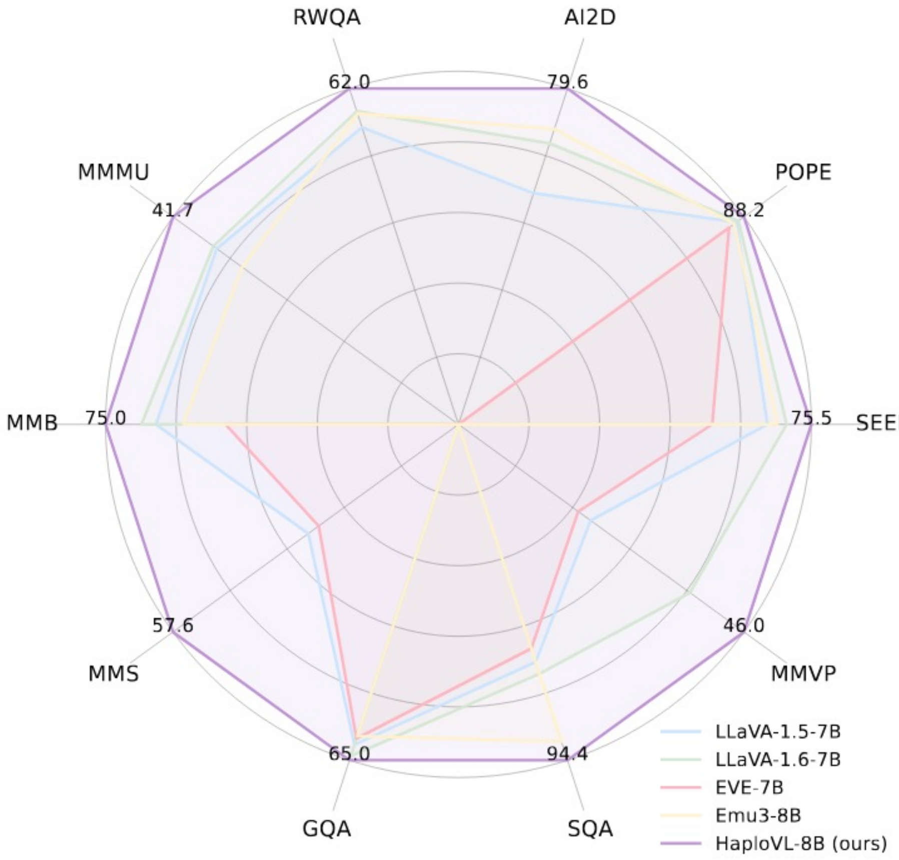
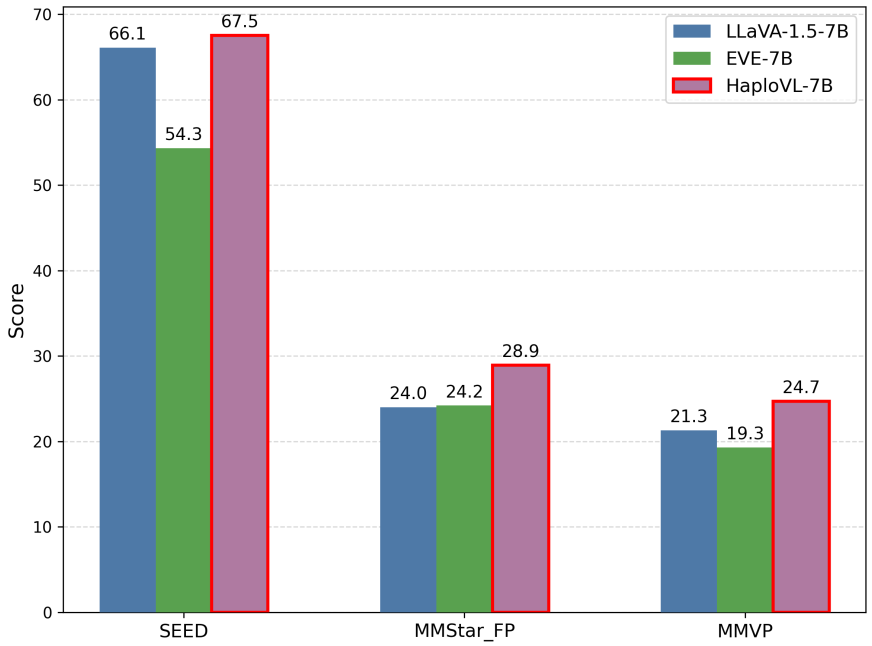
HaploVL: A Single-Transformer Baseline for Multi-Modal Understanding

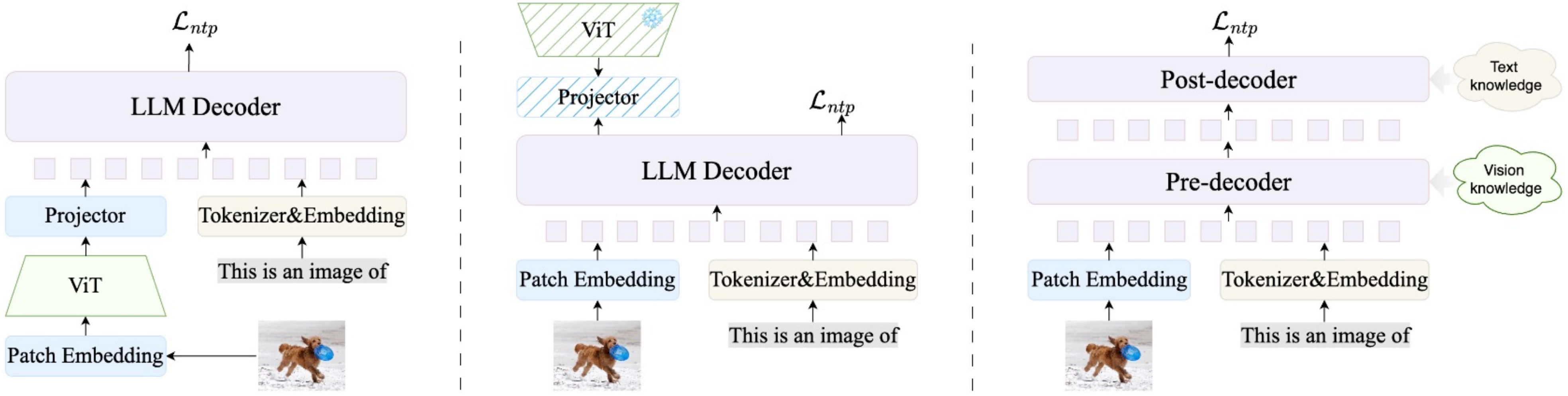
Introduction





Introduction

- Early-fusion LMM



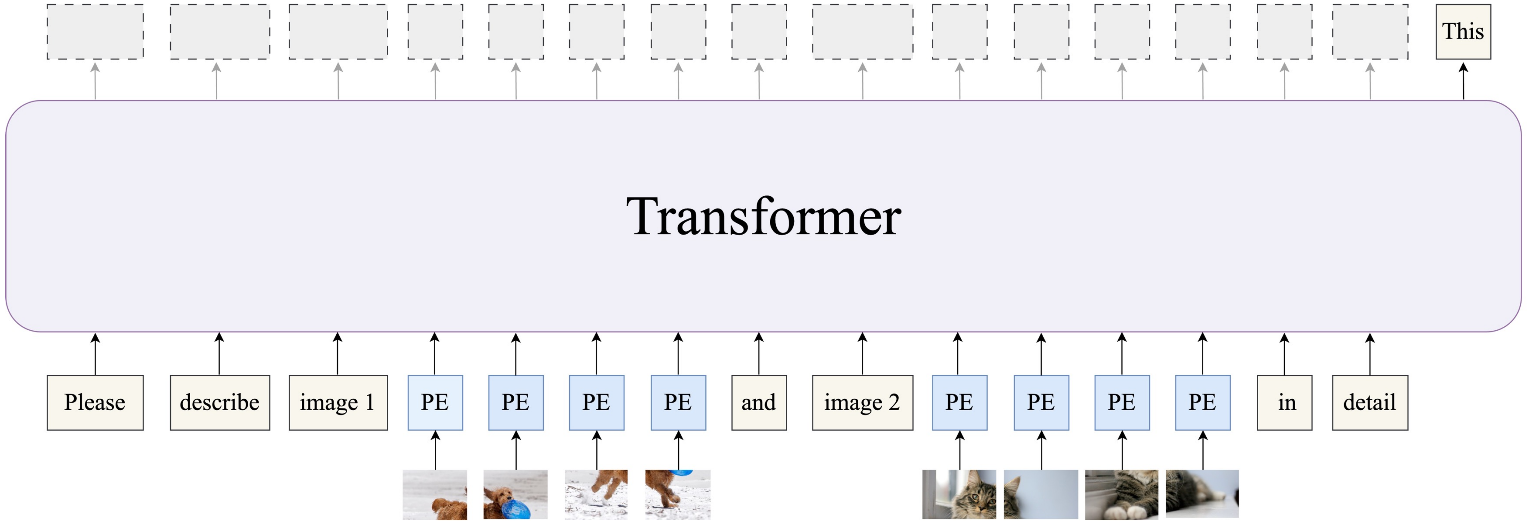
(a) The compositional LMM

(b) EVE

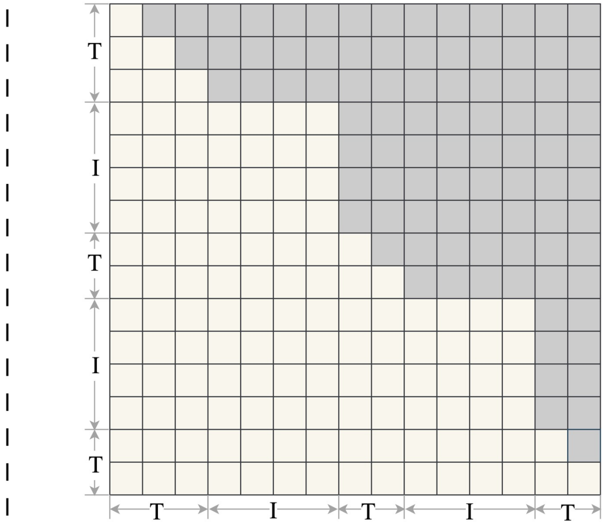
(c) HaploVL (ours)

Approach

- Model Architecture



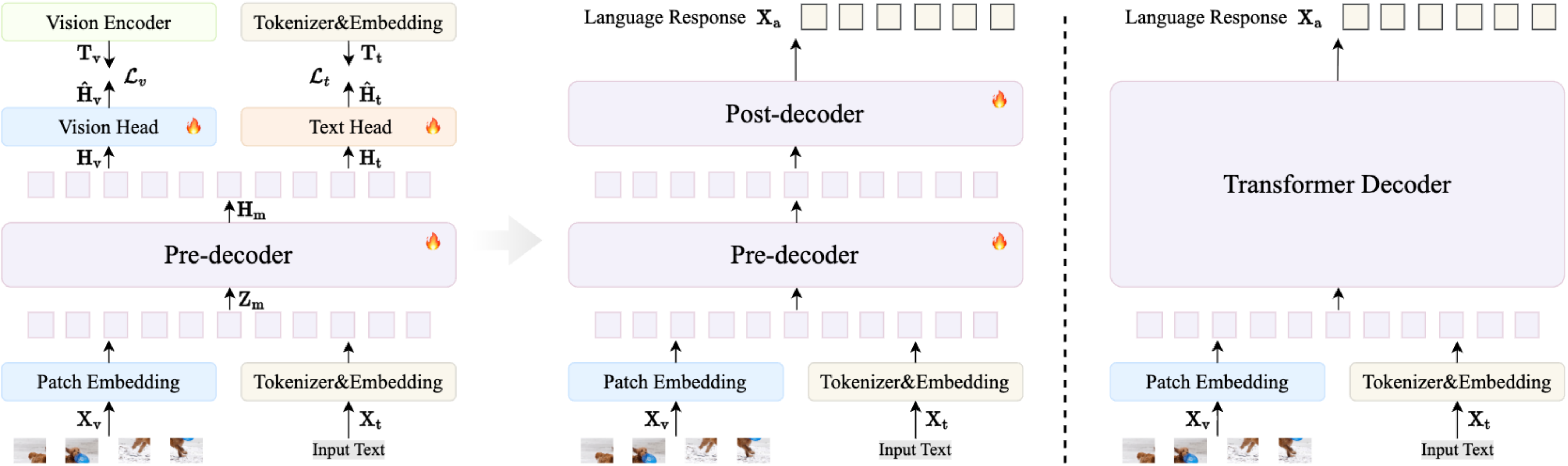
(a) Architecture



(b) Mask Strategy

Approach

- Training Receipt



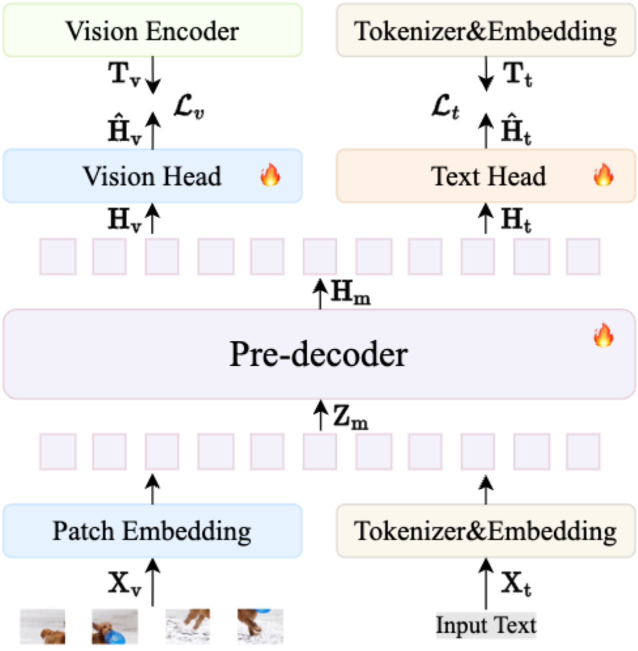
(a) Stage 1: Pre-training

(b) Stage 2: Fully finetuning

(c) Architecture

Approach

- Pretraining



(a) Stage 1: Pre-training

Vision loss:

$$\mathcal{L}_v = 1 - \frac{1}{hw}\sum_{i=1}^{hw}\cos(\hat{\mathrm{H}}_{\mathrm{v},i};\mathrm{T}_{\mathrm{v},i})$$

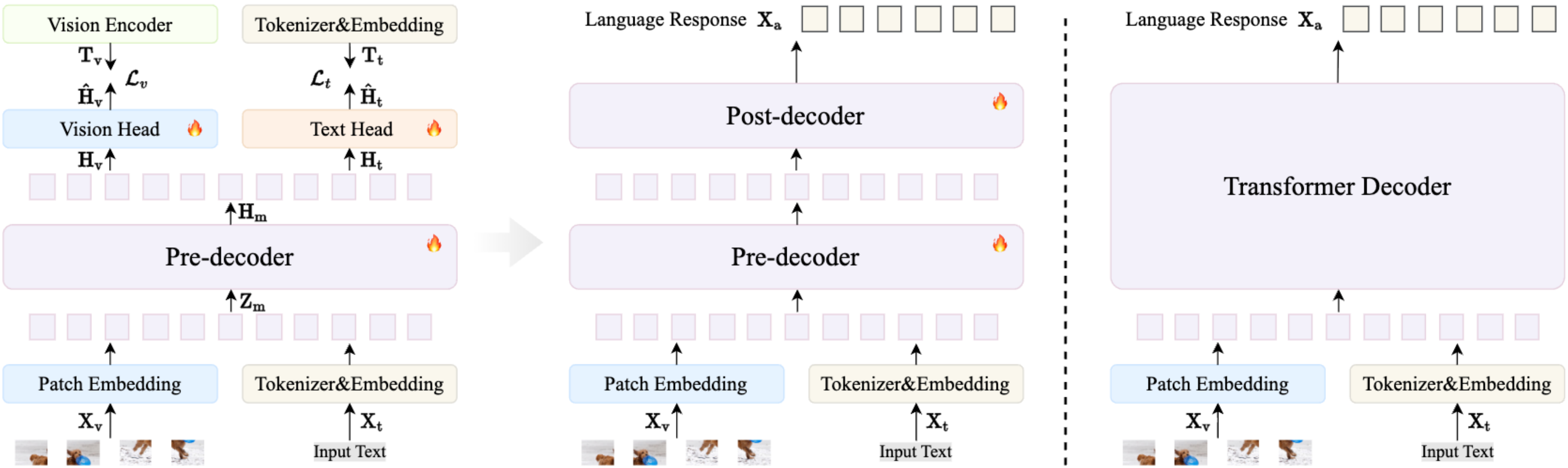
Text loss:

$$\mathcal{L}_{feat} = 1+\frac{1}{S}\sum_{i=1}^{S}[\|\hat{\mathrm{H}}_{\mathrm{t},i}-\mathrm{T}_{\mathrm{t},i}\|_2-\cos(\hat{\mathrm{H}}_{\mathrm{t},i};\mathrm{T}_{\mathrm{t},i})]$$

$$\mathcal{L}_{ctp} = -\frac{1}{S}\sum_{i=1}^{S}\sum_{c=1}^{C} y_{i,c}log(\frac{e^{\frac{x_{i,c}}{\tau}}}{\sum_{j=1}^{C} e^{\frac{x_{i,j}}{\tau}}})$$

Approach

- Training Receipt

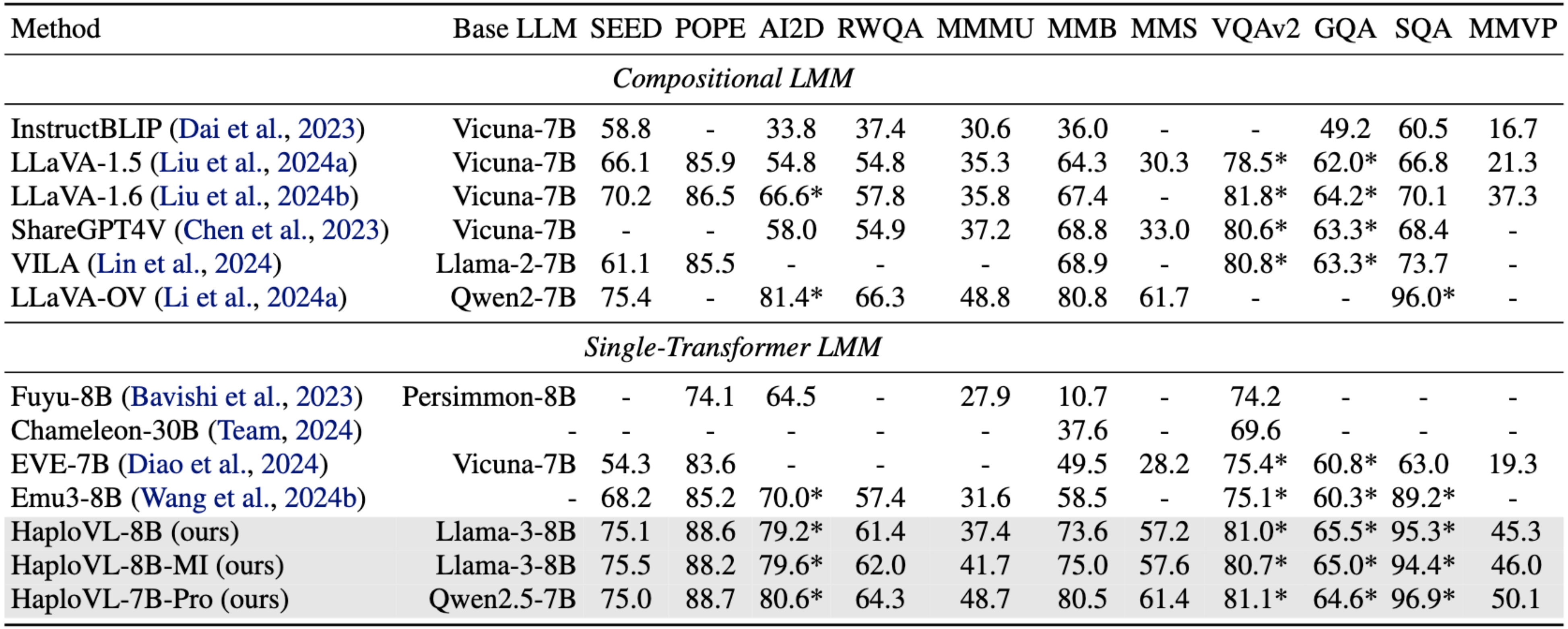


(a) Stage 1: Pre-training

(b) Stage 2: Fully finetuning

(c) Architecture

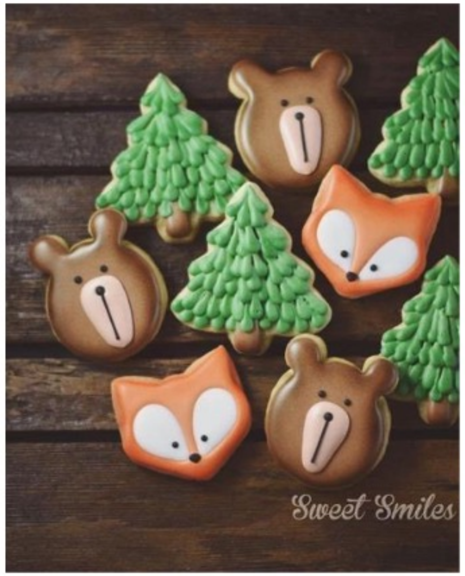
Experiment



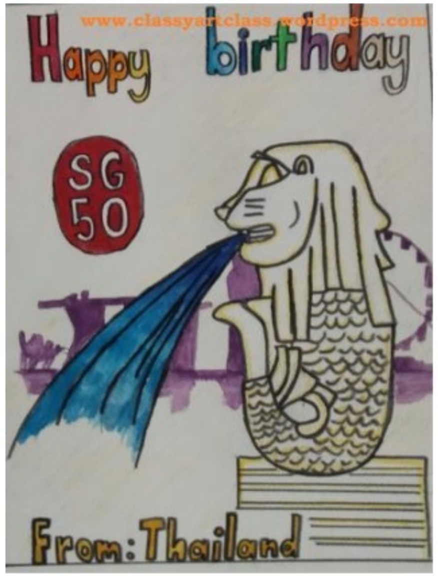
Method	Base LLM	SEED	POPE	AI2D	RWQA	MMMU	MMB	MMS	VQAv2	GQA	SQA	MMVP
Compositional LMM												
InstructBLIP (Dai et al., 2023)	Vicuna-7B	58.8	-	33.8	37.4	30.6	36.0	-	-	49.2	60.5	16.7
LLaVA-1.5 (Liu et al., 2024a)	Vicuna-7B	66.1	85.9	54.8	54.8	35.3	64.3	30.3	78.5*	62.0*	66.8	21.3
LLaVA-1.6 (Liu et al., 2024b)	Vicuna-7B	70.2	86.5	66.6*	57.8	35.8	67.4	-	81.8*	64.2*	70.1	37.3
ShareGPT4V (Chen et al., 2023)	Vicuna-7B	-	-	58.0	54.9	37.2	68.8	33.0	80.6*	63.3*	68.4	-
VILA (Lin et al., 2024)	Llama-2-7B	61.1	85.5	-	-	-	68.9	-	80.8*	63.3*	73.7	-
LLaVA-OV (Li et al., 2024a)	Qwen2-7B	75.4	-	81.4*	66.3	48.8	80.8	61.7	-	-	96.0*	-
Single-Transformer LMM												
Fuyu-8B (Bavishi et al., 2023)	Persimmon-8B	-	74.1	64.5	-	27.9	10.7	-	74.2	-	-	-
Chameleon-30B (Team, 2024)	-	-	-	-	-	-	37.6	-	69.6	-	-	-
EVE-7B (Diao et al., 2024)	Vicuna-7B	54.3	83.6	-	-	-	49.5	28.2	75.4*	60.8*	63.0	19.3
Emu3-8B (Wang et al., 2024b)	-	68.2	85.2	70.0*	57.4	31.6	58.5	-	75.1*	60.3*	89.2*	-
HaploVL-8B (ours)	Llama-3-8B	75.1	88.6	79.2*	61.4	37.4	73.6	57.2	81.0*	65.5*	95.3*	45.3
HaploVL-8B-MI (ours)	Llama-3-8B	75.5	88.2	79.6*	62.0	41.7	75.0	57.6	80.7*	65.0*	94.4*	46.0
HaploVL-7B-Pro (ours)	Qwen2.5-7B	75.0	88.7	80.6*	64.3	48.7	80.5	61.4	81.1*	64.6*	96.9*	50.1

Experiment

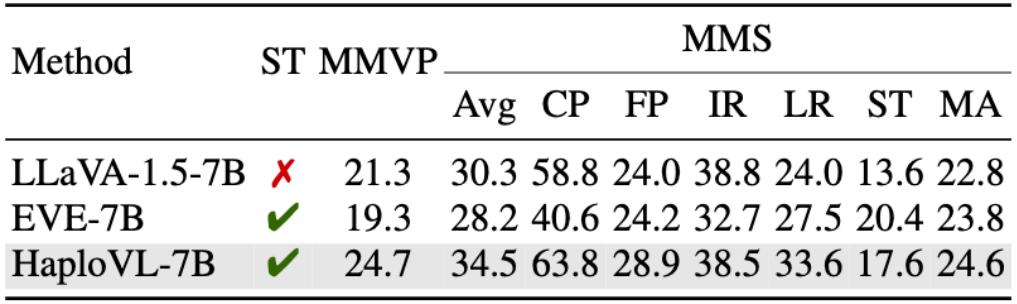
- Same SFT data comparison



Q: How many colors are the eyes of the depicted animals?
A: Two, B: One,
C: Three, D: Four
LLaVA: D ×
Haplo: B √



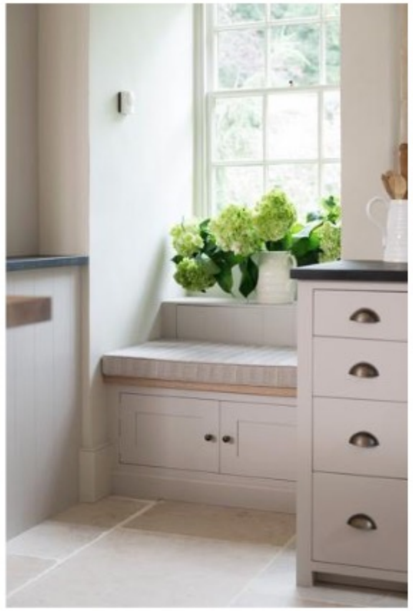
Q: What is the color of the letter in the red circle?
A: Black, B: White,
C: Red, D: Yellow
LLaVA: C ×
Haplo: B √



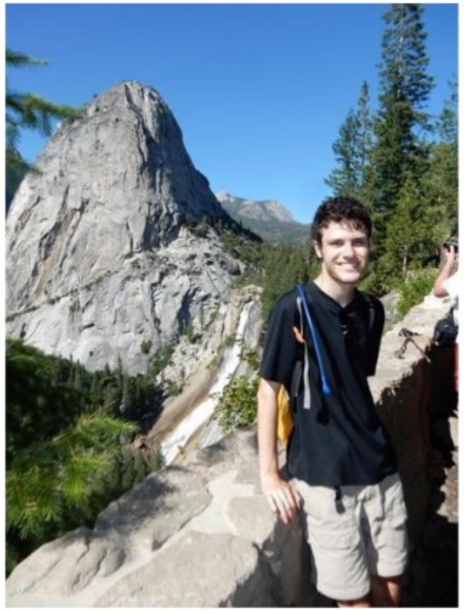
Method	ST	MMVP	MMS						
			Avg	CP	FP	IR	LR	ST	MA
LLaVA-1.5-7B	✗	21.3	30.3	58.8	24.0	38.8	24.0	13.6	22.8
EVE-7B	✓	19.3	28.2	40.6	24.2	32.7	27.5	20.4	23.8
HaploVL-7B	✓	24.7	34.5	63.8	28.9	38.5	33.6	17.6	24.6

Experiment

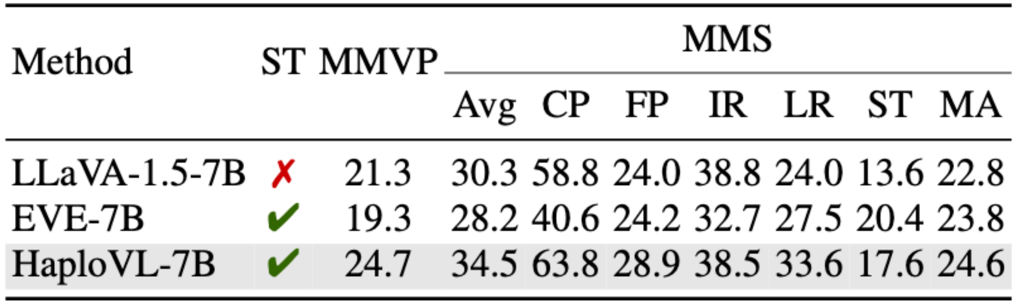
- Same SFT data comparison



Q: How many drawers can be seen in the white kitchen?
A: Three, B: Two,
C: One, D: Four
LLaVA: A ×
Haplo: D √



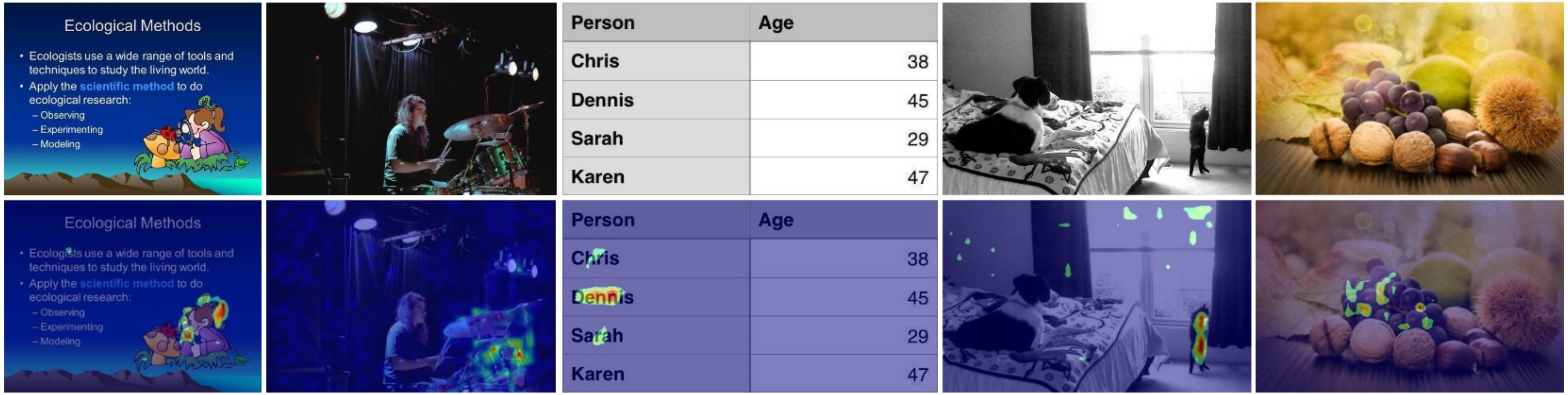
Q: How many people are visible in the image?
A: One, B: Two,
C: Four, D: Three
LLaVA: A ×
Haplo: B √



Method	ST	MMVP	MMS						
			Avg	CP	FP	IR	LR	ST	MA
LLaVA-1.5-7B	✗	21.3	30.3	58.8	24.0	38.8	24.0	13.6	22.8
EVE-7B	✔	19.3	28.2	40.6	24.2	32.7	27.5	20.4	23.8
HaploVL-7B	✔	24.7	34.5	63.8	28.9	38.5	33.6	17.6	24.6

Experiment

- Visualize for the early fusion mechanism



Ecological Methods

- Ecologists use a wide range of tools and techniques to study the living world.
- Apply the scientific method to do ecological research:
  - – Observing
  - – Experimenting
  - – Modeling

Person	Age
Chris	38
Dennis	45
Sarah	29
Karen	47

Q: How many people are there in the image?
HaploVL: 1

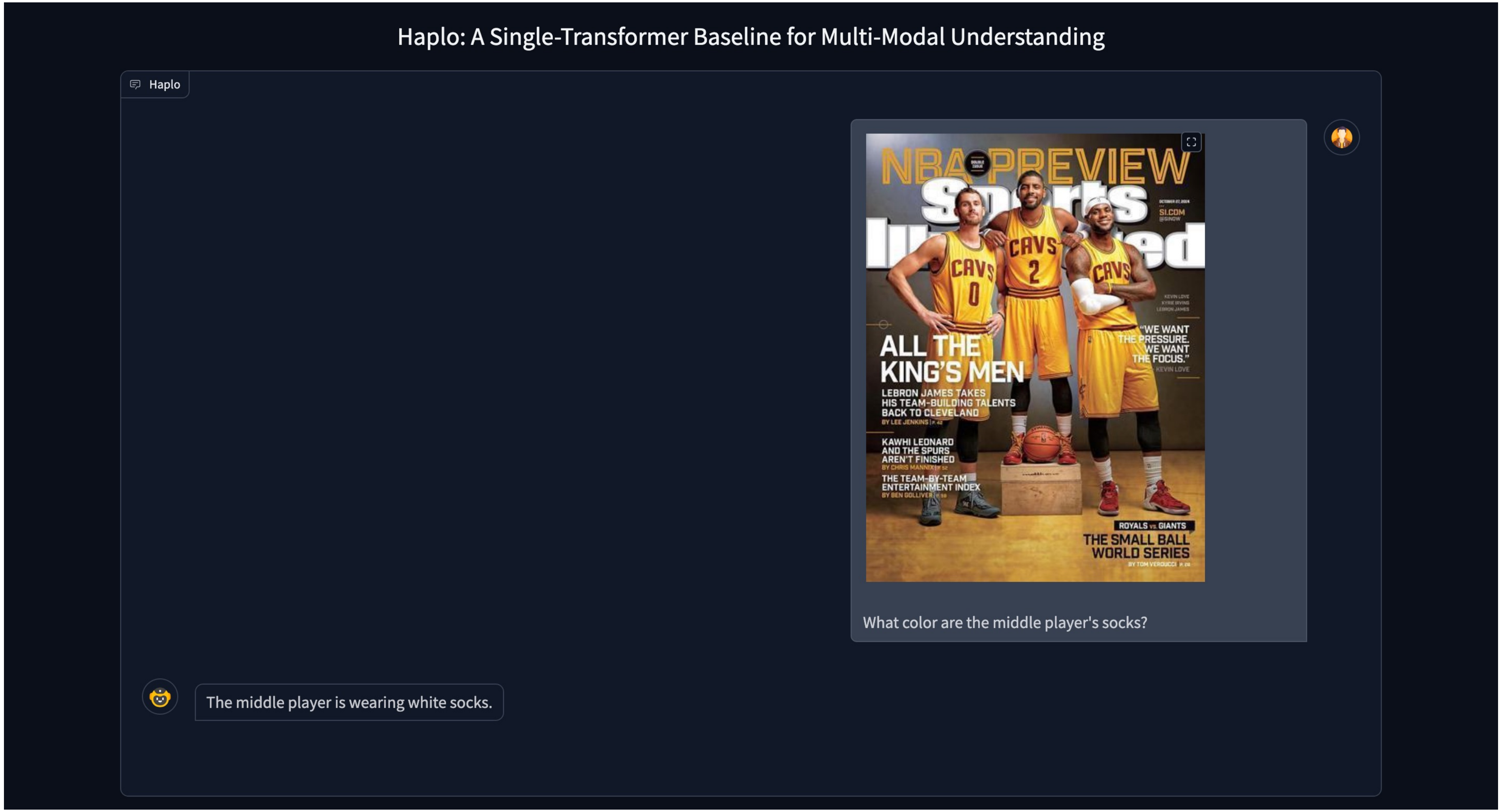
Q: Where is the drum located in the image?
HaploVL: On the right side

Q: According to this picture, how old are Dennis.
HaploVL: 45

Q: where is the cat?
HaploVL: bottom-right

Q:What is the position of the grape in relation to the plate of walnuts?
HaploVL: The grape is on top of the plate of walnuts

Single Image Understanding



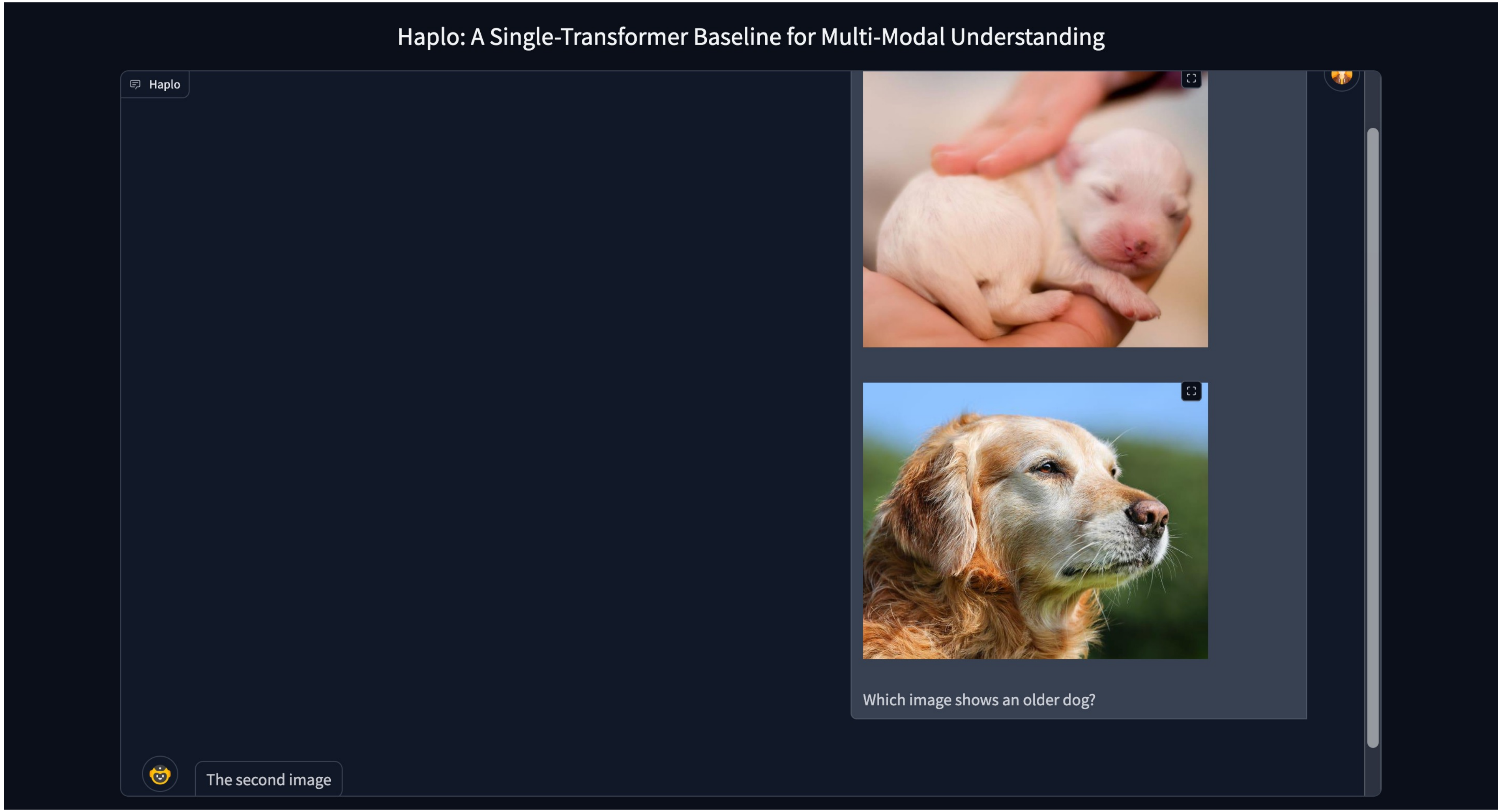
Haplo: A Single-Transformer Baseline for Multi-Modal Understanding

Haplo
NBA PREVIEW
Sports Illustrated
CAVS
ALL THE KING'S MEN
LEBRON JAMES TAKES HIS TEAM-BUILDING TALENTS BACK TO CLEVELAND
KAWHI LEONARD AND THE SPURS AREN'T FINISHED
THE TEAM-BY-TEAM ENTERTAINMENT INDEX
"WE WANT THE PRESSURE. WE WANT THE FOCUS."
ROYALS vs. GIANTS
THE SMALL BALL WORLD SERIES

What color are the middle player's socks?

The middle player is wearing white socks.

Multi-Image Understanding

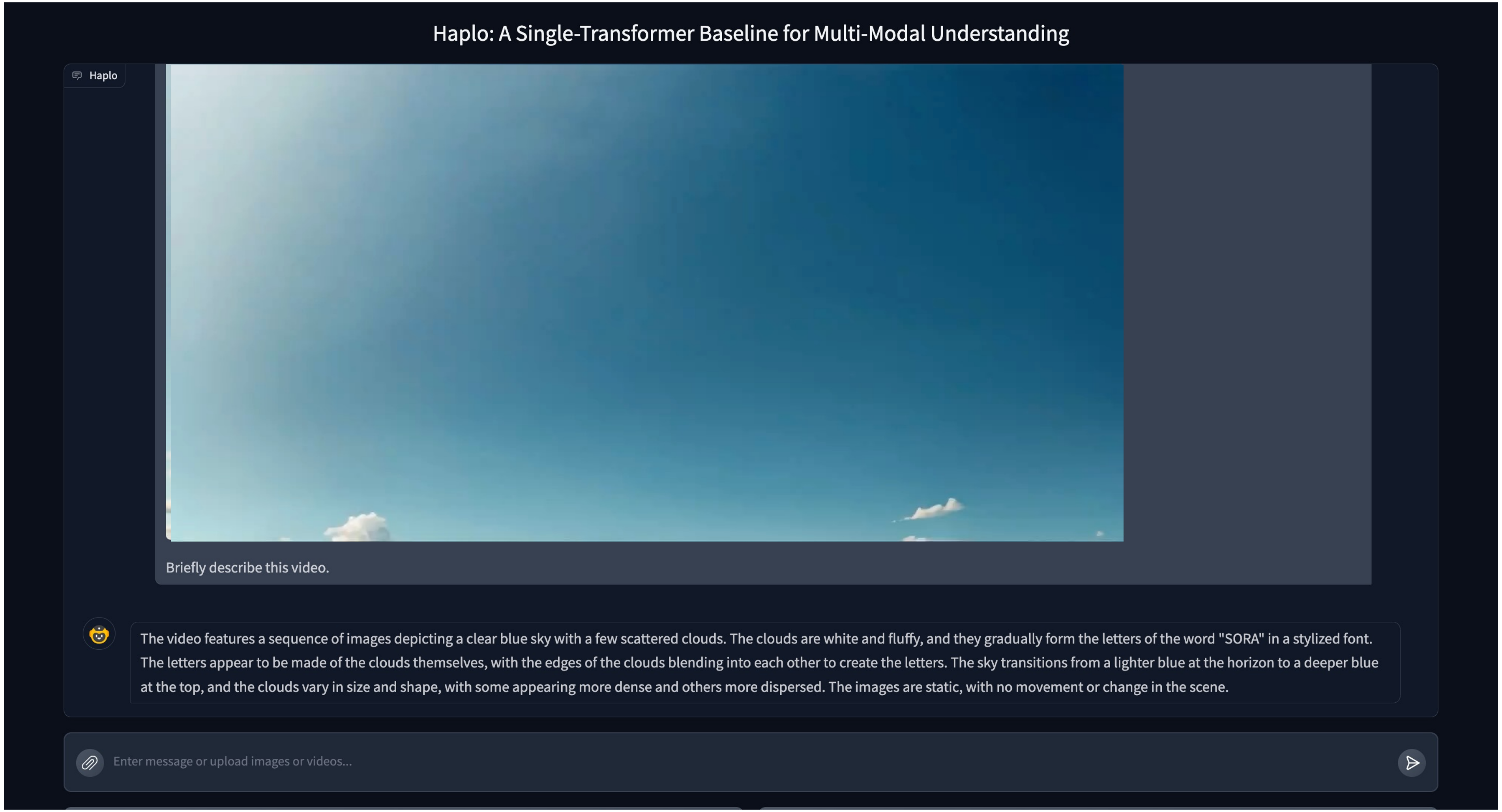


Haplo: A Single-Transformer Baseline for Multi-Modal Understanding

Which image shows an older dog?

The second image

Video Understanding



Haplo: A Single-Transformer Baseline for Multi-Modal Understanding

Briefly describe this video.

The video features a sequence of images depicting a clear blue sky with a few scattered clouds. The clouds are white and fluffy, and they gradually form the letters of the word "SORA" in a stylized font. The letters appear to be made of the clouds themselves, with the edges of the clouds blending into each other to create the letters. The sky transitions from a lighter blue at the horizon to a deeper blue at the top, and the clouds vary in size and shape, with some appearing more dense and others more dispersed. The images are static, with no movement or change in the scene.

Thanks