Introduction to Hallucination Mitigation in Multimodal Large Language Models

2026.08.21

Data **M**ining & **Q**uality **A**nalytics Lab.

김 혜 준

About Me

❖ 김혜준(Kim Hyejun)

- 고려대학교 산업경영공학과 석사과정
- Data Mining & Quality Analytics Lab. (김성범 교수님)

❖ 관심 분야

- Multimodal Large Language Models
- Hallucination Mitigation

❖ Contact

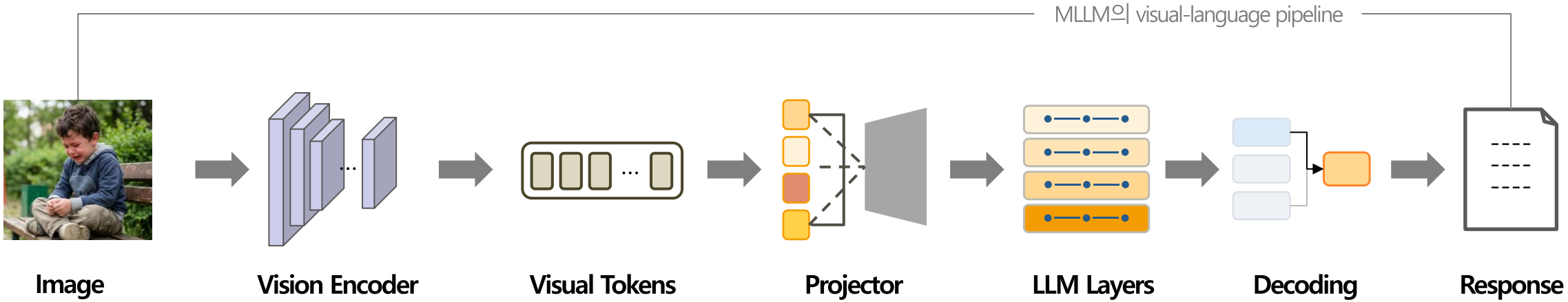
- hj04036@korea.ac.kr

Introduction

Background

❖ MLLM은 이미지를 어떻게 이해하고 답변할까?

- 시각 정보를 추출하고 LLM이 처리할 수 있는 표현으로 변환
- LLM Layers에서 시각·언어 정보를 통합하여 처리
- Decoding을 통해 최종 자연어 응답 생성



Introduction

Background

❖ MLLM은 항상 이미지에 근거해 답변할까?

- 자연스럽고 자세한 문장이라고 해서 이미지 근거와 일치하는 것은 아님
- 특히 복잡한 장면에서 visual evidence와 불일치하는 내용을 생성할 수 있음

MLLM description

?

A boy is crying
because he lost his parents.

이미지에서 확인되지
않는 정보

Introduction

Background

❖ MLLM Hallucination이란 무엇일까?

- **MLLM Hallucination** : 입력 이미지의 **실제 시각적 근거와 일치하지 않는 내용**을 생성하는 현상
- Fluent Response ≠ Visually Grounded Response

MLLM description

A boy is crying
because he lost his parents.

?

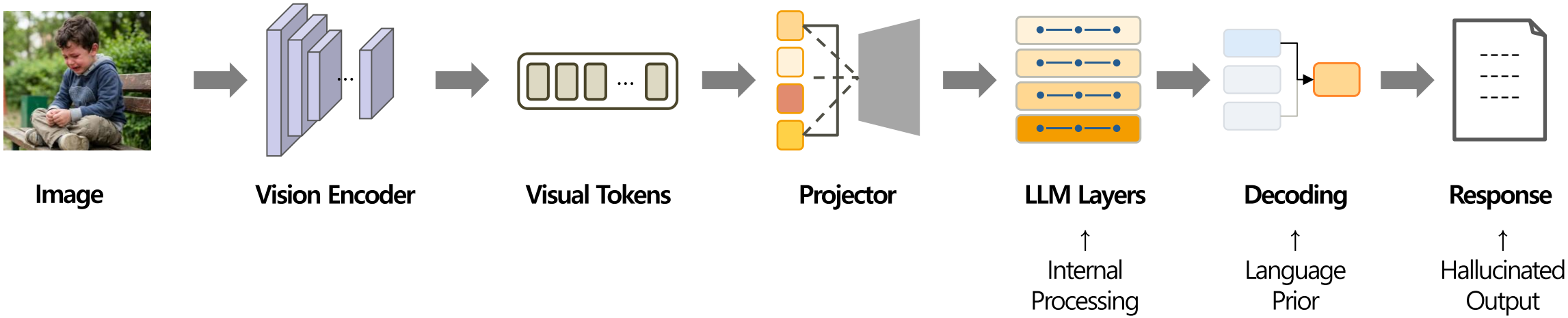
이미지에서 확인되지
않는 정보

Introduction

Background

❖ Hallucination은 최종 응답에서만 발생하는 문제일까?

- **Visual Grounding :** 이미지의 시각적 근거가 충분히 반영되는가?
- **Language Prior :** 시각 정보보다 학습된 언어적 패턴에 과도하게 의존하는가?
- **Internal Processing :** 시각 정보가 LLM layer를 거치며 어떻게 유지·변형되는가?

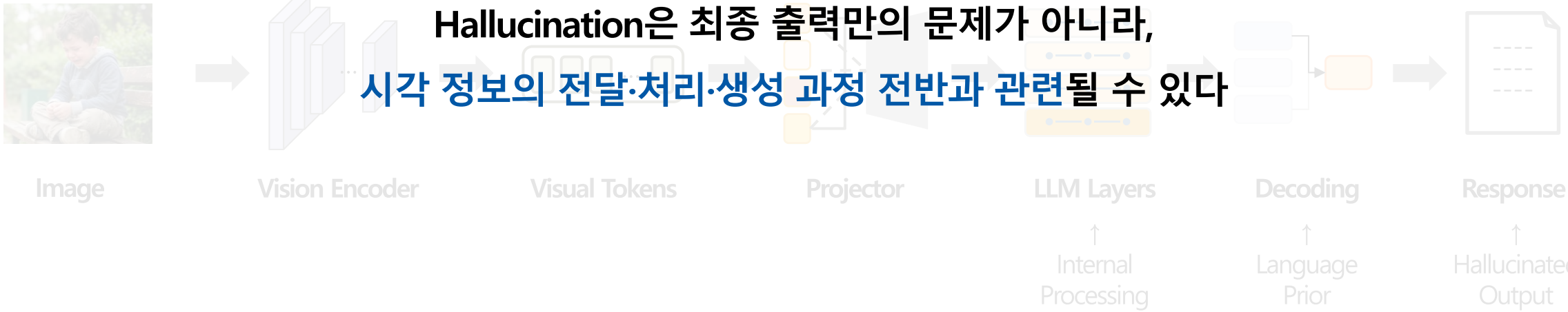


Introduction

Background

❖ Hallucination은 최종 응답에서만 발생하는 문제일까?

- **Visual Grounding :** 이미지의 시각적 근거가 충분히 반영되는가?
- **Language Prior :** 시각 정보보다 학습된 언어적 패턴에 과도하게 의존하는가?
- **Internal Processing :** 시각 정보가 LLM layer를 거치며 어떻게 유지·변형되는가?



Hallucination은 최종 출력만의 문제가 아니라,

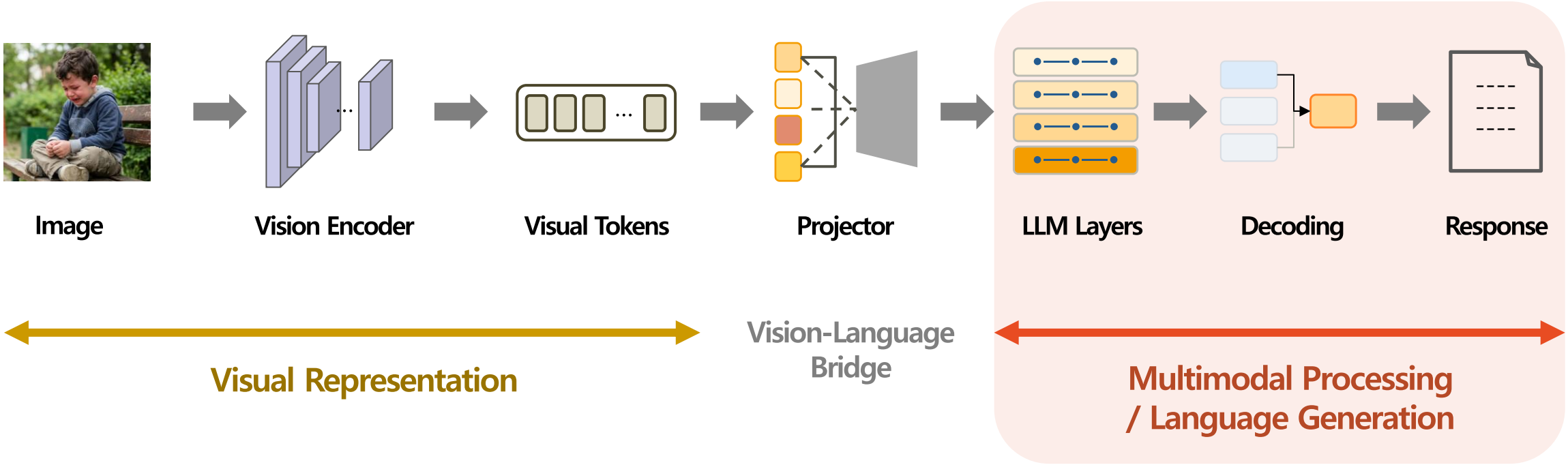
시각 정보의 전달·처리·생성 과정 전반과 관련될 수 있다

Introduction

Background

❖ 그렇다면 Hallucination을 어디에서 완화할 수 있을까?

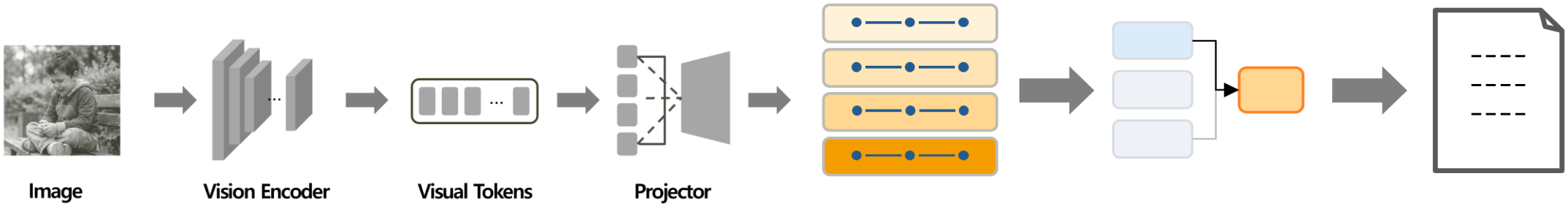
- Hallucination은 MLLM의 전 과정과 관련될 수 있음
- 시각 정보의 표현 및 Vision-Language 연결 과정도 영향을 줄 수 있음
- 본 세미나에서는 LLM 내부의 시각 정보 활용과 생성 과정에 초점을 맞춤



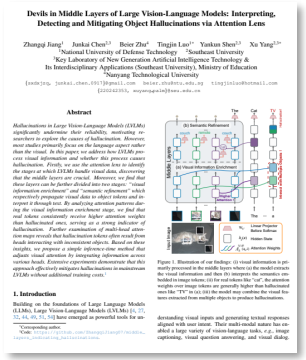
Introduction

Background

❖ 그렇다면 Hallucination을 어디에서 완화할 수 있을까?

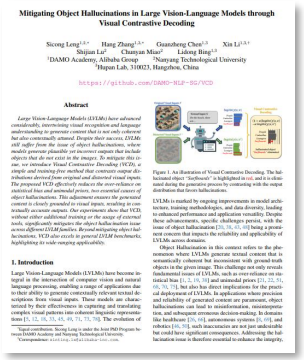


③ Devils in Middle Layers (CVPR 2025)



Why hallucinate?

② VCD (CVPR 2024)



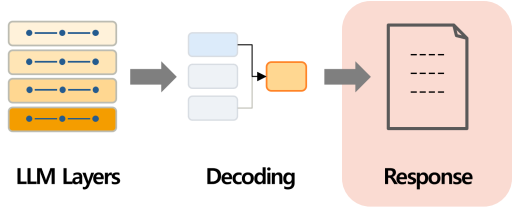
How to generate?

① RLHF-V (CVPR 2024)

What to say?

1. RLHF-V

Behavior Alignment



❖ RLHF-V : Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback (CVPR 2024)

- 기존 preference는 Hallucination의 정확한 위치를 알려주기 어려움
- RLHF-V는 Hallucinated segment를 직접 교정해 생성 행동을 정렬

RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback

Tianyu Yu[1] Yuan Yao[2*] Haoye Zhang[1] Taiwen He[1] Yifeng Han[1]
Ganqu Cui[1] Jinyi Hu[1] Zhiyuan Liu[1*] Hai-Tao Zheng[134*] Maosong Sun[1]
[1]Tsinghua University [2]National University of Singapore
[3]Shenzhen International Graduate School, Tsinghua University
[4]Pengcheng Laboratory, Shenzhen, China
yiranytianyu@gmail.com yaoyuanthu@gmail.com

https://rlhf-v.github.io

Abstract

Multimodal Large Language Models (MLLMs) have recently demonstrated impressive capabilities in multimodal understanding, reasoning, and interaction. However, existing MLLMs prevalently suffer from serious hallucination problems, generating text that is not factually grounded in associated images. The problem makes existing MLLMs untrustworthy and thus impractical in real-world (especially high-stakes) applications. To address the challenge, we present RLHF-V, which enhances MLLM trustworthiness via behavior alignment from fine-grained correctional human feedback. Specifically, RLHF-V collects human preference in the form of segment-level corrections on hallucinations, and performs dense direct preference optimization over the human feedback. Comprehensive experiments on five benchmarks in both automatic and human evaluation show that, RLHF-V can enable substantially more trustworthy MLLM behaviors with promising data and computation efficiency. Remarkably, using 1.4k annotated data samples, RLHF-V significantly reduces the hallucination rate of the base MLLM by 34.8%, outperforming the concurrent LLaVA-RLHF trained on 10k annotated data. The final model achieves state-of-the-art performance in trustworthiness among open-source MLLMs, and shows better robustness than GPT-4V in preventing hallucinations aroused from over-generalization.

1. Introduction

The recent success of Multimodal Large Language Models (MLLMs) marks a significant milestone in AI research [2, 4, 11, 12, 19, 27, 29, 42, 51]. By connecting visual signals and Large Language Models (LLMs), MLLMs show unprecedented capabilities in multimodal understanding, reasoning, and interaction [28, 29, 44]. The models are typically pre-trained on large-scale image-text data to learn the foundational multimodal knowledge and capabilities [2, 4, 12, 19]. To steer the model behavior, most MLLMs are further fine-tuned with instruction tuning (also known as supervised fine-tuning), which supervises models to clone the behavior from demonstration data, enabling MLLMs to engage users in various open-world tasks [4, 11, 25, 27, 47].

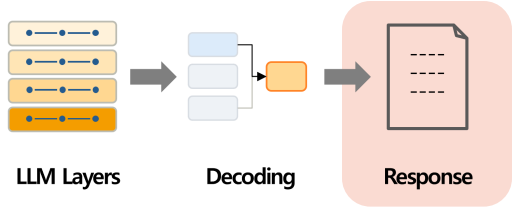
However, current MLLM behaviors are not well aligned with human preferences. A glaring issue is their tendency to produce *hallucinations* — responses that are not factually grounded in the associated images [21, 25, 29, 37]. This typically includes descriptions of non-existing visual contents and errors in descriptions. As shown in Figure 1, current MLLMs can hallucinate about objects, attributes, numbers, positions, actions, etc. Quantitatively, our human evaluation shows that the problem is prevalent among state-of-the-art MLLMs, where even the most advanced GPT-4V [29] contains obvious hallucinations in 45.9% responses. The problem makes existing MLLMs untrustworthy and thus impractical in real-world (especially high-stakes) applications, such as guiding visually impaired individuals [29] or autonomous driving systems [43].

We argue that the problem arises from the lack of positive/negative human feedback in instruction-tuned models, making it challenging to learn the precise behavior boundaries to exclude hallucination. To address the problem, we propose RLHF-V, a novel framework that aligns MLLM behavior by learning from human feedback. However, simply applying traditional Reinforcement Learning from Hu-

*Corresponding authors

1. RLHF-V

Behavior Alignment



❖ RLHF-V는 Hallucination을 어떻게 교정할까?

- Hallucinated segment를 직접 교정하고, 수정된 응답을 바탕으로 DDPO를 수행

Prompt : Please describe the image in detail.

Inputs

A

The image shows a boy crying on a bench. He seems sad **because he lost his parents.** A swing set is visible in the background.

B

A young boy is sitting alone on a bench outdoors. He is crying, and a woman is standing near **the swings watching him.** Trees and a swing set can be seen in the background.

MLLM outputs

A*

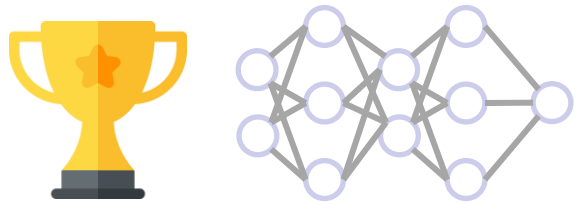
The image shows a boy crying on a bench. He seems sad **and is looking down at his hands.** A swing set is visible in the background.

B*

A young boy is sitting alone on a bench outdoors. He is crying, and **the swing set behind him is empty.** Trees and a swing set can be seen in the background.

Fine-grained Correctional Human Feedback

Dense Direct Preference Optimization (DDPO)

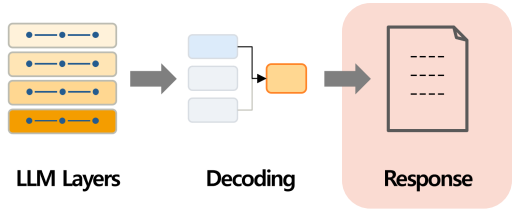


Final MLLM

(1) Human Feedback Collection → **(2) Human Preference Learning**

1. RLHF-V

Behavior Alignment



❖ 왜 전체 응답이 아니라, 수정된 부분에 집중할까?

- 기존 preference는 응답 전체의 선호만 비교
- RLHF-V는 Hallucinated segment를 직접 교정해 오류 위치와 수정 방향을 명확히 제공

A

The image shows a boy crying on a bench. He seems sad **because he lost his parents.** A swing set is visible in the background.

B

A young boy is sitting alone on a bench outdoors. He is crying, and a woman is standing near **the swings watching him.** Trees and a swing set can be seen in the background.

MLLM outputs

A*

The image shows a boy crying on a bench. He seems sad **and is looking down at his hands.** A swing set is visible in the background.

B*

A young boy is sitting alone on a bench outdoors. He is crying, and **the swing set behind him is empty.** Trees and a swing set can be seen in the background.

Fine-grained Correctional Human Feedback

기존

- **Response A > Response B**
- 전체 응답 단위 선호 → 오류 위치는 불명확

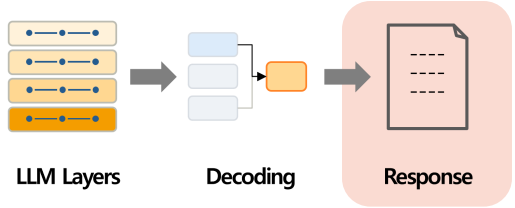
RLHF-V

- **Hallucinated segment → Corrected segment**
- 잘못된 부분만 직접 교정
- 오류 위치와 수정 방향이 명확

오류 위치가 명확한 Fine-grained Human Feedback

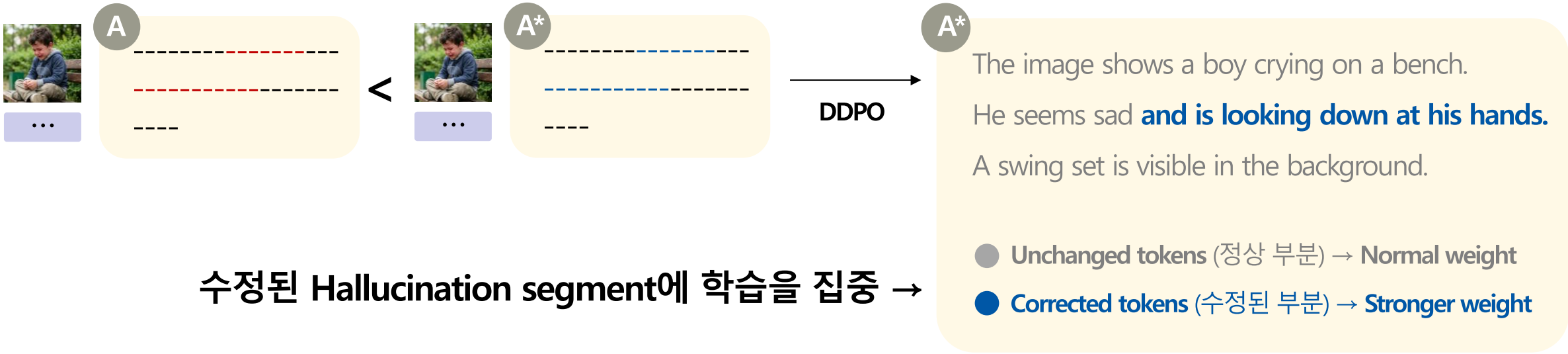
1. RLHF-V

Behavior Alignment



❖ 수정된 부분을 더 강하게 학습한다면?

- Original–Corrected response를 preference pair로 구성
- Corrected segment에 더 강한 optimization weight를 부여

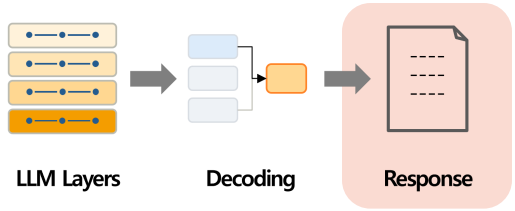


수정된 Hallucination segment에 학습을 집중 →

Final MLLM

1. RLHF-V

Behavior Alignment



❖ Fine-grained Alignment는 Hallucination을 실제로 줄였을까?

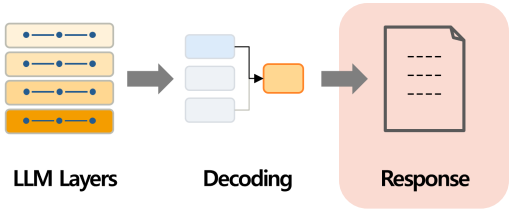
- Key Results
  - ➢ Object Hallucination 크게 감소
  - ➢ Fine-grained feedback으로 기존 RLHF보다 낮은 Hallucination
  - ➢ 적은 correction data로도 효과적인 alignment

Model	Object HalBench ↓		MHumanEval ↓				MMHal-Bench		LLaVA Bench			VQAv2
	Resp.	Mention	Object	Position	Number	All	Info.	Resp.↓	Conv.	Detail	Comp.	testdev
LLaVA [27]	63.0	29.5	46.6	21.2	19.9	80.8	31.9	70.8	85.4	74.3	96.3	-
Muffin [47]	50.5	24.5	33.6	16.4	26.0	74.7	33.4	68.8	89.3	**79.7**	97.7	-
LRV [25]	32.3	22.3	43.2	11.6	19.2	82.9	22.2	78.1	61.7	47.3	55.0	-
LLaVA-RLHF [37]	38.1	18.9	37.7	17.8	18.5	72.6	39.9	65.6	**93.8**	74.3	**111.4**	-
InstructBLIP [11]	25.9	14.3	30.8	15.1	17.1	63.7	29.5	64.4	83.2	67.6	90.6	-
Qwen-VL-Chat [4]	43.8	20.0	34.9	16.4	15.8	61.0	38.5	**52.1**	81.9	77.1	92.3	79.5
LLaVA 1.5 [26]	46.3	22.6	30.8	17.8	17.1	61.0	39.2	**52.1**	81.6	75.5	95.2	**80.0**
RLHF-V	**12.2**	**7.5**	**21.9**	**7.5**	**14.4**	**55.5**	**40.0**	**52.1**	93.1	75.3	91.6	**80.0**
GPT-4V [29]	13.6	7.3	22.6	12.3	11.0	45.9	47.6	31.3	96.0	102.5	106.7	77.2*

Fine-grained Feedback → Less Hallucination

1. RLHF-V

Behavior Alignment



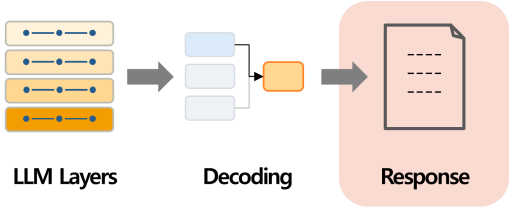
❖ Fine-grained Alignment는 Hallucination을 실제로 줄였을까?

- Key Results
  - ➢ Object Hallucination 크게 감소
  - ➢ Fine-grained feedback으로 기존 RLHF보다 낮은 Hallucination
  - ➢ 적은 correction data로도 효과적인 alignment

- Limitation
  - ➢ Human correction annotation 필요
  - ➢ 추가 post-training 필요

1. RLHF-V

Behavior Alignment



❖ Fine-grained Alignment는 Hallucination을 실제로 줄였을까?

- Key Results
  - Object Hallucination 크게 감소
  - Fine-grained feedback으로 기존 RLHF보다 낮은 Hallucination
  - 적은 correction data로도 효과적인 alignment
- Limitation
  - 그렇다면 추가 Human Feedback과 재학습 없이도 Hallucination을 줄일 수 있을까?
  - 추가 post-training 필요

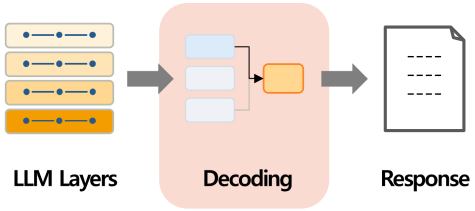
그렇다면 추가 Human Feedback과 재학습 없이도 Hallucination을 줄일 수 있을까?

→ VCD (CVPR 2024)

Model	Object HalBench ↓			MHumanEval ↓			MMHal-Bench		LLaVA Bench			VQAv2
					Number	All	Info.	Resp.↓	Conv.	Detail	Comp.	testdev
LLaVA [27]	63.0	29.5	46.6	21.2	19.9	80.8	31.9	70.8	85.4	74.3	96.3	-
Muffin [47]	50.5	24.5	33.6	16.4	26.0	74.7	33.4	68.8	89.3	**79.7**	97.7	-
LRV [25]	32.3	22.3	43.2	11.6	19.2	82.9	22.2	78.1	61.7	47.3	55.0	-
LLaVA-RLHF [37]	38.1	18.9	37.7	17.8	18.5	72.6	39.9	65.6	**93.8**	74.3	**111.4**	-
InstructBLIP [11]	25.9	14.3	30.8	15.1	17.1	63.7	29.5	64.4	83.2	67.6	90.6	-
Qwen-VL-Chat [4]	43.8	20.0	34.9	16.4	15.8	61.0	38.5	**52.1**	81.9	77.1	92.3	79.5
LLaVA 1.5 [26]	46.3	22.6	30.8	17.8	17.1	61.0	39.2	**52.1**	81.6	75.5	95.2	**80.0**
RLHF-V	**12.2**	**7.5**	**21.9**	**7.5**	**14.4**	**55.5**	**40.0**	**52.1**	93.1	75.3	91.6	**80.0**
GPT-4V [29]	13.6	7.3	22.6	12.3	11.0	45.9	47.6	31.3	96.0	102.5	106.7	77.2*

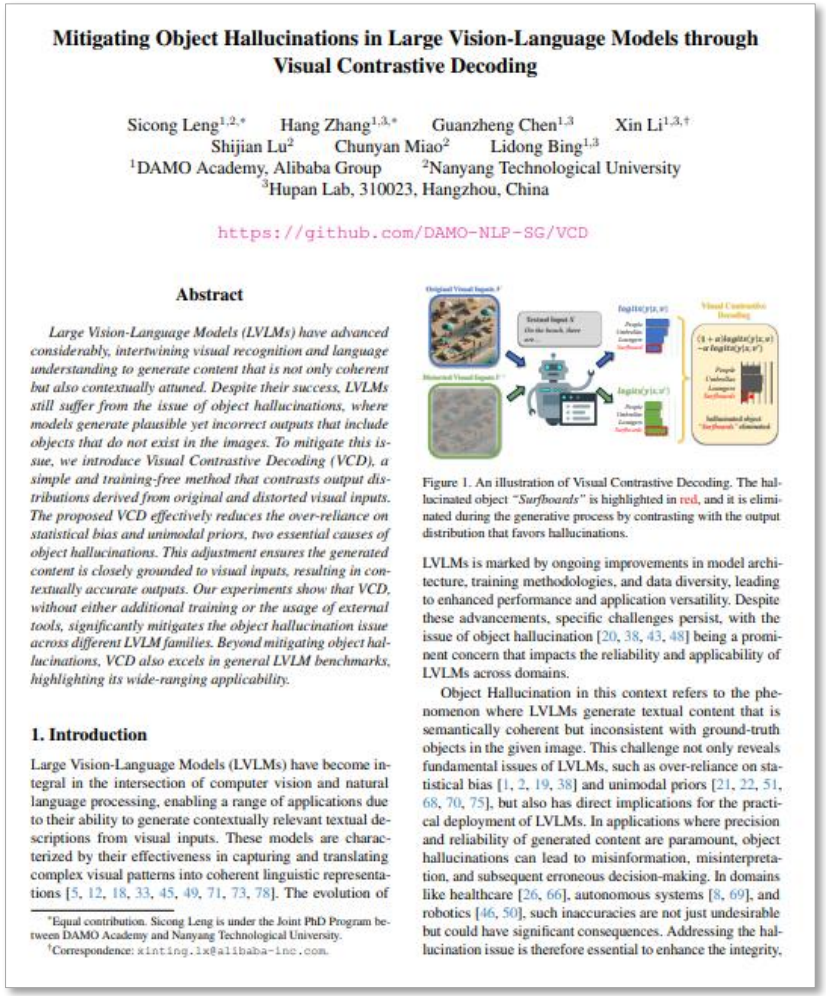
2. VCD

Decoding Intervention



❖ Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding (CVPR 2024)

- 추가 학습 없이 Decoding 단계에서 Hallucination을 완화
- Original / Distorted visual input의 prediction 차이를 활용



Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding

Sicong Leng[1,2,*] Hang Zhang[1,3,*] Guanzheng Chen[1,3] Xin Li[1,3,†]
Shijian Lu[2] Chunyan Miao[2] Lidong Bing[1,3]
[1]DAMO Academy, Alibaba Group [2]Nanyang Technological University
[3]Hupan Lab, 310023, Hangzhou, China

https://github.com/DAMO-NLP-SG/VCD

Abstract

Large Vision-Language Models (LVLMs) have advanced considerably, intertwining visual recognition and language understanding to generate content that is not only coherent but also contextually attuned. Despite their success, LVLMs still suffer from the issue of object hallucinations, where models generate plausible yet incorrect outputs that include objects that do not exist in the images. To mitigate this issue, we introduce Visual Contrastive Decoding (VCD), a simple and training-free method that contrasts output distributions derived from original and distorted visual inputs. The proposed VCD effectively reduces the over-reliance on statistical bias and unimodal priors, two essential causes of object hallucinations. This adjustment ensures the generated content is closely grounded to visual inputs, resulting in contextually accurate outputs. Our experiments show that VCD, without either additional training or the usage of external tools, significantly mitigates the object hallucination issue across different LVLM families. Beyond mitigating object hallucinations, VCD also excels in general LVLM benchmarks, highlighting its wide-ranging applicability.

1. Introduction

Large Vision-Language Models (LVLMs) have become integral in the intersection of computer vision and natural language processing, enabling a range of applications due to their ability to generate contextually relevant textual descriptions from visual inputs. These models are characterized by their effectiveness in capturing and translating complex visual patterns into coherent linguistic representations [5, 12, 18, 33, 45, 49, 71, 73, 78]. The evolution of

*Equal contribution. Sicong Leng is under the Joint PhD Program between DAMO Academy and Nanyang Technological University.
†Correspondence: xinting.lx@alibaba-inc.com.

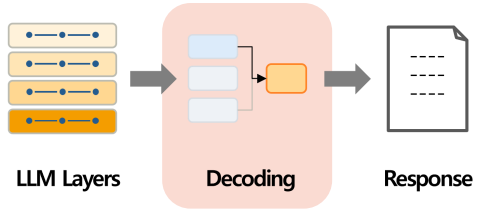
Figure 1. An illustration of Visual Contrastive Decoding. The hallucinated object *"Surfboards"* is highlighted in red, and it is eliminated during the generative process by contrasting with the output distribution that favors hallucinations.

LVLMs is marked by ongoing improvements in model architecture, training methodologies, and data diversity, leading to enhanced performance and application versatility. Despite these advancements, specific challenges persist, with the issue of object hallucination [20, 38, 43, 48] being a prominent concern that impacts the reliability and applicability of LVLMs across domains.

Object Hallucination in this context refers to the phenomenon where LVLMs generate textual content that is semantically coherent but inconsistent with ground-truth objects in the given image. This challenge not only reveals fundamental issues of LVLMs, such as over-reliance on statistical bias [1, 2, 19, 38] and unimodal priors [21, 22, 51, 68, 70, 75], but also has direct implications for the practical deployment of LVLMs. In applications where precision and reliability of generated content are paramount, object hallucinations can lead to misinformation, misinterpretation, and subsequent erroneous decision-making. In domains like healthcare [26, 66], autonomous systems [8, 69], and robotics [46, 50], such inaccuracies are not just undesirable but could have significant consequences. Addressing the hallucination issue is therefore essential to enhance the integrity,

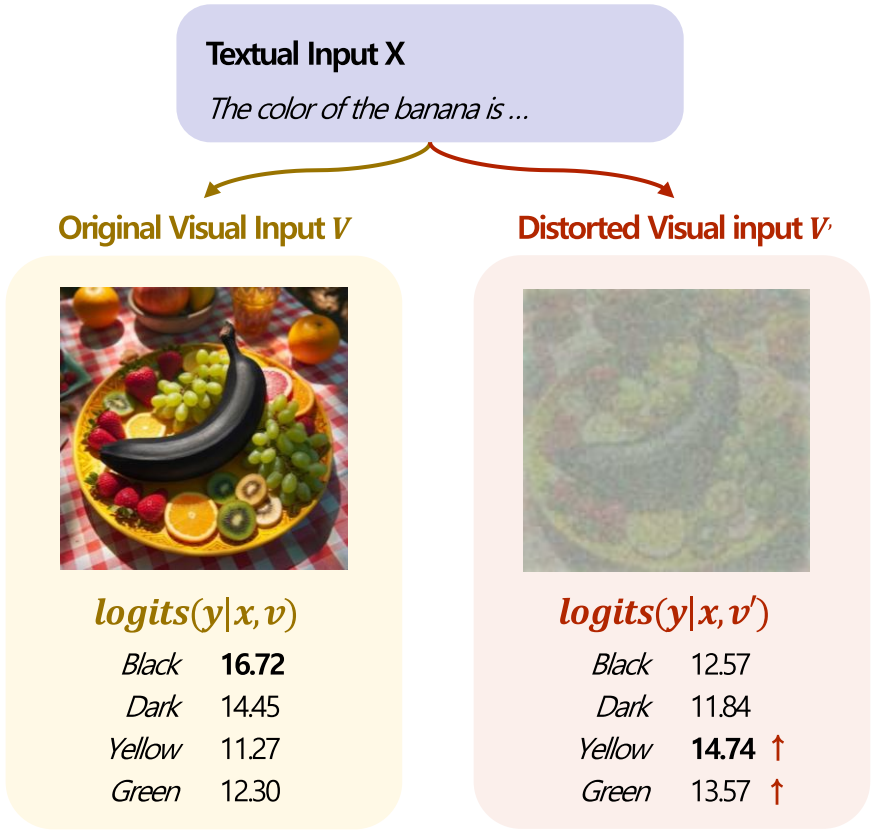
2. VCD

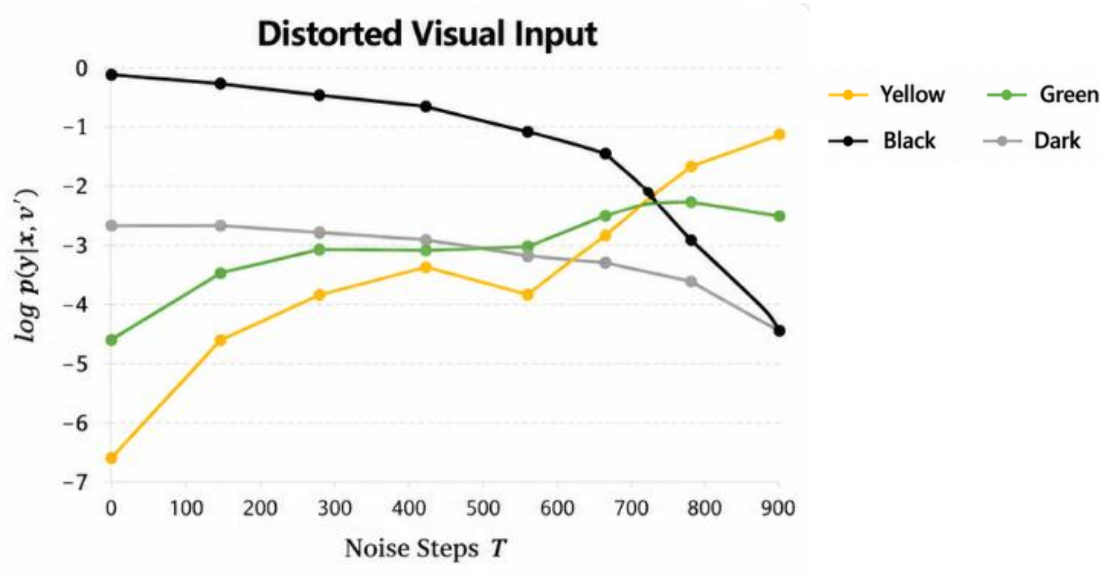
Decoding Intervention



❖ 시각 정보가 불확실해질수록 Language Prior가 강해질까?

- Visual Evidence ↓ → Language Prior / Statistical Bias ↑



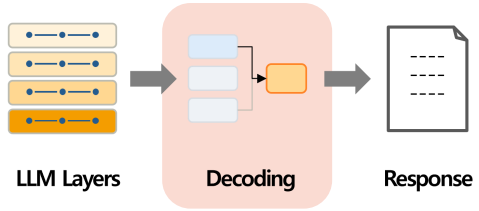


- Noise ↑ → **Black** probability ↓
- **Yellow** / **Green** 과 같은 typical color bias ↑

시각적 근거가 약해질수록 Language Prior의 영향이 커짐

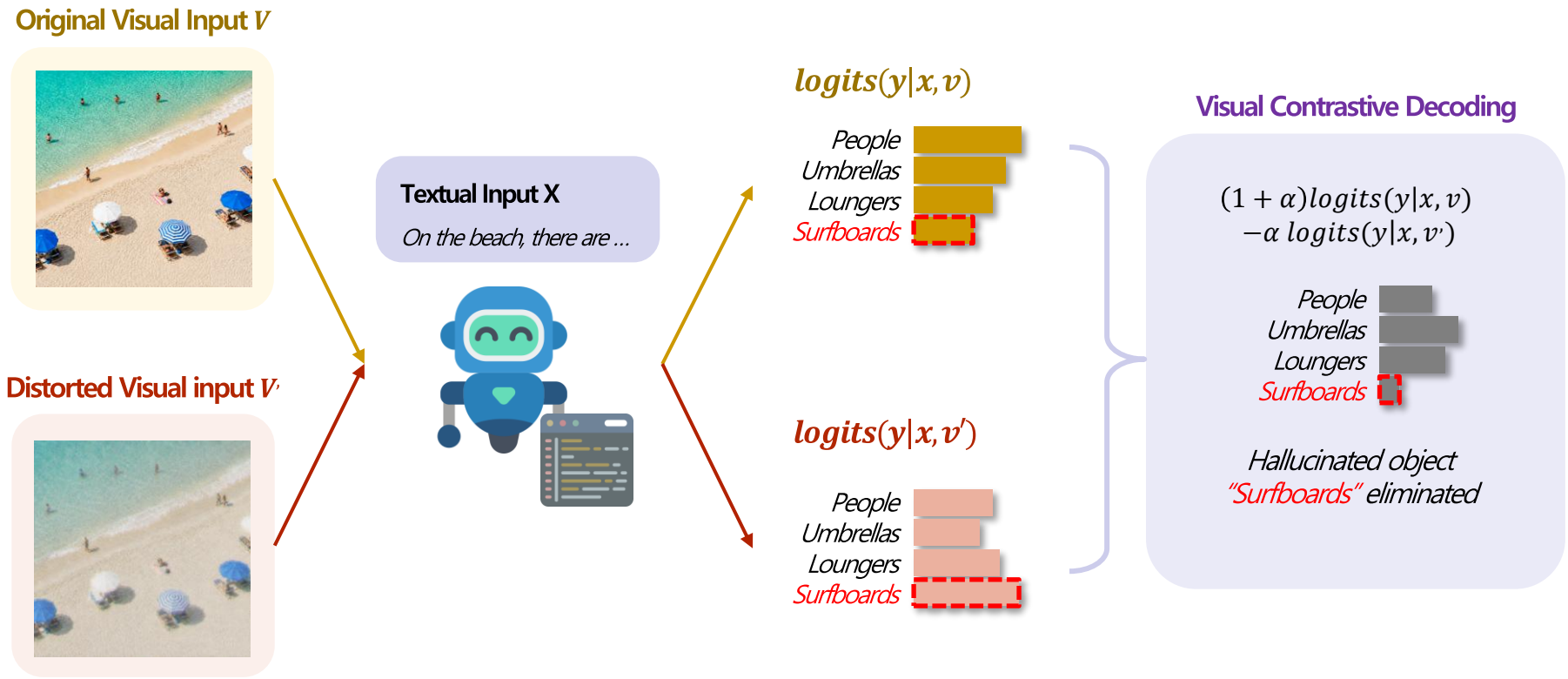
2. VCD

Decoding Intervention



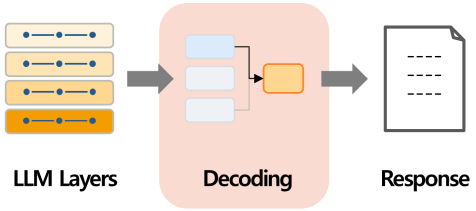
❖ VCD는 어떻게 Hallucination을 줄일까?

- 원본·손상 이미지에서 각각 output distribution 계산
- 원본·손상 이미지의 예측 차이를 이용해 시각적 근거는 강화하고, 언어적 편향은 억제



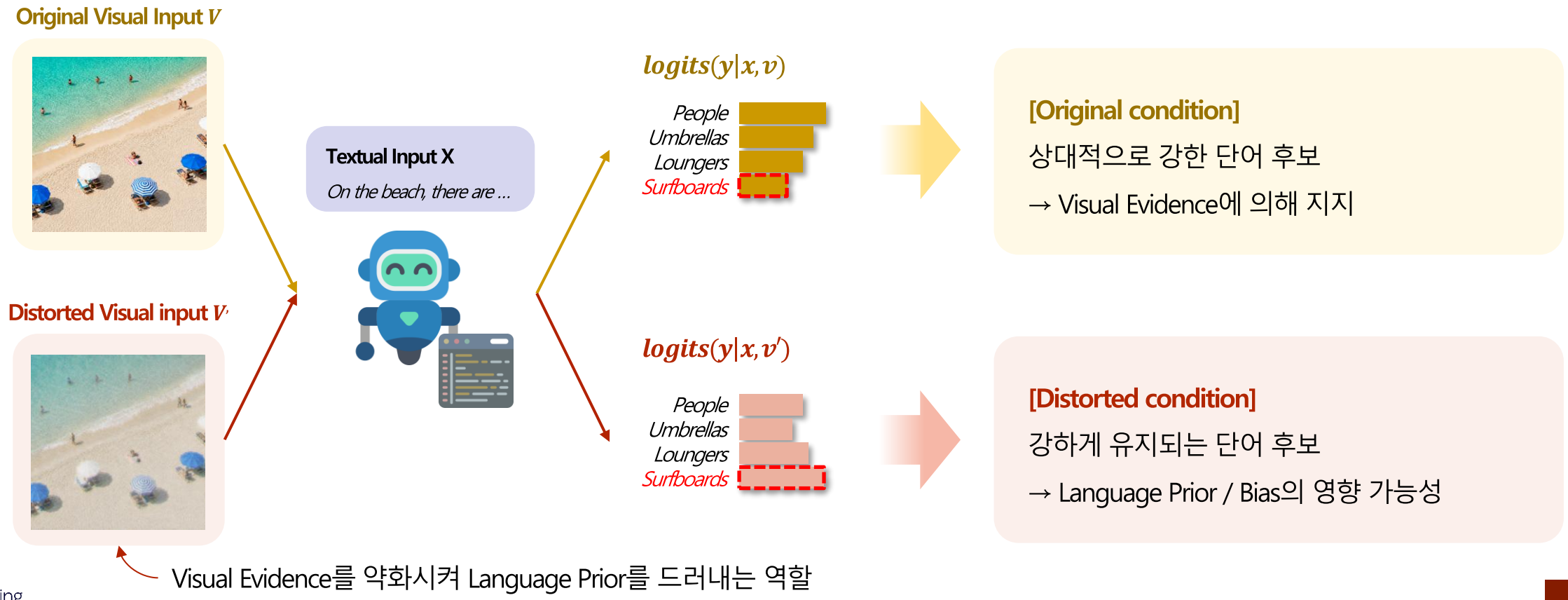
2. VCD

Decoding Intervention



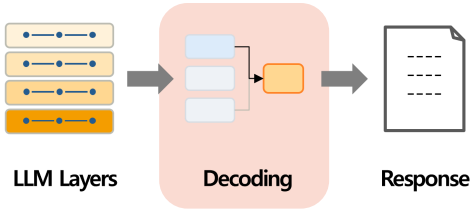
❖ 손상된 이미지에서도 강한 Token은 무엇을 의미할까?

- 원본 이미지 (Original condition) : 시각 정보가 정상적으로 포함된 예측
- 손상된 이미지 (Distorted condition) : 시각 정보는 약해지고, **언어적 편향의 영향이 상대적으로 증가**



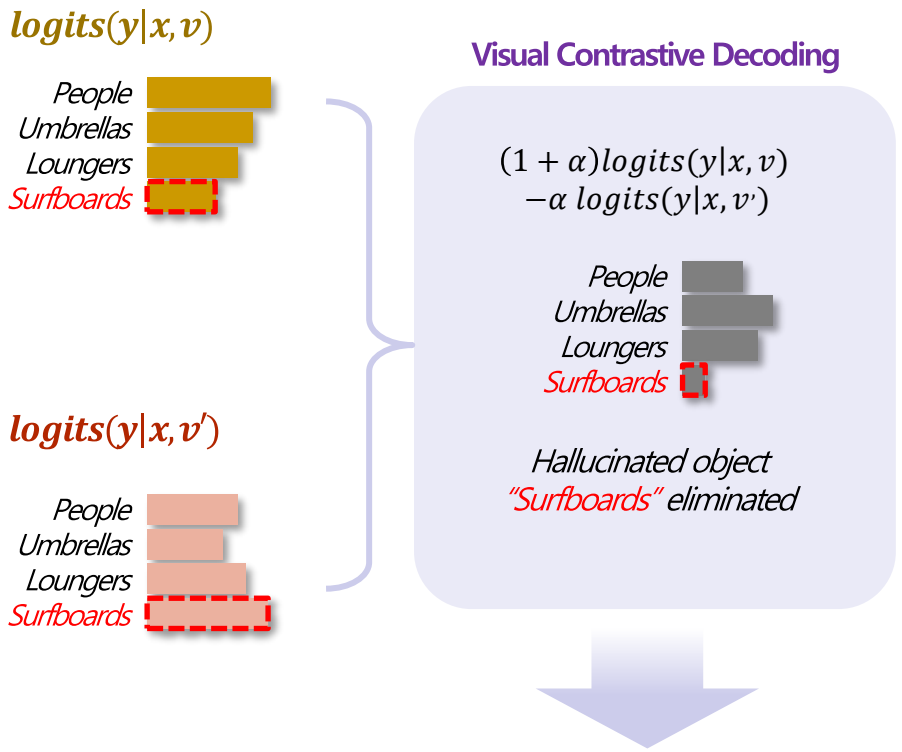
2. VCD

Decoding Intervention



❖ 원본 이미지와 손상된 이미지의 예측을 비교하면 무엇이 달라질까?

- 원본·손상 이미지의 예측 차이를 이용해 시각적 근거는 강화하고, 언어적 편향은 억제



Visual Contrastive Decoding

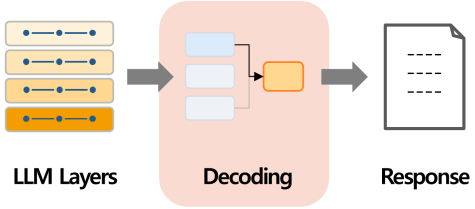
- 손상된 이미지에서도 강한 단어 후보 → **상대적으로 억제**
- 원본 이미지에서 더 강한 단어 후보 → **상대적으로 강화**

시각적 근거(Visual Evidence) ↑ 언어적 편향(Language Prior) ↓

- **Contrast 이전 :** Surfboards probability ↑
- **Contrast 이후 :** Surfboards suppressed / eliminated

2. VCD

Decoding Intervention



❖ Decoding만 바꿔도 Hallucination이 줄어들까?

- Visual Contrastive Decoding을 적용하면 이미지에 근거하지 않은 object generation이 감소

[Regular Decoding]

A beach with people, umbrellas, loungers, and **surfboards**

[VCD]

A beach with people, umbrellas, and loungers

→ Regular Decoding에서 생성된 근거없는 표현이 VCD에서는 제거됨

Prompt: Describe this photo in detail.

Regular Decoding: The image features a painting of a brown dog sitting in a chair, with a turkey feather in its mouth, giving it a unique, conspicuous appearance. The dog's expression is one of mischief as it looks directly into the camera. The scene is set in a room with a dining table and a fork next to the table. A rock, likely from the surrounding landscape, can be seen beside the table. A person's face is visible near the bottom left corner of the painting, partially covered by the dog's furry coat.

Visual Contrastive Decoding: The image features a unique and amusing painting of a dog with a large brush mustache, giving it an artistic human-like appearance. The dog is sitting on a couch, with a red background behind it, adding to the artistic flair of the painting. The couch has an orange and green tint to it. In addition to the dog's human-like mustache, the painting also includes a monochromatic style with the background, giving the painting a striking visual appeal. The dog appears to be wearing a hat, further emphasizing the artistic interpretation of the scene. Overall, the painting captures the dog's whimsical and charming personality.

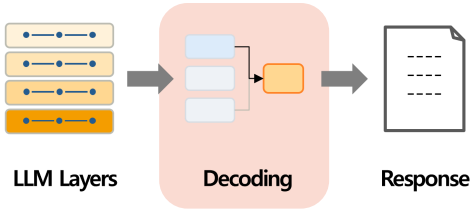
Prompt: Describe this photo in detail.

Regular Decoding: In the image, a man is standing on the back of a Subaru in a city with the color of a Ford SUV parked in the background. He appears to be ironing clothes while standing on a wine rack located under the vehicle. He is wearing a yellow shirt. There is a white and blue taxi parked nearby the yellow Subaru. Additionally, there are several buildings located in the cityscape, perhaps in the background of the Subaru and the taxi, and possibly overlapping with the yellow Subaru.

Visual Contrastive Decoding: In the photo, a man is ironing clothes on a table placed on the back of a taxi in a busy city street. The man is wearing a yellow shirt, suggesting he is a taxi driver or a part of the yellow cabs' team. The yellow vehicle in the background is an SUV-like vehicle, which is not a typical yellow cab. The table on the back of the car is set up with a pile of clothes, which the man is ironing. The man is working hard to get the clothes as wrinkle-free as possible.

2. VCD

Decoding Intervention



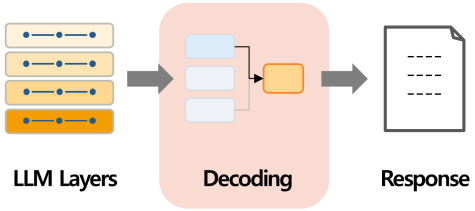
❖ Decoding만 바꿔도 Hallucination이 줄어들까?

- Key Results
  - ➢ 여러 LVLM에서 Regular Decoding 대비 성능 개선
  - ➢ Object Hallucination 감소
  - ➢ 추가 학습 없이 inference 단계에서 적용

Dataset	Setting	Model	Decoding	Accuracy↑	Precision	Recall	F1 Score↑
MSCOCO	*Random*	LLaVA1.5	Regular	$83.29_{(\pm 0.35)}$	$92.13_{(\pm 0.54)}$	$72.80_{(\pm 0.57)}$	$81.33_{(\pm 0.41)}$
			VCD	$\mathbf{87.73}_{(\pm 0.40)}$	$91.42_{(\pm 0.55)}$	$83.28_{(\pm 0.42)}$	$\mathbf{87.16}_{(\pm 0.41)}$
		Qwen-VL	Regular	$84.73_{(\pm 0.36)}$	$95.61_{(\pm 0.45)}$	$72.81_{(\pm 0.38)}$	$82.67_{(\pm 0.41)}$
			VCD	$\mathbf{88.63}_{(\pm 0.10)}$	$94.64_{(\pm 0.25)}$	$81.91_{(\pm 0.19)}$	$\mathbf{87.81}_{(\pm 0.11)}$
		InstructBLIP	Regular	$80.71_{(\pm 0.73)}$	$81.67_{(\pm 0.67)}$	$79.19_{(\pm 1.14)}$	$80.41_{(\pm 0.80)}$
			VCD	$\mathbf{84.53}_{(\pm 0.38)}$	$88.55_{(\pm 0.54)}$	$79.32_{(\pm 0.44)}$	$\mathbf{83.68}_{(\pm 0.40)}$
	Popular	LLaVA1.5	Regular	$81.88_{(\pm 0.48)}$	$88.93_{(\pm 0.60)}$	$72.80_{(\pm 0.57)}$	$80.06_{(\pm 0.05)}$
			VCD	$\mathbf{85.38}_{(\pm 0.38)}$	$86.92_{(\pm 0.53)}$	$83.28_{(\pm 0.42)}$	$\mathbf{85.06}_{(\pm 0.37)}$
		Qwen-VL	Regular	$84.13_{(\pm 0.18)}$	$94.31_{(\pm 0.43)}$	$72.64_{(\pm 0.45)}$	$82.06_{(\pm 0.23)}$
			VCD	$\mathbf{87.12}_{(\pm 0.07)}$	$91.49_{(\pm 0.10)}$	$81.85_{(\pm 0.19)}$	$\mathbf{86.40}_{(\pm 0.09)}$
		InstructBLIP	Regular	$78.22_{(\pm 0.84)}$	$77.87_{(\pm 1.03)}$	$78.85_{(\pm 0.52)}$	$78.36_{(\pm 0.76)}$
			VCD	$\mathbf{81.47}_{(\pm 0.42)}$	$82.89_{(\pm 0.64)}$	$79.32_{(\pm 0.44)}$	$\mathbf{81.07}_{(\pm 0.39)}$
	Adversarial	LLaVA1.5	Regular	$78.96_{(\pm 0.52)}$	$83.06_{(\pm 0.58)}$	$72.75_{(\pm 0.59)}$	$77.57_{(\pm 0.57)}$
			VCD	$\mathbf{80.88}_{(\pm 0.33)}$	$79.45_{(\pm 0.29)}$	$83.29_{(\pm 0.43)}$	$\mathbf{81.33}_{(\pm 0.34)}$
		Qwen-VL	Regular	$82.26_{(\pm 0.30)}$	$89.97_{(\pm 0.33)}$	$72.61_{(\pm 0.50)}$	$80.37_{(\pm 0.37)}$
			VCD	$\mathbf{84.26}_{(\pm 0.39)}$	$85.84_{(\pm 0.45)}$	$82.05_{(\pm 0.39)}$	$\mathbf{83.90}_{(\pm 0.39)}$
		InstructBLIP	Regular	$75.84_{(\pm 0.45)}$	$74.30_{(\pm 0.63)}$	$79.03_{(\pm 0.68)}$	$76.59_{(\pm 0.40)}$
			VCD	$\mathbf{79.56}_{(\pm 0.41)}$	$79.67_{(\pm 0.59)}$	$79.39_{(\pm 0.50)}$	$\mathbf{79.52}_{(\pm 0.38)}$

2. VCD

Decoding Intervention



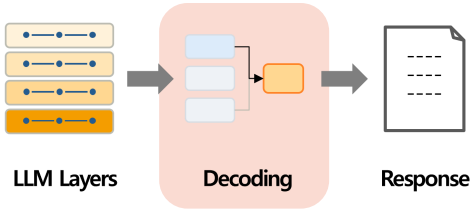
❖ Decoding만 바꿔도 Hallucination이 줄어들까?

- Key Results
  - ➢ 여러 LVLM에서 Regular Decoding 대비 성능 개선
  - ➢ Object Hallucination 감소
  - ➢ 추가 학습 없이 inference 단계에서 적용

- Limitation
  - ➢ Original / Distorted 두 input을 사용해 추론 비용 증가
  - ➢ Hallucination의 내부 발생 과정 자체를 직접 분석하지는 않음

2. VCD

Decoding Intervention



❖ **Decoding만 바꿔도 Hallucination이 줄어들까?**

- Key Results
  - ➢ 여러 LVLM에서 Regular Decoding 대비 성능 개선
  - ➢ Object Hallucination 감소
  - ➢ 추가 학습 없이 inference 단계에서 적용

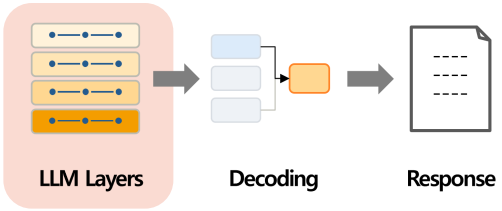
- Limitation
  - ➢ Original / Distorted 두 input을 사용해 추론 비용 증가
  - ➢ Hallucination의 내부 발생 과정 자체를 직접 설명하지는 않음

그렇다면 Hallucination이 출력되기 전, LLM 내부에서 그 징후를 찾을 수 없을까?

→ Devils in Middle Layers (CVPR 2025)

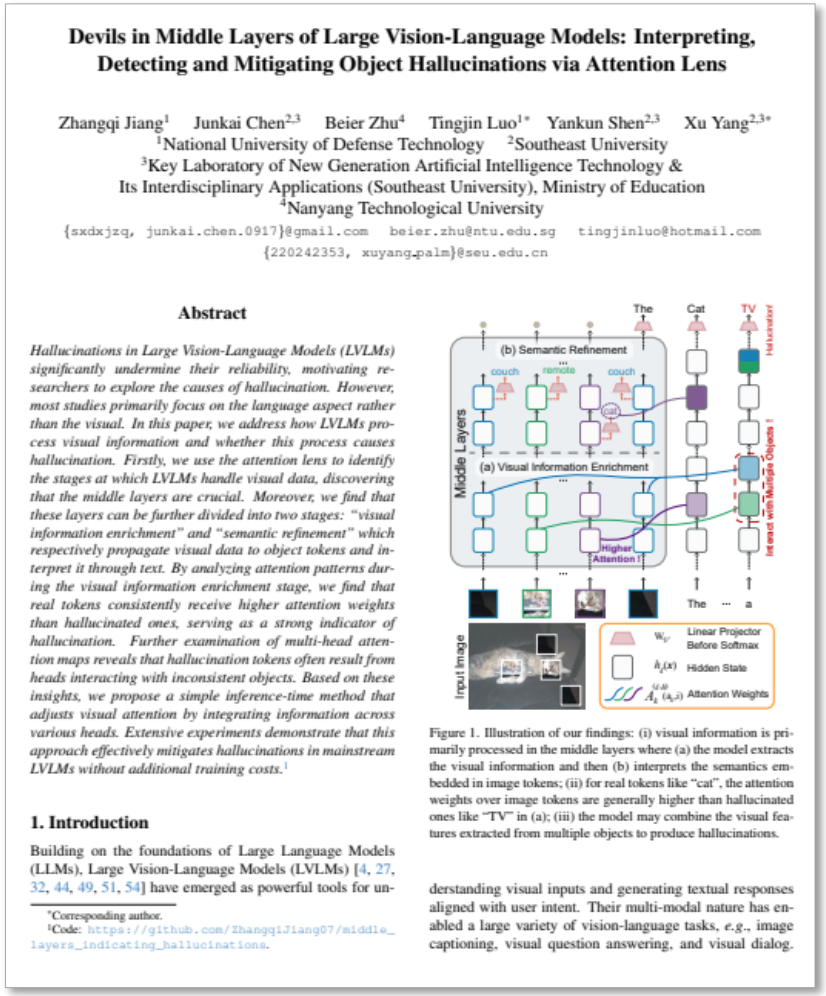
3. Devils in Middle Layers

Internal Mechanism



❖ **Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens (CVPR 2025)**

- 기존 연구는 주로 최종 출력(RLHF-V)이나 decoding 과정(VCD)에서 Hallucination을 완화
- 본 논문은 LLM Layer별 Visual Attention을 분석해 Hallucination의 내부 메커니즘을 해석



Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens

Zhangqi Jiang[1] Junkai Chen[2,3] Beier Zhu[4] Tingjin Luo[1*] Yankun Shen[2,3] Xu Yang[2,3*]
[1]National University of Defense Technology [2]Southeast University
[3]Key Laboratory of New Generation Artificial Intelligence Technology & Its Interdisciplinary Applications (Southeast University), Ministry of Education
[4]Nanyang Technological University
{sxdxjzq, junkai.chen.0917}@gmail.com beier.zhu@ntu.edu.sg tingjinluo@hotmail.com
{220242353, xuyang_palm}@seu.edu.cn

Abstract

Hallucinations in Large Vision-Language Models (LVLMs) significantly undermine their reliability, motivating researchers to explore the causes of hallucination. However, most studies primarily focus on the language aspect rather than the visual. In this paper, we address how LVLMs process visual information and whether this process causes hallucination. Firstly, we use the attention lens to identify the stages at which LVLMs handle visual data, discovering that the middle layers are crucial. Moreover, we find that these layers can be further divided into two stages: "visual information enrichment" and "semantic refinement" which respectively propagate visual data to object tokens and interpret it through text. By analyzing attention patterns during the visual information enrichment stage, we find that real tokens consistently receive higher attention weights than hallucinated ones, serving as a strong indicator of hallucination. Further examination of multi-head attention maps reveals that hallucination tokens often result from heads interacting with inconsistent objects. Based on these insights, we propose a simple inference-time method that adjusts visual attention by integrating information across various heads. Extensive experiments demonstrate that this approach effectively mitigates hallucinations in mainstream LVLMs without additional training costs.[1]

1. Introduction

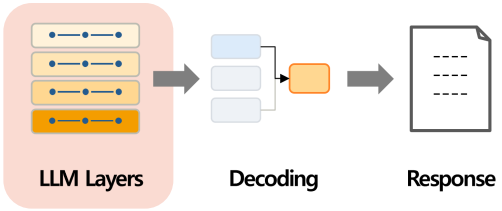
Building on the foundations of Large Language Models (LLMs), Large Vision-Language Models (LVLMs) [4, 27, 32, 44, 49, 51, 54] have emerged as powerful tools for understanding visual inputs and generating textual responses aligned with user intent. Their multi-modal nature has enabled a large variety of vision-language tasks, *e.g.*, image captioning, visual question answering, and visual dialog.

*Corresponding author.
[1]Code: https://github.com/ZhangqiJiang07/middle_layers_indicating_hallucinations.

Figure 1. Illustration of our findings: (i) visual information is primarily processed in the middle layers where (a) the model extracts the visual information and then (b) interprets the semantics embedded in image tokens; (ii) for real tokens like "cat", the attention weights over image tokens are generally higher than hallucinated ones like "TV" in (a); (iii) the model may combine the visual features extracted from multiple objects to produce hallucinations.

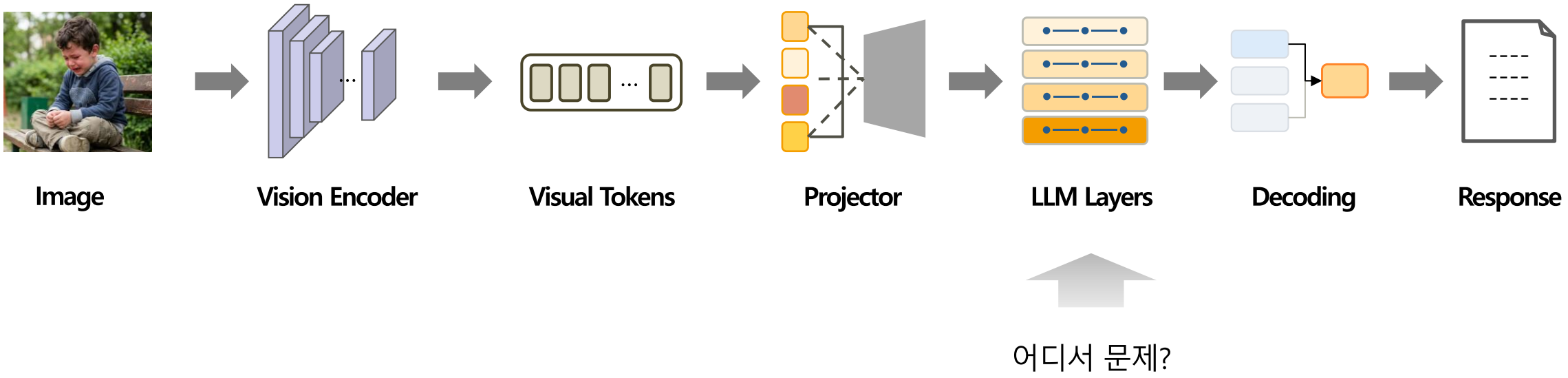
3. Devils in Middle Layers

Internal Mechanism



❖ **Hallucination은 모델 내부의 어느 단계에서 나타날까?**

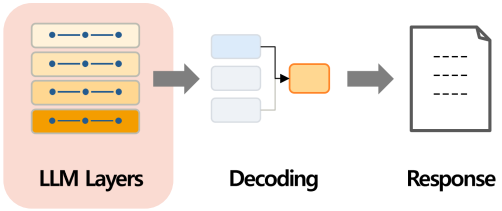
- 기존 연구는 주로 **최종 출력이나 decoding 과정**에서 Hallucination을 완화
- 하지만 MLLM 내부에서 **시각 정보가 어떤 layer에서 처리되고, 어디서 오류가 발생하는지는 명확하지 않음**



최종 출력이 아니라, LLM 내부의 시각 정보 처리 과정에 주목해보자!

3. Devils in Middle Layers

Internal Mechanism



❖ 모든 LLM Layer가 시각 정보를 동일하게 처리할까?

- Attention을 이용해 **layer별 시각 정보 처리 과정**을 분석
- 시각 정보의 핵심 처리가 **중간 layer**에 집중되는 것을 확인

< Middle Layers >

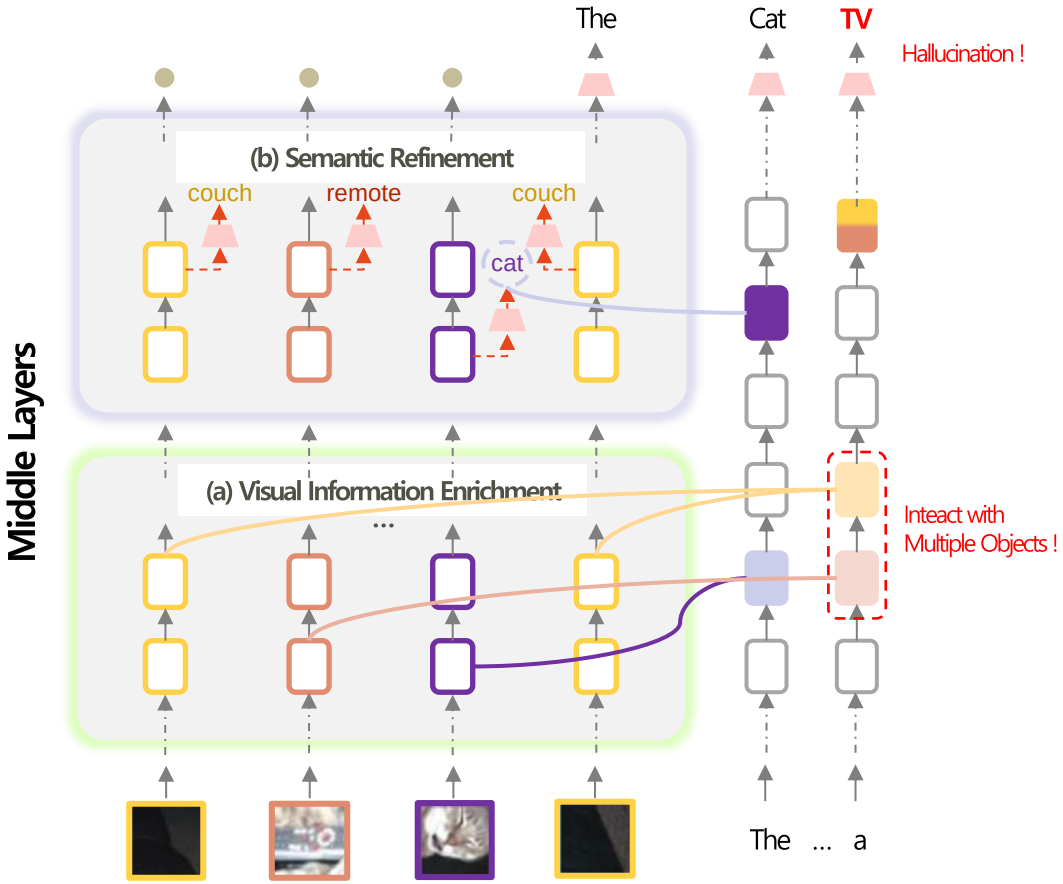
① 시각 정보 강화

→ 이미지 정보를 객체 표현에 반영

② 의미 정제

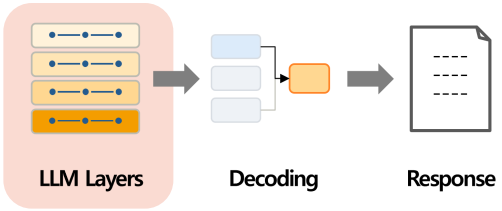
→ 시각 정보를 다음 token 예측에 활용

→ Middle Layers가 시각 정보 처리의 핵심 구간



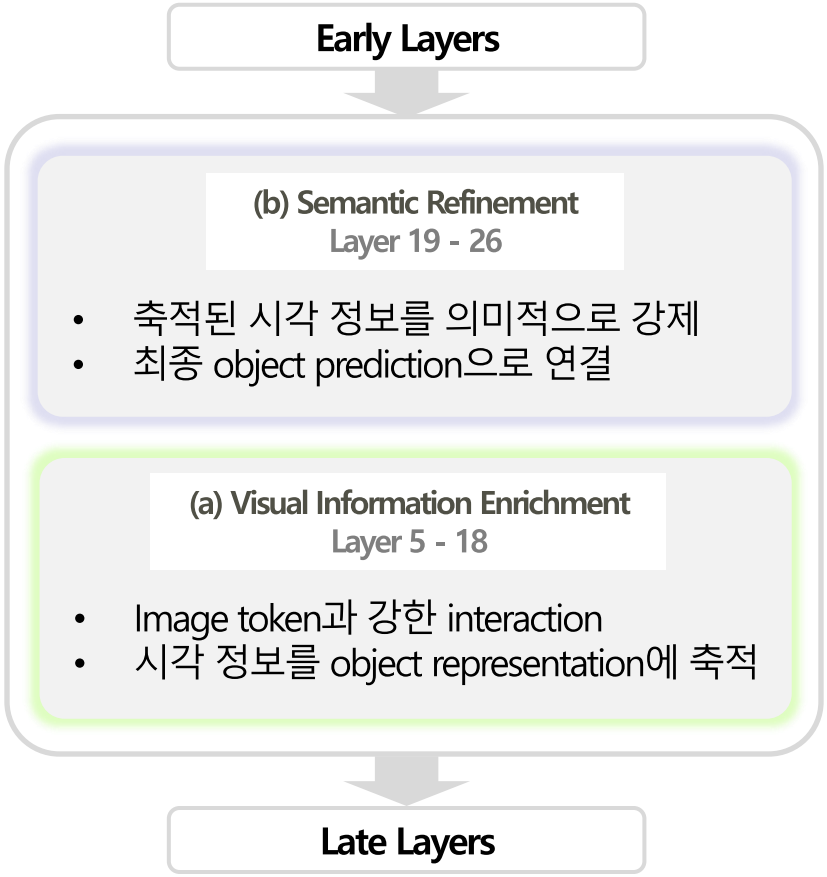
3. Devils in Middle Layers

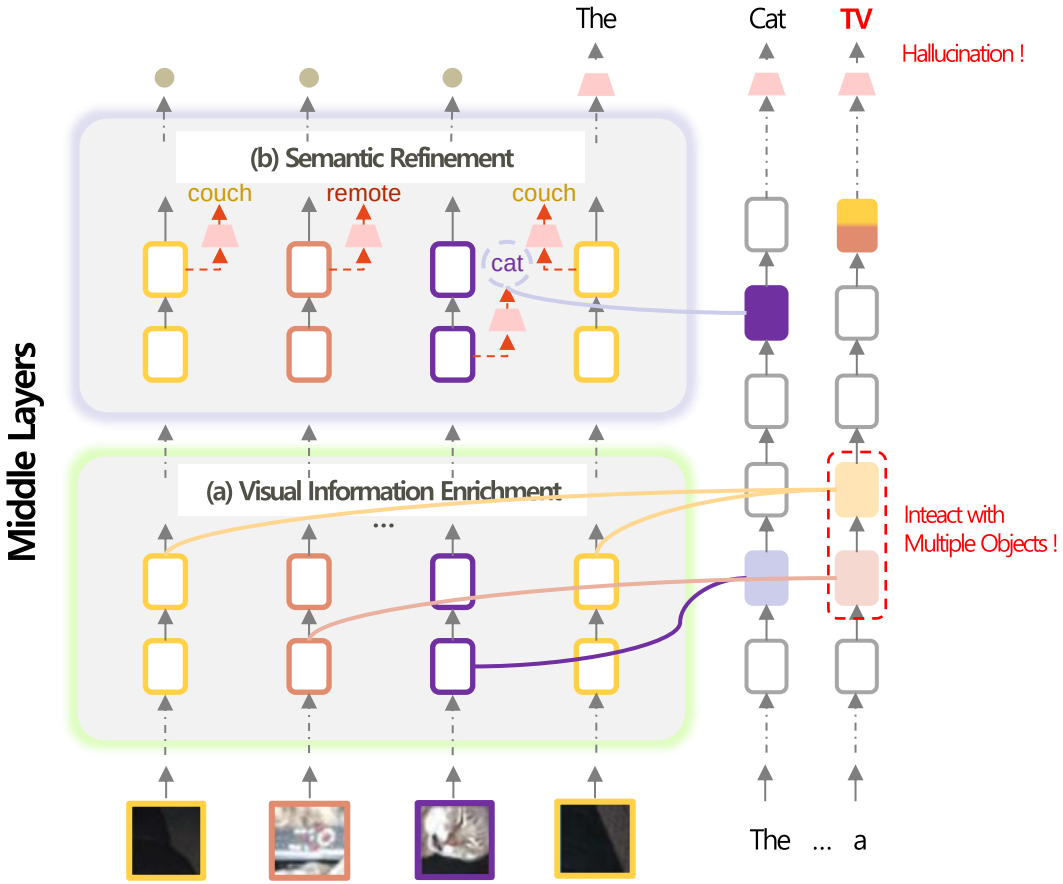
Internal Mechanism



❖ Middle Layers에서는 시각 정보가 어떻게 처리될까?

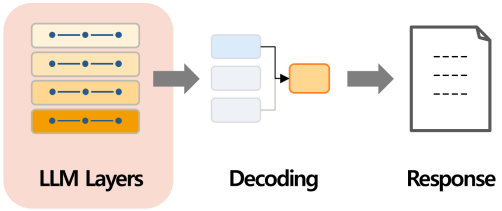
- Middle Layers의 시각 정보 처리는 두 단계로 나타남





3. Devils in Middle Layers

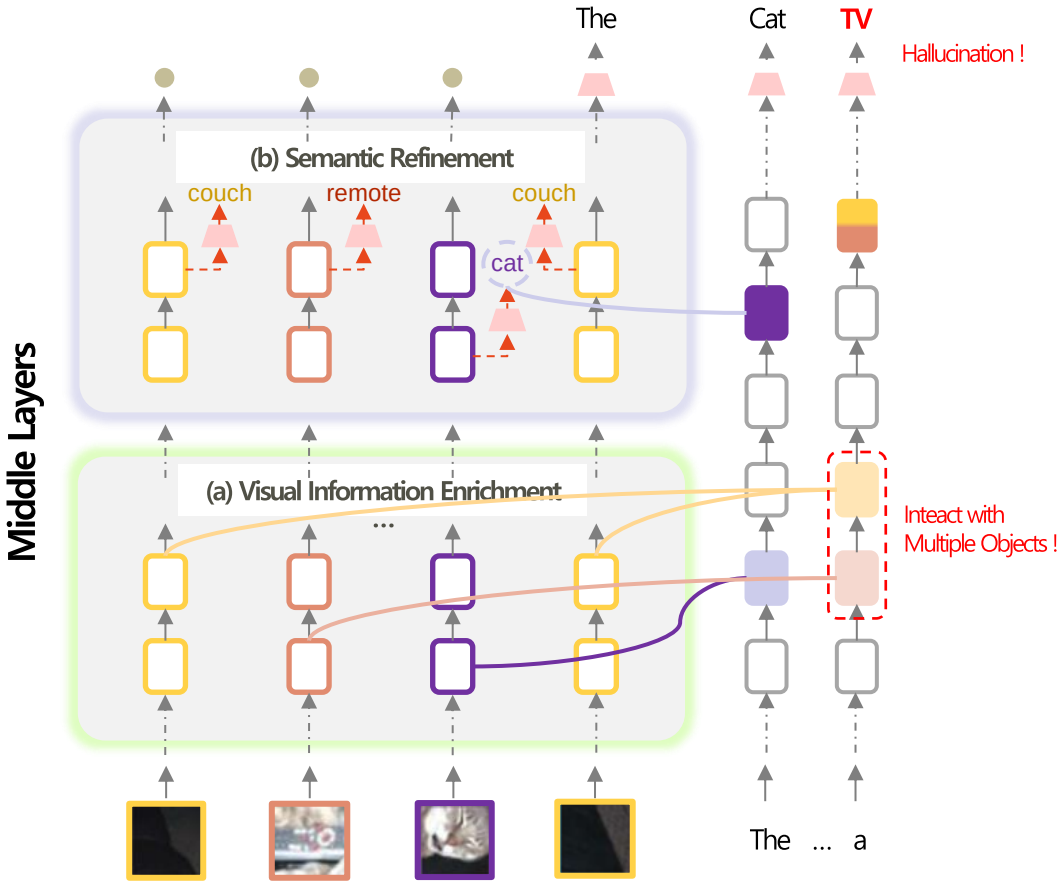
Internal Mechanism



❖ Middle Layers에서는 시각 정보가 어떻게 처리될까?

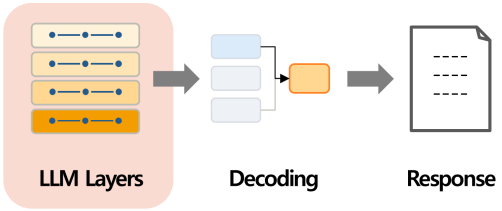
- Middle Layers의 시각 정보 처리는 두 단계로 나타남

시각 정보는 중간 layer에서 먼저 축적되고, 이후 의미적으로 정제



3. Devils in Middle Layers

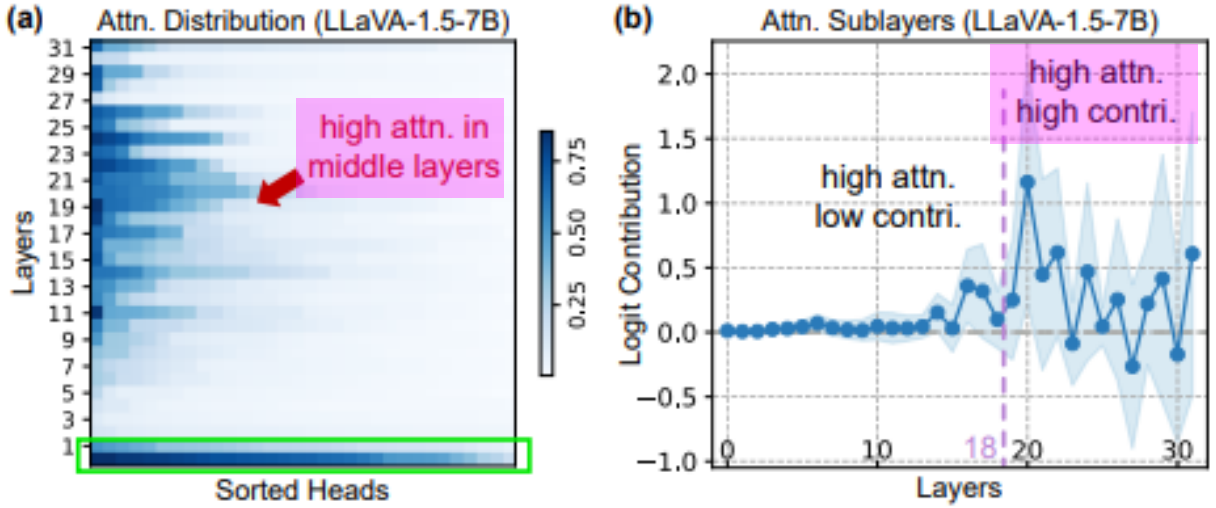
Internal Mechanism



❖ 실제로 Middle Layers에서 시각 정보가 집중될까?

- Layer별 visual attention과 logit contribution을 분석

< Visual attention is concentrated in middle layers >

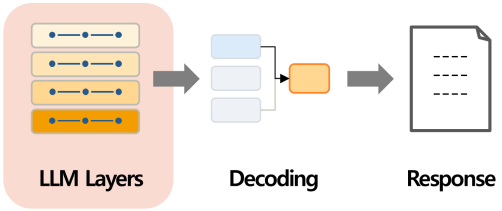


- **Early Layers :** visual interaction 상대적으로 약함
- **Middle Layers :** visual attention과 contribution이 크게 증가
- **Late Layers :** 최종 prediction으로 전달

Middle Layers가 visual information processing의 핵심 구간으로 관찰됨

3. Devils in Middle Layers

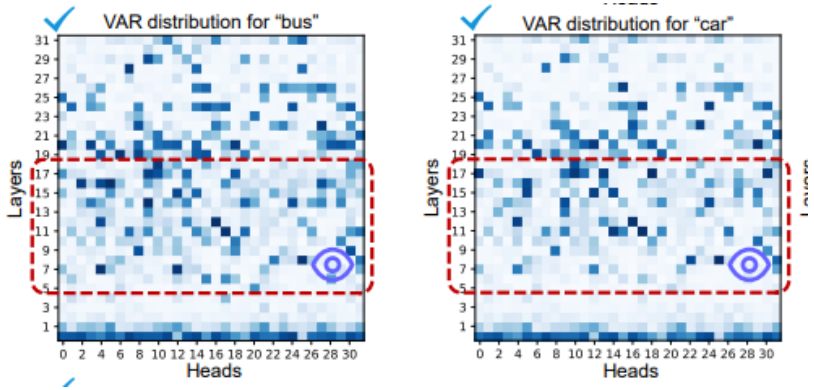
Internal Mechanism



❖ 실제 객체와 Hallucinated 객체는 Attention이 어떻게 다를까?

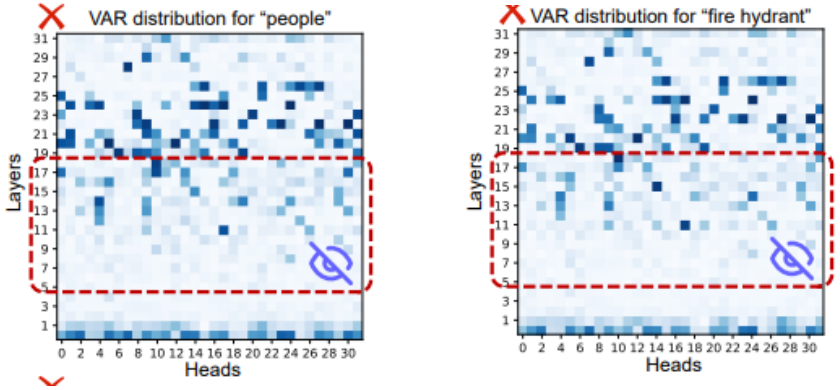
- Layer별 visual attention과 logit contribution을 분석

< Real Object Tokens >



Active visual attention

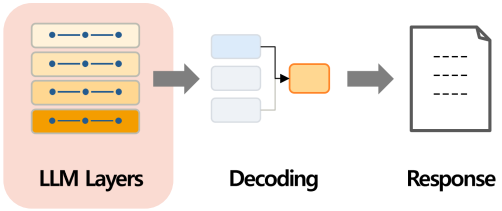
< Hallucinated Object Tokens >



Inactive visual attention

3. Devils in Middle Layers

Internal Mechanism

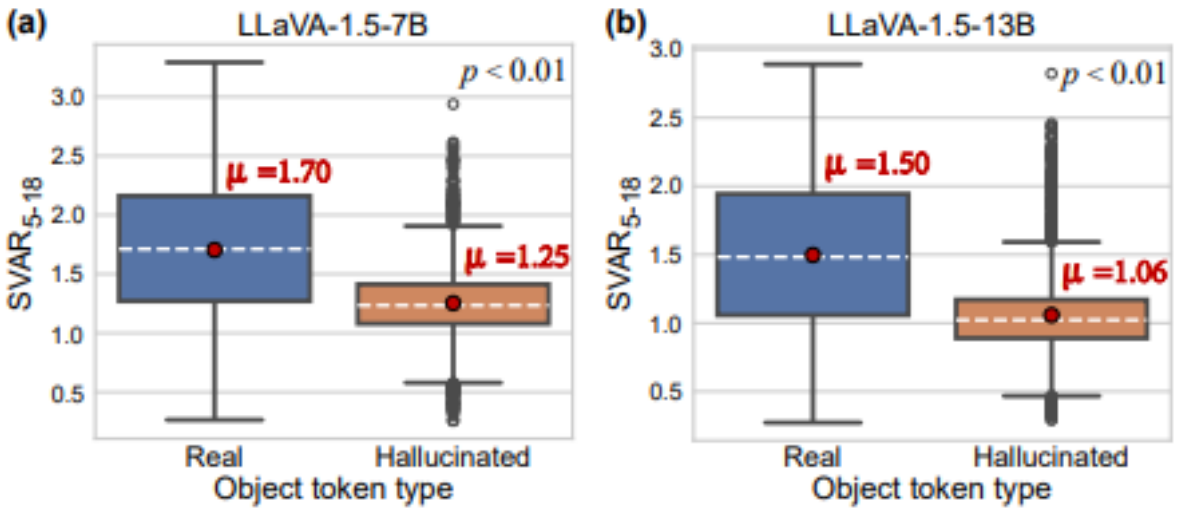


❖ 실제 객체와 Hallucinated 객체는 Attention이 어떻게 다를까?

- Layer별 visual attention과 logit contribution을 분석

<Visual Attention Ratio>

생성 token이 image token을 얼마나 참고하는지 나타내는 지표



Real token :

enrichment stage에서 visual attention이 높음

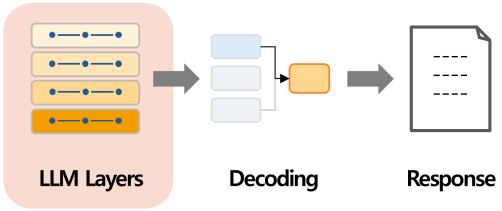
Hallucinated token :

같은 구간에서 visual attention이 상대적으로 낮음

Hallucinated token은 Visual Information Enrichment 단계에서 낮은 visual attention을 보임

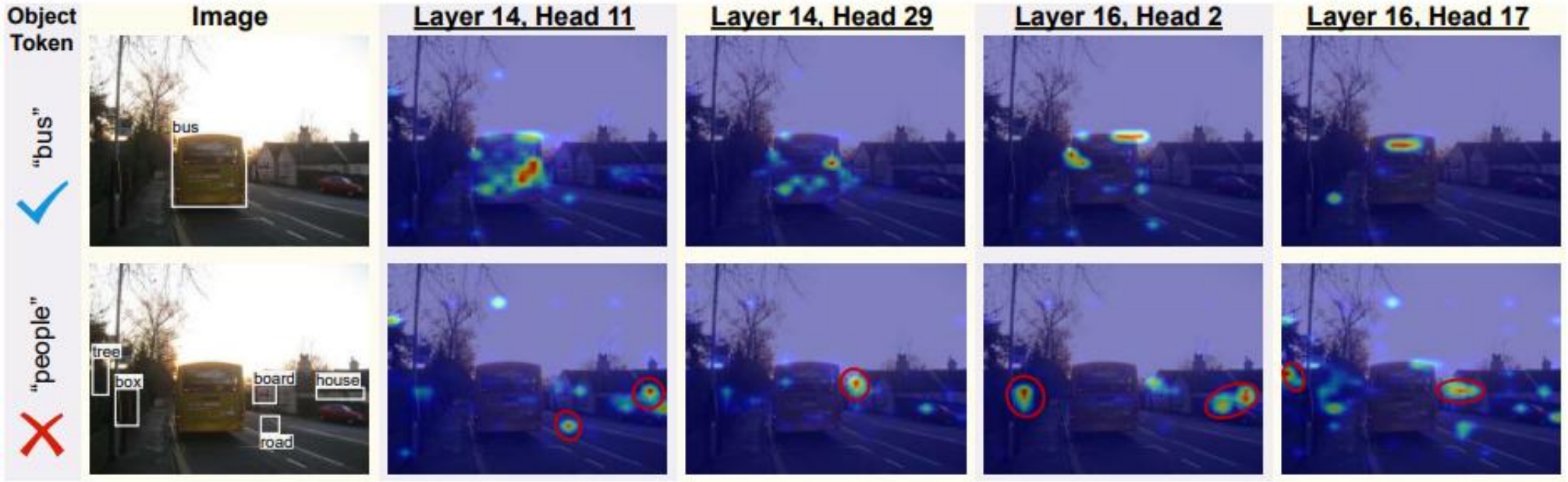
3. Devils in Middle Layers

Internal Mechanism



❖ **그런데 attention이 단순히 약한 것만이 문제일까?**

- Real token : 여러 head가 실제 객체 영역을 비교적 일관되게 참조
- Hallucinated token : head마다 서로 다른 객체·영역을 참조



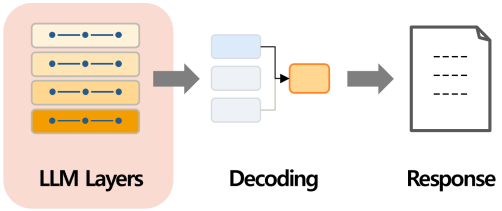
Real Object

Hallucinated Object

→ Hallucinated token은 낮은 visual attention뿐 아니라, Head 간 불일치한 visual interaction을 보임

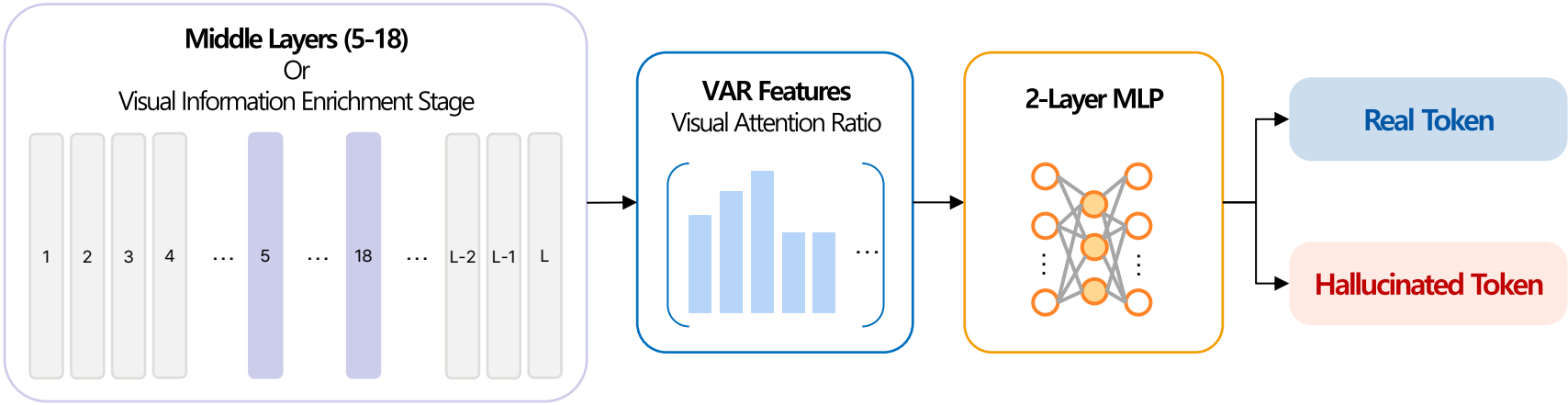
3. Devils in Middle Layers

Internal Mechanism



❖ Middle-layer Attention만으로 Hallucination을 구분할 수 있을까?

- Middle Layers의 Visual Attention Ratio를 Hallucination signal로 활용
- 2-layer MLP로 Real / Hallucinated token 분류



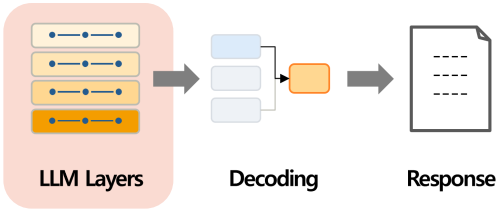
<Detection Performance by Layer Range>

Layers	0-4	5-18	19-26	27-31	0-31
Accuracy	82.61	**84.46**	81.49	77.72	84.46
AUROC	89.57	**90.19**	87.34	84.30	90.28
Recall	66.22	**72.34**	68.88	63.56	71.28

Best single-stage

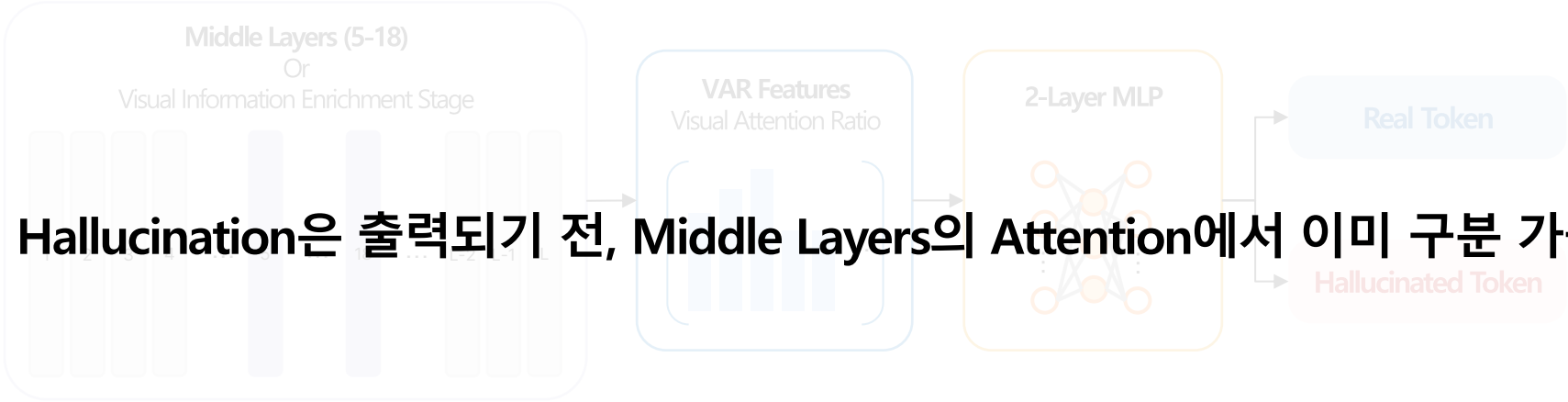
3. Devils in Middle Layers

Internal Mechanism



❖ **Middle-layer Attention만으로 Hallucination을 구분할 수 있을까?**

- Middle Layers의 Visual Attention Ratio를 Hallucination signal로 활용
- 2-layer MLP로 Real / Hallucinated token 분류



Hallucination은 출력되기 전, Middle Layers의 Attention에서 이미 구분 가능한 신호로 나타남

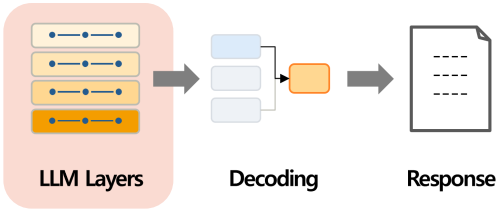
<Detection Performance by Layer Range>

Layers	0-4	5-18	19-26	27-31	0-31
Accuracy	82.61	84.46	81.49	77.72	84.46
AUROC	89.57	90.19	87.34	84.30	90.28
Recall	66.22	72.34	68.88	63.56	71.28

Best single-stage

3. Devils in Middle Layers

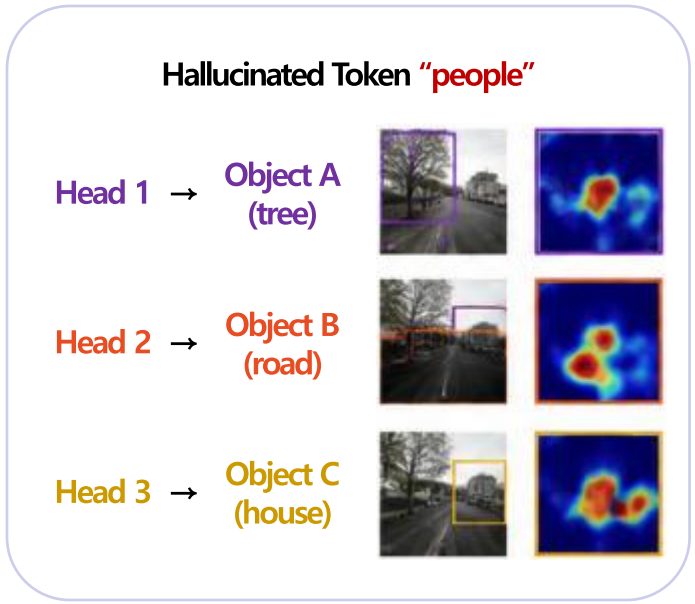
Internal Mechanism



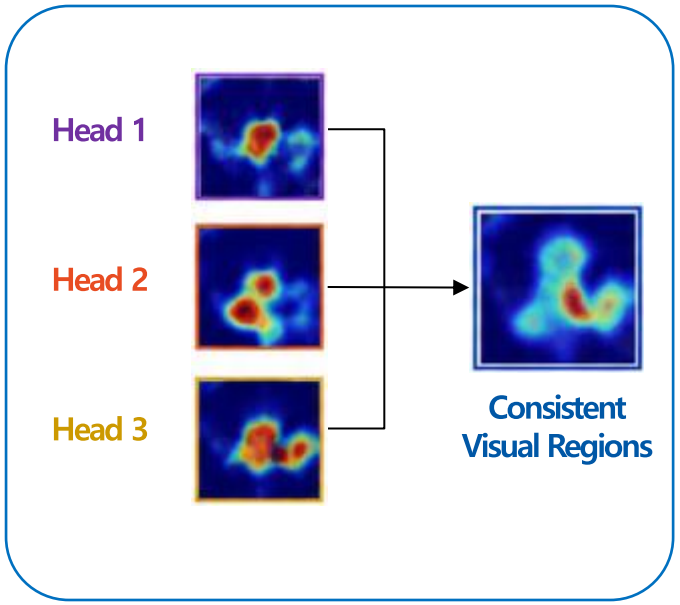
❖ Hallucination을 유발하는 Attention을 직접 보정한다면?

- 여러 Head에서 공통적으로 강조되는 image region을 통합
- Enrichment stage의 visual attention을 공통 영역 방향으로 강화

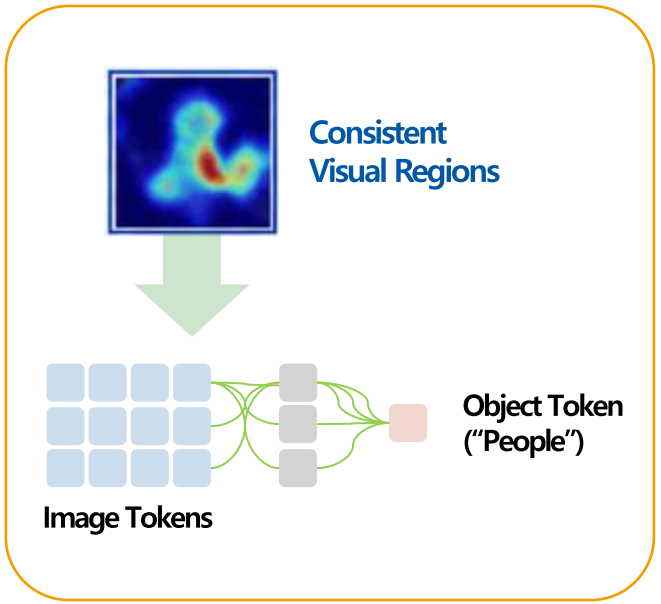
① Inconsistent Head Interaction



② Across-head Aggregation



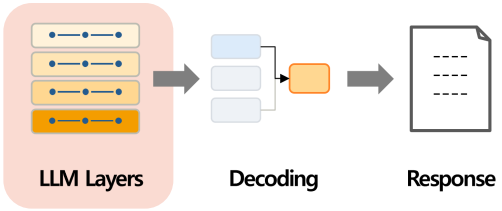
③ Attention Intervention



공통 visual evidence를 강화해 더 일관된 object prediction을 유도

3. Devils in Middle Layers

Internal Mechanism



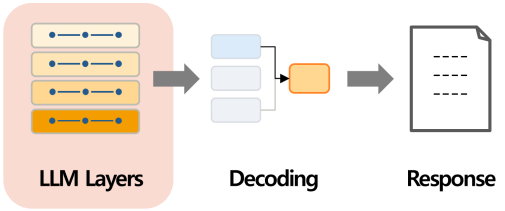
❖ Attention을 직접 보정하면 Hallucination도 줄어들까?

- Key Results
  - ➢ Middle-layer Attention 보정만으로 Hallucination 지표 감소
  - ➢ 여러 LVLM에서 공통적으로 개선 효과 확인
  - ➢ 재학습 없이 inference 단계에서 바로 적용 가능

Method	LLaVA-1.5-7B			LLaVA-1.5-13B			Shikra-7B			MiniGPT-4-7B			Avg.	
	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$
Decoding Strategy														
Greedy	53.0	15.6	76.7	49.8	14.6	78.2	57.6	15.7	75.3	31.8	12.0	71.1	48.1	14.5
Beam	55.6	15.4	77.5	50.4	13.8	79.0	59.0	16.3	74.6	29.2	9.9	71.2	48.6	13.9
OPERA	45.6	13.1	79.1	42.6	13.2	77.8	41.4	13.7	73.5	25.4	9.6	71.2	38.8	12.4
Contrastive Decoding														
VCD†	58.6	18.2	72.8	53.6	15.3	75.8	56.4	15.5	75.2	41.4	14.1	68.2	52.5	15.8
PAI†	**24.2**	7.1	75.2	33.0	9.2	77.8	38.6	10.1	76.2	23.2	8.2	71.4	29.8	8.7
Ours†	25.0	**6.7**	76.1	**25.8**	**8.8**	77.3	**23.8**	**9.4**	72.7	**21.4**	**8.0**	70.8	**24.0**	**8.2**
Δ%	↑3.3%	↓5.6%		↓21.8%	↓4.3%		↓38.3%	↓6.9%		↓7.8%	↓2.4%		↓19.5%	↓5.7%

3. Devils in Middle Layers

Internal Mechanism



❖ Attention을 직접 보정하면 Hallucination도 줄어들까?

- Key Results
  - ➢ Middle-layer Attention 보정만으로 Hallucination 지표 감소
  - ➢ 여러 LVLM에서 공통적으로 개선 효과 확인
  - ➢ 재학습 없이 inference 단계에서 바로 적용 가능

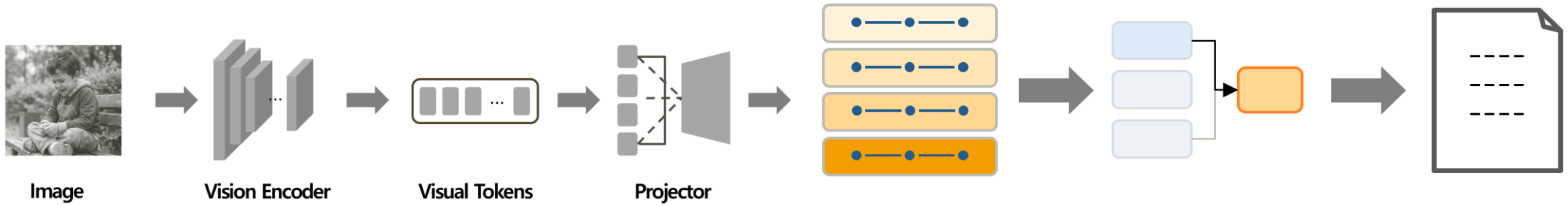
- Limitation
  - ➢ 모델별로 visual-processing layer 범위가 달라질 수 있음
  - ➢ Object Hallucination 중심 분석

Middle Layers의 Visual Attention은 Hallucination을 해석·탐지할 뿐 아니라
직접 완화하는 signal로 활용 가능

Conclusion

이거 뭐라하지?

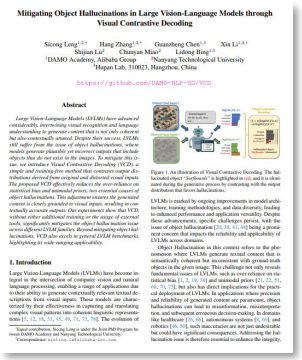
❖ 세 논문은 Hallucination을 어디에서 바라봤을까?



③ Devils in Middle Layers (CVPR 2025)

Why hallucinate?

② VCD (CVPR 2024)



How to generate?

① RLHF-V (CVPR 2024)

What to say?

Conclusion

이거 뭐라하지?

❖ **Hallucination을 줄이기 위해 결국 중요한 것은 무엇일까?**

- **Output만 고치는 것으로 끝나지 않는다**

  → 어떤 내용을 생성하도록 학습하는지가 중요 (RLHF-V)

- **생성 순간에도 Visual Evidence를 지켜야 한다**

  → Language Prior가 시각 정보보다 우세해지지 않도록 제어 (VCD)

- **내부에서 Visual Evidence가 실제로 활용되는지가 중요하다**

  → Hallucination signal은 최종 출력 이전에도 나타남 (Devils in Middle Layers)

Hallucination은 최종 응답만의 문제가 아니라,

Visual Evidence가 전달되고 활용되어 응답으로 이어지는 전 과정과 관련된다.

고맙습니다

Appendix

❖ Hallucination을 어떻게 발견하고 측정할까? : Benchmark

- Output-level benchmark뿐 아니라, Devils에서는 Attention-level signal까지 Hallucination 평가에 활용

평가	이번 발표에서의 역할	핵심
Object HalBench	RLHF-V	생성 응답의 Object Hallucination 평가
POPE	VCD	객체 존재 여부 기반 Hallucination 평가
CHAIR	VCD / Devils 관련 결과 해석	생성 caption 내 hallucinated object 평가
VAR / SVAR	Devils	Visual Attention 기반 내부 Hallucination signal