Don (Fried Rice with Tuna): 9.00  
379. Kani Don (Fried Rice with Tuna): 9.00  
380. Kani Don (Fried Rice with Tuna): 9.00  
381. Kani Don (Fried Rice with Tuna): 9.00  
382. Kani Don (Fried Rice with Tuna): 9.00  
383. Kani Don (Fried Rice with Tuna): 9.00  
384. Kani Don (Fried Rice with Tuna): 9.00  
385. Kani Don (Fried Rice with Tuna): 9.00  
386. Kani Don (Fried Rice with Tuna): 9.00  
387. Kani Don (Fried Rice with Tuna): 9.00  
388. Kani Don (Fried Rice with Tuna): 9.00  
389. Kani Don (Fried Rice with Tuna): 9.00  
390. Kani Don (Fried Rice with Tuna): 9.00  
391. Kani Don (Fried Rice with Tuna): 9.00  
392. Kani Don (Fried Rice with Tuna): 9.00  
393. Kani Don (Fried Rice with Tuna

---

02 **MODEL**

Mixed Preference Optimization(MPO)



**InternVL2-8B-MPO**

---

# Model with Mixed Preference Optimization

Only DPO → positive 응답이 상대적으로 더 좋아 보이지만,  
사실상 절대적으로 생성 확률이 떨어질 수 있음!

**Training Objective**  $\mathcal{L} = w_p \mathcal{L}_p + w_q \mathcal{L}_q + w_g \mathcal{L}_g$

## PREFERENCE LOSS

선호 쌍 학습용

$$\mathcal{L}_p = -\log \sigma \left( \beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} - \beta \log \frac{\pi_\theta(y_r | x)}{\pi_0(y_r | x)} \right)$$

DPO (Direct Preference Optimization) loss로,  
chosen 응답과 rejected 응답 중 어느 쪽  
이 더 나은지 **상대적인 선호 정도**를 학습

## QUALITY LOSS

응답 품질 평가용

$$\mathcal{L}_q = \mathcal{L}_q^+ + \mathcal{L}_q^-$$

$$\mathcal{L}_q^+ = -\log \sigma \left( \beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} - \delta \right),$$

$$\mathcal{L}_q^- = -\log \sigma \left( - \left( \beta \log \frac{\pi_\theta(y_r | x)}{\pi_0(y_r | x)} - \delta \right) \right)$$

응답의 **절대적 품질(quality)**을 평가해서  
학습하는 loss.

Binary classifier (예: BCO)가  
chosen/rejected 응답을  
분류하도록 학습

## GENERATION LOSS

정답 생성 loss

$$\mathcal{L}_g = -\frac{\log \pi_\theta(y_c | x)}{|y_c|}.$$

SFT (Supervised Fine-Tuning)  
loss로, chosen 응답이 잘 생성되도록  
**cross-entropy 기반**으로 학습

# Model with Mixed Preference Optimization

시그모이드 함수로 soft binary 분류 확률로 바꾸고,  $w_p \mathcal{L}_p + w_q \mathcal{L}_q + w_g \mathcal{L}_g$

negative log-likelihood로 loss 형태로 변환

chosen > rejected 라고 판단할수록 loss ↓

chosen - rejected

PREFERENCE LOSS

PREFERENCE LOSS

GENERATION LOSS

현재 학습 중인 모델  $\theta$  가 instruction  $x$  에 대해 응답  $y$  를 생성할 확률

$$\mathcal{L}_p = -\log \sigma \left( \beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} \right)$$

$$\mathcal{L}_p = -\log \sigma \left( \beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} - \beta \log \frac{\pi_\theta(y_r | x)}{\pi_0(y_r | x)} \right) = -\frac{\log \pi_\theta(y_c | x)}{|y_c|}$$

DPO (Direct Preference Optimization) loss로,

chosen 응답과 rejected 응답 중 어느 쪽이 더 나은지 상대적인 선호 정도를 학습

$$\mathcal{L}_q = -\log \sigma \left( - \left( \beta \log \frac{\pi_\theta(y_r | x)}{\pi_0(y_r | x)} - \beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} \right) \right)$$

초기 모델 (예: 사전학습된 모델)에서의 동일한 생성 확률

loss로, chosen 응답이 잘 생성되도록 cross-entropy 기반으로 학습

하나의 절대적 품질(quality)을 평가해서 학습하는 loss.

Binary classifier (예: BCO)가



현재 모델이 chosen 응답을 rejected 응답보다 더 선호하도록 직접적으로 학습하는 DPO(Direct Preference Optimization) 기반 loss

# Model with Mixed Preference Optimization

Training Objective  $\mathcal{L} = w_p \mathcal{L}_p + w_q \mathcal{L}_q + w_g \mathcal{L}_g$

## QUALITY LOSS

### PREFERENCE

$$\mathcal{L}_q = \mathcal{L}_q^+ + \mathcal{L}_q^-$$

chosen 응답이 positive로 분류되도록 학습

rejected 응답이 negative로 분류되도록 학습

### GENERATION LOSS

시그모이드 함수

$$\mathcal{L}_q^+ = -\log \sigma \left( \beta \log \frac{\pi_\theta(y_c | x)}{\pi_0(y_c | x)} - \delta \right)$$

$$\mathcal{L}_q^- = -\log \sigma \left( - \left( \beta \log \frac{\pi_\theta(y_r | x)}{\pi_0(y_r | x)} - \delta \right) \right)$$

$$\mathcal{L}_g = -\frac{\log \pi_\theta(y_c | x)}{|y_c|}$$

<reward shift의 역할>

평균 reward를 계산

전체 score 분포를 이동(shift)시키는 값 (Revised Fine-Tuning)

- gradient 안정화 : 너무 큰/작은 logit 값으로 gradient 소실 방지
- 스케일 정규화: 전체 score 평균 baseline을 맞춰 relative 차이에 집중
- preference loss와 차이: ranking만 중요한게 아닌 절대 score로 binary 분류라  $\delta$  필요

Binary classifier (예. BCO)가

chosen/rejected 응답을

분류하도록 학습

# Model with Mixed Preference Optimization

Training Objective  $\mathcal{L} = w_p \mathcal{L}_p + w_q \mathcal{L}_q + w_g \mathcal{L}_g$

## GENERATION LOSS

### PREFERENCE LOSS

$$\mathcal{L}_p = -\log \sigma \left( \beta \log \frac{\pi_\theta(y_c | x)}{\pi_\theta(y_r | x)} - \beta \log \frac{\pi_0(y_c | x)}{\pi_0(y_r | x)} \right)$$

DPO (Direct Preference Optimization) loss로, chosen 응답과 rejected 응답 중 어느 쪽이 더 나은지 상대적인 선호 정도를 학습

### QUALITY LOSS

현재 모델  $\theta$  가 instruction  $x$ 에 대해 chosen response  $y_c$  를 생성할 확률

$$\mathcal{L}_g = -\frac{\log \pi_\theta(y_c | x)}{|y_c|}$$

$\mathcal{L}_q = -\log \sigma \left( -\left( \beta \log \frac{\pi_\theta(y_c | x)}{\pi_\theta(y_r | x)} - \delta \right) \right)$   
 $y_c$ 의 길이 (토큰 수, sequence length)

응답 하나하나의 절대적 품질(quality)을 평가해서 학습하는 loss.  
chosen response의 log-likelihood (로그 생성 확률)를 토큰당 평균으로 계산한 값

### GENERATION LOSS

$$\mathcal{L}_g = -\frac{\log \pi_\theta(y_c | x)}{|y_c|}$$

SFT (Supervised Fine-Tuning) loss로, chosen 응답이 잘 생성되도록 cross-entropy 기반으로 학습



generation loss는 chosen 응답을 정확히 생성하도록 negative log-likelihood로 학습시키는 SFT용 손실로, preference loss와 quality loss가 주지 못하는 absolute copy 능력을 강화하는 역할

Binary classifier (예: BCO)가 chosen/rejected 응답을 분류하도록 학습

# Chain-of-thought with Multimodal Input

	방식 설명	적용 도메인
Background Knowledge-based CoT	문제/이미지와 관련된 배경 지식 소개 → reasoning 단계 → 최종 답변	Science
Visual Content-based CoT	이미지 내용(예: 그래프, 문서, 숫자 등) 분석 → reasoning 단계 → 최종 답변	Chart, OCR, Document
Grounded CoT	텍스트 응답 생성 중, 언급된 모든 객체들을 이미지 속 특정 영역(region)과 연결 (예: "이 사과(이미지 속 좌측 상단)를 참조함")	General VQA

# Main Results

그룹	벤치마크	내용	평가 방식
Multimodal reasoning	M3CoT	Multimodal CoT reasoning 평가	Accuracy
	MathVista	Multimodal 수학 문제 해결 평가	Accuracy
	MathVision	실수학경시대회 문제 기반 고난도 수학 추론 평가	Accuracy
Complex VQA	MM-Vet	Visual conversation across diverse tasks	GPT-4 correctness/helpfulness score
	LLaVA-Bench	Multimodal 대화, 묘사, reasoning 평가	GPT-4 correctness/helpfulness score


Model Name	Reasoning			General VQA		Hallucination Evaluation		
	M3CoT	MathVista	MathVision	MMVet	LLaVA-Bench	POPE	CRPE	MMHalBench
<i>Closed-Source Models</i>								
Gemini-1.5-Pro [78]	-	63.9	19.2	-	-	-	-	-
GPT-4o [71]	64.3	63.8	30.4	69.1	97.6	86.9	76.6	4.0
GPT-4o-Mini [71]	61.9	52.4	27.3	66.9	95.4	85.1	73.1	3.6
<i>Open-Source Models</i>								
LLaVA-1.5-13B [52]	39.5	27.6	11.1	36.3	70.7	85.9	55.6	2.4
Qwen2-VL-7B [96]	57.8	58.2	21.1	60.6	67.7	88.1	74.4	3.4
MiniCPM-V-2-6-8B [105]	56.0	60.6	23.4	57.4	83.4	87.3	75.2	3.6
LLaVA-OneVision-7B [44]	52.3	63.2	18.4	51.4	79.9	88.4	73.7	3.1
<i>InternVL Models</i>								
InternVL2-26B [20]	58.2	59.4	23.4	62.1	92.3	88.0	75.6	3.7
InternVL2-40B [20]	63.6	63.7	21.4	65.5	100.5	88.4	77.3	3.9
InternVL2-76B [20]	65.4	67.2	23.7	65.7	99.3	89.0	77.8	3.8
InternVL2-Pro [20]	65.6	66.3	18.8	69.4	99.5	88.2	77.6	3.7
InternVL2-8B [20]	59.3	58.3	20.4	54.2	73.2	86.9	75.0	3.3
InternVL2-8B-MPO (ours)	79.2	67.0	25.7	56.2	76.7	88.1	75.4	3.5

Table 2. **Results on 8 multimodal benchmarks.** We report the overall score of MM-Vet and LLaVA-Bench evaluated by GPT-4-Turbo. Our InternVL2-8B-MPO demonstrates superior performance compared to InternVL2-8B across multimodal reasoning, VQA, and hallucination evaluation benchmarks. Notably, our model even achieves reasoning performance comparable to the 10× larger InternVL2-76B.

# Hallucination evaluation

## POPE

### Pointing Out the Presence of Entities



**Instruction-based evaluation**

Provide a detailed description of the given image.

The image features a table with a variety of food items displayed in bowls. There are two bowls of food, one containing a mix of vegetables, such as **broccoli** and **carrots**, and the other containing meat. The bowl with vegetables is placed closer to the front, while the **meat bowl** is situated behind it. In addition to the main dishes, there is an **apple** placed on the table, adding a touch of fruit to the meal. A **bottle** can also be seen on the table, possibly containing a **beverage** or **condiment**. The table is neatly arranged, showcasing the different food items in an appetizing manner.

**POPE**

*Random settings*

Is there a **bottle** in the image?

Yes, there is a bottle in the image.

*Popular settings*

Is there a **knife** in the image?

Yes, there is a knife in the image.

*Adversarial settings*

Is there a **pear** in the image?

Yes, there is a pear in the image.

Figure 1: Cases of object hallucination in LVLMs. **Bold** objects are ground-truth objects in the annotations and **red** objects are hallucinated objects by LVLMs. The left case is from the traditional instruction-based evaluation method, and the right cases are from three variants of POPE.

object의 존재 여부에 대한 hallucination 평가  
모델에게 이미지나 멀티모달 입력에 기반한  
Yes/No 질문을 던짐

> 모델이 실제로 존재하지 않는 객체를 있다고 말하면 →  
hallucination으로 간주  
metric: F1 score

## CRPE

### Checking Relations for Perceptual Entities

객체 간의 관계에 대한 hallucination 평가  
multiple-choice questions(객관식)

ex.

“이미지에서 강아지는 어디에 있나요?”  
A. 소파 위 B. 바닥에 C. 주인의 무릎 위

잘못된 관계(예: 개는 소파에 없는데 있다고 함)를  
말하면 → hallucination

metric: accuracy

## MMHalBench

### Multimodal Hallucination Benchmark

### Example 3:

- Image Contents: Fashion accessory, Mammal, Dog, Umbrella
- Question: Describe this image in details.
- Standard Human-Generated Answer: In this picture I can see a dog laying under a yellow umbrella on a beach, and there is a rope on the right side. In the background I can see few persons talking or lying . At the top of the image, I can see the blue and clear sky.
- LMM Response to Evaluate: The image features a brown dog laying on the sand at the beach, underneath a yellow umbrella. The dog is positioned near the water, and there are people in the background. The beach is sandy, and the sun is shining, creating a pleasant atmosphere. The dog is enjoying the shade provided by the umbrella, which is a common way to protect pets from the sun's harmful rays. The presence of people in the background suggests that the beach is a popular spot for leisure activities, such as sunbathing, swimming, and relaxing.
- Analysis: The LMM's response is false because it claims that the dog is positioned near the water, but there is no water in the image.
- Hallucination: Yes.

open-ended multimodal 질문에서:

hallucination rate (환각률)

informativeness (정보량, 유용성)

모델 답변과 인간 답변을 비교 →

GPT-4가 심사위원(judge)로 나서서:

어떤 답이 더 덜 hallucinated인지,  
어떤 답이 더 informative한지 평가.

# Main Results

Model Name	Reasoning			General VQA		Hallucination Evaluation		
	M3CoT	MathVista	MathVision	MMVet	LLaVA-Bench	POPE	CRPE	MMHalBench
<i>Closed-Source Models</i>								
Gemini-1.5-Pro [78]	-	63.9	19.2	-	-	-	-	-
GPT-4o [71]	64.3	63.8	30.4	69.1	97.6	86.9	76.6	4.0
GPT-4o-Mini [71]	61.9	52.4	27.3	66.9	95.4	85.1	73.1	3.6
<i>Open-Source Models</i>								
LLaVA-1.5-13B [52]	39.5	27.6	11.1	36.3	70.7	85.9	55.6	2.4
Qwen2-VL-7B [96]	57.8	58.2	21.1	60.6	67.7	88.1	74.4	3.4
MiniCPM-V-2-6-8B [105]	56.0	60.6	23.4	57.4	83.4	87.3	75.2	3.6
LLaVA-OneVision-7B [44]	52.3	63.2	18.4	51.4	79.9	88.4	73.7	3.1
<i>InternVL Models</i>								
InternVL2-26B [20]	58.2	59.4	23.4	62.1	92.3	88.0	75.6	3.7
InternVL2-40B [20]	63.6	63.7	21.4	65.5	100.5	88.4	77.3	3.9
InternVL2-76B [20]	65.4	67.2	23.7	65.7	99.3	89.0	77.8	3.8
InternVL2-Pro [20]	65.6	66.3	18.8	69.4	99.5	88.2	77.6	3.7
InternVL2-8B [20]	59.3	58.3	20.4	54.2	73.2	86.9	75.0	3.3
InternVL2-8B-MPO (ours)	79.2	67.0	25.7	56.2	76.7	88.1	75.4	3.5

Table 2. **Results on 8 multimodal benchmarks.** We report the overall score of MM-Vet and LLaVA-Bench evaluated by GPT-4-Turbo. Our InternVL2-8B-MPO demonstrates superior performance compared to InternVL2-8B across multimodal reasoning, VQA, and hallucination evaluation benchmarks. Notably, our model even achieves reasoning performance comparable to the 10× larger InternVL2-76B.

# Main Results

Model Name	Reasoning			General VQA		Hallucination Evaluation		
	M3CoT	MathVista	MathVision	MMVet	LLaVA-Bench	POPE	CRPE	MMHalBench
<i>Closed-Source Models</i>								
Gemini-1.5-Pro [78]	-	63.9	19.2	-	-	-	-	-
GPT-4o [71]	64.3	63.8	30.4	69.1	97.6	86.9	76.6	4.0
GPT-4o-Mini [71]	61.9	52.4	27.3	66.9	95.4	85.1	73.1	3.6
<i>Open-Source Models</i>								
LLaVA-1.5-13B [52]	39.5	27.6	11.1	36.3	70.7	85.9	55.6	2.4
Qwen2-VL-7B [96]	57.8	58.2	21.1	60.6	67.7	88.1	74.4	3.4
MiniCPM-V-2-6-8B [105]	56.0	60.6	23.4	57.4	83.4	87.3	75.2	3.6
LLaVA-OneVision-7B [44]	52.3	63.2	18.4	51.4	79.9	88.4	73.7	3.1
<i>InternVL Models</i>								
InternVL2-26B [20]	55.2	59.4	21.4	62.1	92.3	88.0	75.6	3.7
InternVL2-40B [20]	63.6	63.7	21.4	65.5	100.5	88.4	77.3	3.9
InternVL2-76B [20]	65.4	67.2	23.7	65.7	99.3	89.0	77.8	3.8
InternVL2-Pro [20]	65.6	66.3	18.8	69.4	99.5	88.2	77.6	3.7
InternVL2-8B [20]	59.3	58.3	20.4	54.2	73.2	86.9	75.0	3.3
InternVL2-8B-MPO (ours)	79.2	67.0	25.7	56.2	76.7	88.1	75.4	3.5

Table 2. **Results on 8 multimodal benchmarks.** We report the overall score of MM-Vet and LLaVA-Bench evaluated by GPT-4-Turbo. Our InternVL2-8B-MPO demonstrates superior performance compared to InternVL2-8B across multimodal reasoning, VQA, and hallucination evaluation benchmarks. Notably, our model even achieves reasoning performance comparable to the 10× larger InternVL2-76B.

# Main Results

reasoning 쪽이 상대적으로 높음

Model Name	Reasoning			General VQA		Hallucination Evaluation		
	M3CoT	MathVista	MathVision	MMVet	LLaVA-Bench	POPE	CRPE	MMHalBench
<i>Closed-Source Models</i>								
Gemini-1.5-Pro [78]	-	63.9	19.2	-	-	-	-	-
GPT-4o [71]	64.3	63.8	30.4	69.1	97.6	86.9	76.6	4.0
GPT-4o-Mini [71]	61.9	52.4	27.3	66.9	95.4	85.1	73.1	3.6
<i>Open-Source Models</i>								
LLaVA-1.5-13B [52]	39.5	27.6	11.1	36.3	70.7	85.9	55.6	2.4
Qwen2-VL-7B [96]	57.8	58.2	21.1	60.6	67.7	88.1	74.4	3.4
MiniCPM-V-2-6-8B [105]	56.0	60.6	23.4	57.4	83.4	87.3	75.2	3.6
LLaVA-OneVision-7B [44]	52.3	63.2	18.4	51.4	79.9	88.4	73.7	3.1
<i>InternVL Models</i>								
InternVL2-26B [20]	58.2	59.4	23.4	62.1	92.3	88.0	75.6	3.7
InternVL2-40B [20]	63.6	63.7	21.4	65.5	100.5	88.4	77.3	3.9
InternVL2-76B [20]	65.4	67.2	23.7	65.7	99.3	89.0	77.8	3.8
InternVL2-Pro [20]	65.6	66.3	18.8	69.4	99.5	88.2	77.6	3.7
InternVL2-8B [20]	59.3	58.3	20.4	54.2	73.2	86.9	75.0	3.3
InternVL2-8B-MPO (ours)	79.2	67.0	25.7	56.2	76.7	88.1	75.4	3.5

Table 2. **Results on 8 multimodal benchmarks.** We report the overall score of MM-Vet and LLaVA-Bench evaluated by GPT-4-Turbo. Our InternVL2-8B-MPO demonstrates superior performance compared to InternVL2-8B across multimodal reasoning, VQA, and hallucination evaluation benchmarks. Notably, our model even achieves reasoning performance comparable to the 10× larger InternVL2-76B.

# Main Results

POPE : object의 존재 여부에 대한 hallucination 평가

Model Name	Reasoning			General VQA		Hallucination Evaluation		
	M3CoT	MathVista	MathVision	MMVet	LLaVA-Bench	POPE	CRPE	MMHalBench
<i>Closed-Source Models</i>								
Gemini-1.5-Pro [78]	-	63.9	19.2	-	-	-	-	-
GPT-4o [71]	64.3	63.8	30.4	69.1	97.6	86.9	76.6	4.0
GPT-4o-Mini [71]	61.9	52.4	27.3	66.9	95.4	85.1	73.1	3.6
<i>Open-Source Models</i>								
LLaVA-1.5-13B [52]	39.5	27.6	11.1	36.3	70.7	85.9	55.6	2.4
Qwen2-VL-7B [96]	57.8	58.2	21.1	60.6	67.7	88.1	74.4	3.4
MiniCPM-V-2-6-8B [105]	56.0	60.6	23.4	57.4	83.4	87.3	75.2	3.6
LLaVA-OneVision-7B [44]	52.3	63.2	18.4	51.4	79.9	88.4	73.7	3.1
<i>InternVL Models</i>								
InternVL2-26B [20]	58.2	59.4	23.4	62.1	92.3	88.0	75.6	3.7
InternVL2-40B [20]	63.6	63.7	21.4	65.5	100.5	88.4	77.3	3.9
InternVL2-76B [20]	65.4	67.2	23.7	65.7	99.3	89.0	77.8	3.8
InternVL2-Pro [20]	65.6	66.3	18.8	69.4	99.5	88.2	77.6	3.7
InternVL2-8B [20]	59.3	58.3	20.4	54.2	73.2	86.9	75.0	3.3
InternVL2-8B-MPO (ours)	79.2	67.0	25.7	56.2	76.7	88.1	75.4	3.5

Table 2. **Results on 8 multimodal benchmarks.** We report the overall score of MM-Vet and LLaVA-Bench evaluated by GPT-4-Turbo. Our InternVL2-8B-MPO demonstrates superior performance compared to InternVL2-8B across multimodal reasoning, VQA, and hallucination evaluation benchmarks. Notably, our model even achieves reasoning performance comparable to the 10× larger InternVL2-76B.

divide-and-conquer approach란?

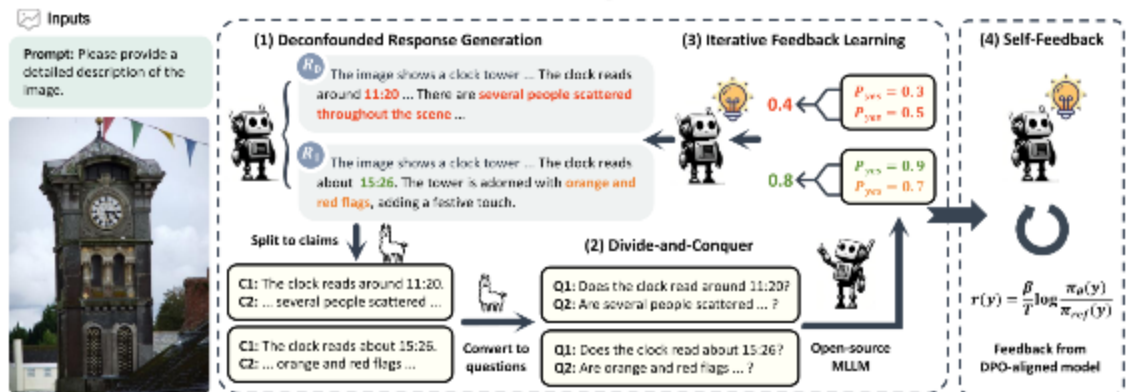
여러 candidate 응답을 생성 →

divide-and-conquer로 각 응답을 atomic claim 단위로 쪼갬 → claim별 yes/no 평가 →

전체 응답의 신뢰도 점수 계산 → 최종적으로 ranking → chosen/rejected pair 구성.

# Ablation Study

: 모델, 시스템, 알고리즘의 구성 요소별 중요성을 평가하기 위해, 일부 요소를 제거(ablate)하거나 교체하면서 성능 변화를 분석



## 1. Comparison between MPO & SFT **Model 비교**

Model Name	Setting	M3CoT	MathVista	MMVet	POPE
InternVL2-8B	Direct	59.3	58.3	54.2	86.9
	CoT	57.0	56.8	54.7	82.9
InternVL2-8B-SFT	Direct	63.9	62.7	54.7	86.5
	CoT	67.8	64.2	53.8	84.0
InternVL2-8B-MPO	Direct	77.2	64.5	55.1	87.0
	CoT	79.2	67.0	56.2	88.1

SFT only: MMPR dataset에서 chosen response 만 학습 (x, yc)

MPO(SFT + MMPR): MMPR positive /negative (x, yc, yn)

## 2. Comparison with RLAIIF-V **Dataset 비교**

Dropout Next-Token Prediction (Dropout NTP) 방식과

RLAIIF-V의 divide-and-conquer approach를 비교

전체 응답 중 잘못된 응답 비율    응답 안에서 언급된 객체들 중 잘못 언급된 비율

Method	Object HalBench		MM HalBench	
	Resp. (↓)	Ment. (↓)	Score	Hall. (↓)
InternVL2-8B	18.4	8.7	3.3	40.6
RLAIIF-V [107]	7.3	3.9	3.5	32.3
DropoutNTP (ours)	7.6	4.1	3.6	31.3

공정한 비교를 위해 RLAIIF-V에서 사용한 것과 동일한 prompt와 chosen response를 쓰되, rejected response는 Dropout NTP 방식으로 (즉, 이미지 없이 continuation 생성으로) 대체

둘 다 멀티모달 모델의 hallucination(환각) 감소와 trustworthiness(신뢰성) 향상을 위한 preference data 생성 방식으로 쓰임

# Ablation Study

: 모델, 시스템, 알고리즘의 구성 요소별 중요성을 평가하기 위해, 일부 요소를 제거(ablate)하거나 교체하면서 성능 변화를 분석

## 3. Effects of optimization algorithms

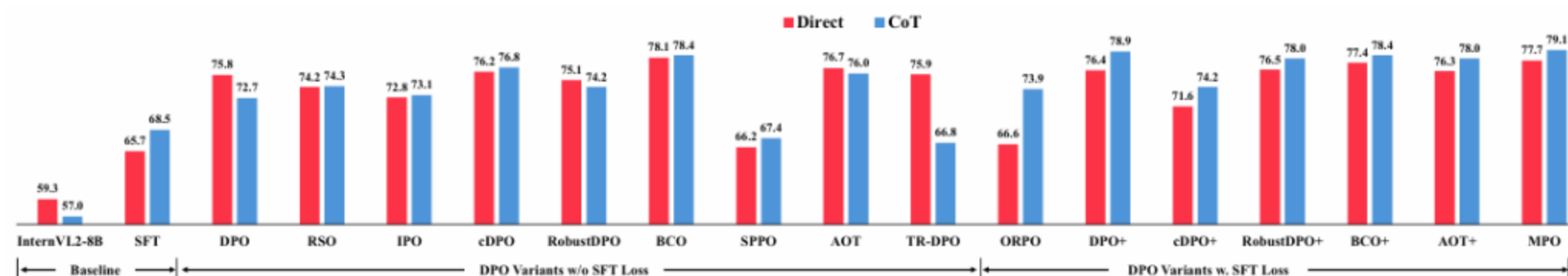


Figure 3. Results of models trained with different preference optimization algorithms on M3CoT. The algorithm X extended with the SFT loss is called X+ for brevity. For instance, DPO+ denotes the combination of DPO loss and SFT loss.

DPO (Direct Preference Optimization):

reward 함수를 따로 만들지 않고, offline preference pair로 직접 fine-tuning.

RSO (Relative Score Optimization): DPO의 sigmoid loss 대신 hinge loss 사용.

IPO (Implicit Preference Optimization): DPO의 overfitting 문제 완화: log-likelihood 평균화 + chosen/rejected gap 조절용  $\beta$  (베타) 파라미터 도입.

cDPO (confident DPO): preference 데이터에 label noise(오류) 가능성을 고려한 DPO 수정 버전.

RobustDPO: noisy label을 가정하고 DPO loss의 unbiased(편향 없는) 추정치 제공.

BCO (Behavior Cloning Objective): binary classifier 학습  $\rightarrow$  logit 값을 reward로 사용.

SPPO (Self-Play Preference Optimization): chosen reward는 +1/2,

rejected reward는 -1/2로 밀어넣어 Nash equilibrium(내쉬 균형) 근사, data sparsity 문제 완화.

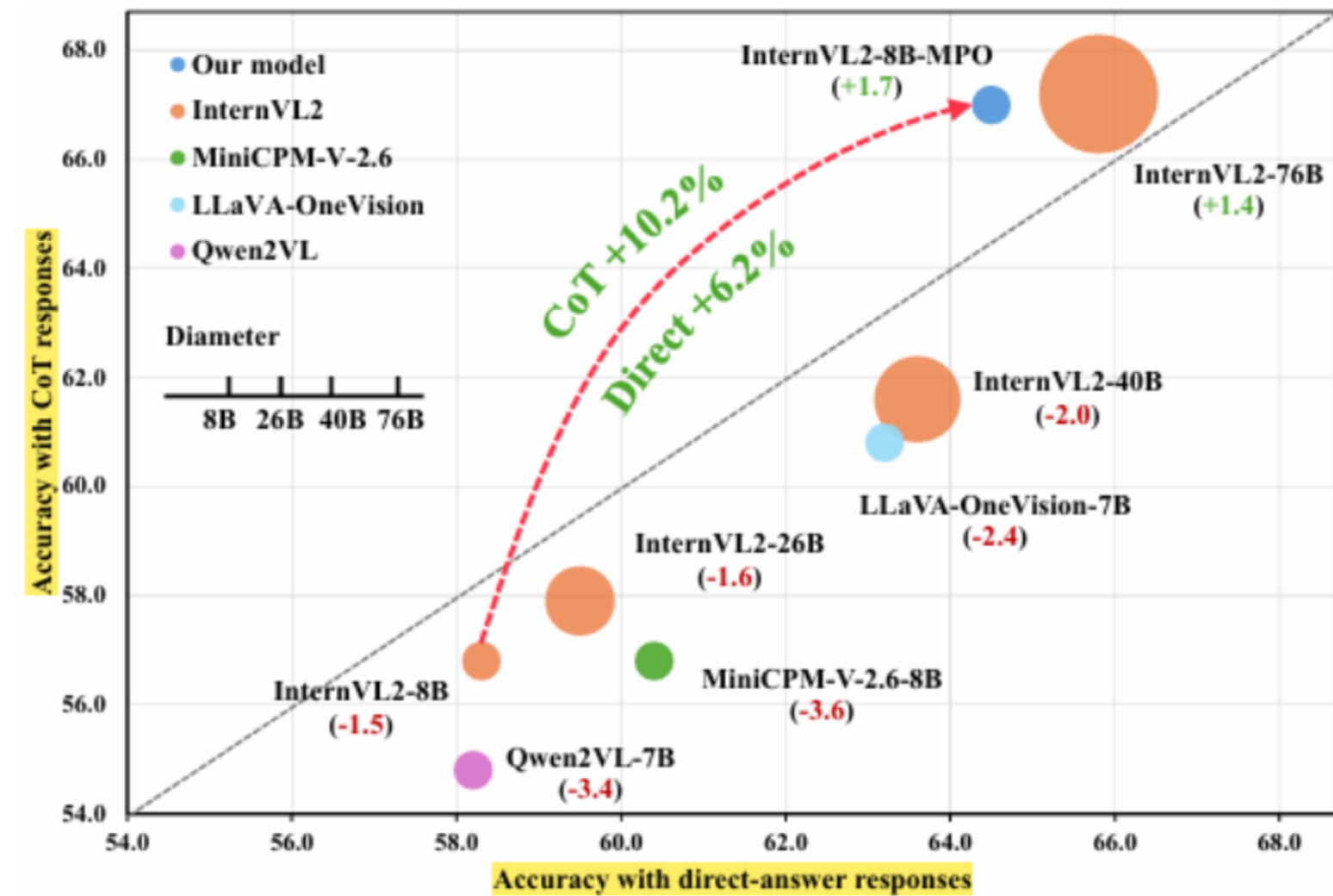
AOT (Alignment via Optimal Transport): optimal transport를 이용해 distributional preference alignment 적용.

TR-DPO (Teacher-Referenced DPO) : DPO 훈련 중 몇 스텝마다 reference model과 동기화해 overfitting 완화.

ORPO (Odds Ratio Preference Optimization): reference model 없이 log-odds ratio penalty를 NLL(loss)에 추가, 별도 preference alignment phase 없이 fine-tuning.

- learning rate =  $5e-6$
- 각 논문의 hyper-parameter 설정 그대로 사용
- 각 알고리즘에 SFT loss도 확장 적용  $\rightarrow$  SFT의 영향 분석
- baseline : reasoning preference data에서 chosen response만으로 학습한 SFT 모델 포함

# Conclusion



---

# Q&A

---



논문에서는 SFT (supervised fine-tuning)를 사용했는데,  
만약 SFT + Dropout NTP 대신  
self-supervised fine-tuning을 쓰면  
distribution shift 문제가 해결되지 않을까?

---