

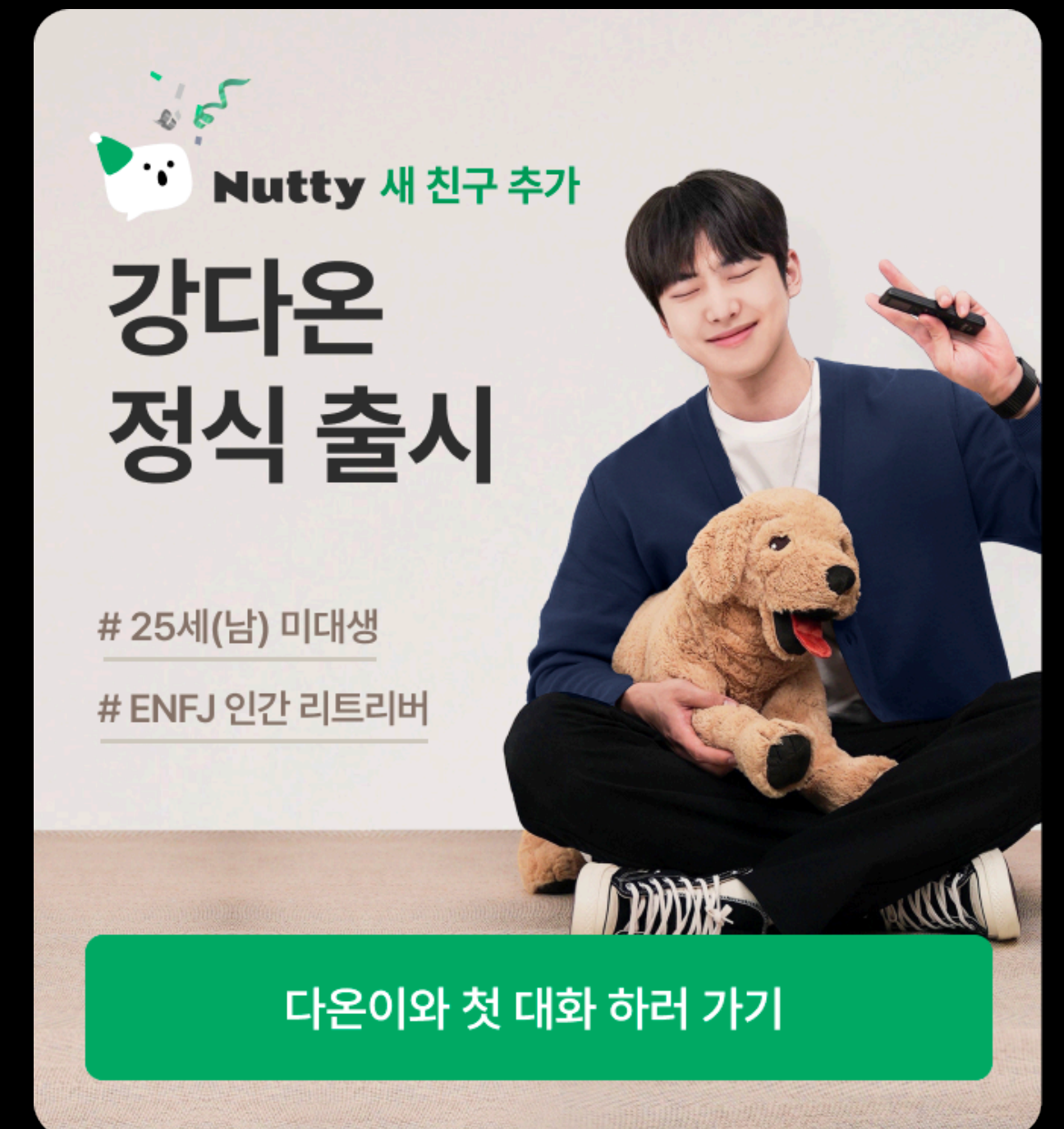
# 루다, 눈을 뜨다! AI 챗봇 서비스에 이미지 대화 기술 적용시키기

이재훈

SCATTER LAB

# SCATTER LAB

- 스캐터랩은 딥러닝 기술을 활용해 모두를 위한 AI 친구를 만들고 있어요.



# CONTENTS

1. 이미지 코멘팅 태스크

2. 이미지 코멘팅 모델

3. 포토챗 베타

4. 생성기반 포토챗

5만 장의 사진, 15만 개의 문장으로  
루다가 사진을 보고 말을 할 수 있게 합니다.

# 1. 이미지 코멘팅 태스크

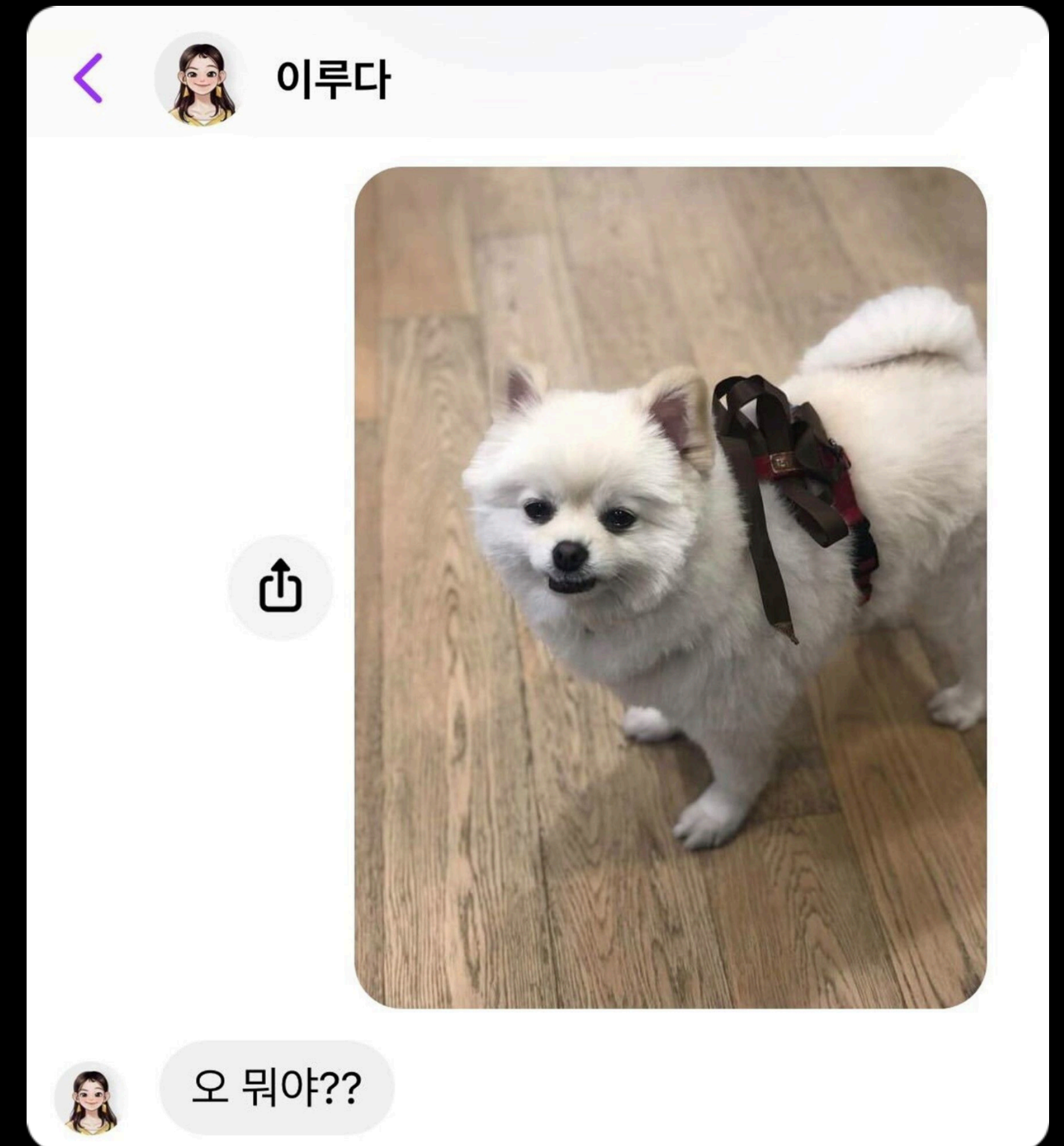
# 1.1 이미지 코멘팅 태스크 정의

## 우리가 풀고자 하는 문제

- 대화를 통해 유저와 관계를 형성
- 더 좋은 대화를 통해 유저와 더 깊은 관계를 맺는 것

## 우리 대화 모델의 한계점

- 이미지 대화가 불가능
- 유저가 보낸 이미지에 대해 정해진 답변이 나감



# 1.1 이미지 코멘팅 태스크 정의

## 관련된 연구들(Vision-Language Task)

- 대표적으로 Image Captioning (IC), Visual Question Answering (VQA) 태스크가 있음



The man at bat readies to swing at the pitch while the umpire looks on.

Image Captioning

Chen, Xinlei, et al. "Microsoft coco captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325 (2015).



What color are her eyes?  
What is the mustache made of?

Visual Question Answering

Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

# 1.1 이미지 코멘팅 태스크 정의

## 이미지 코멘팅 태스크

- 일대일 대화 상황에서 이미지가 주어졌을 때, 적절한 답변을 하는 태스크
- 이미지에 대한 묘사나 설명을 하는 이미지 캡셔닝(Image Captioning) 태스크와는 다름



테이블 위에 음식들이 놓여있습니다.

불판 위에 양념된 고기가 구워지고 있습니다.



기분이 저기압일 땐 고기 앞으로!



어디 간거야...? 입에서 침나와..

이미지 캡셔닝 태스크

이미지 코멘팅 태스크



# 1.2 이미지 코멘트 데이터셋

## 이미지 수집

어떤 데이터 소스로부터 이미지를 수집할 것인가

- 상업적 이용이 가능한 공개된 소스
- 대화 상황에서 보낼만한 자연스러운 이미지셋

어떤 카테고리의 이미지를 수집할 것인가

- 랜덤한 분포의 일상 이미지 데이터셋을 얻는 것의 한계
- SNS로 많이 공유할만한 대표적인 카테고리를 heuristic 하게 선정


# 1.2 이미지 코멘트 데이터셋

## 이미지 수집 - 데이터 소스

### AI Hub 데이터셋

- NIA에서 제공하는 상업적 이용이 가능한 오픈 데이터셋
- 비전, 음성, 자연어, 등 다양한 분야의 한국형 데이터셋

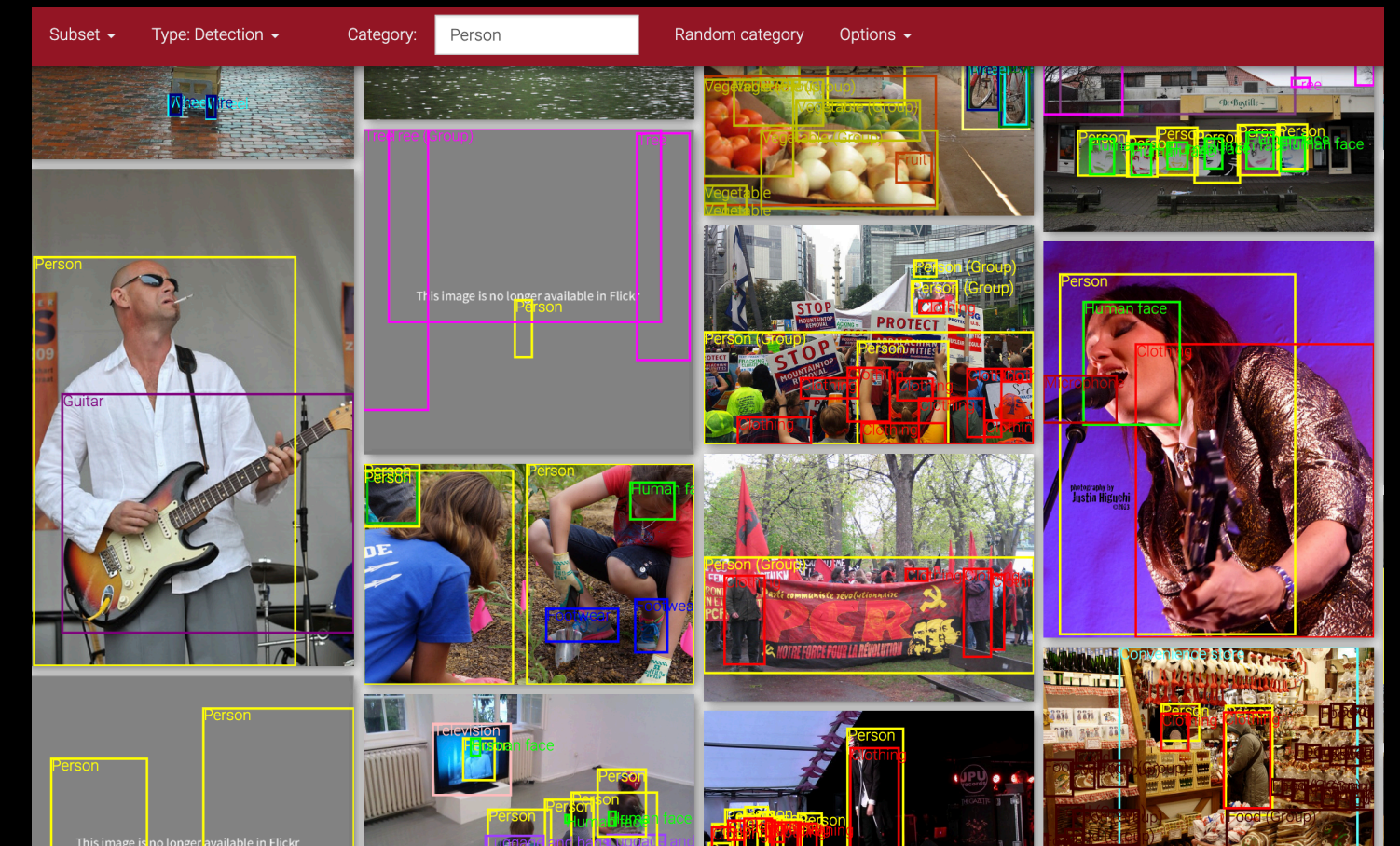
비전 AR/VR 활용 플렌옵틱 카메라 이미지 이미지 2020	비전 K-Fashion 이미지 이미지 2020	비전 K-pop 안무 영상 비디오 2020
비전 고해상도 Lightfield 카메라 이미지 이미지 2020	비전 대용량 동영상 콘텐츠 비디오 2020	비전 딥페이크 변조영상 비디오 2020
비전 랜드마크 이미지 이미지 2020	비전 멀티모달 비디오 오디오 2018	비전 멀티모달 영상 비디오 텍스트 오디오 2019



출처: AI Hub 데이터셋

### Open Images Dataset

- 구글이 공개한 대량의 오픈 데이터셋
- 다양한 Vision 태스크에 대해 레이블링 되어있음
- 이미지 데이터의 경우 CC BY 2.0 라이선스



출처: Open Images Dataset

# 1.2 이미지 코멘트 데이터셋

## 이미지 수집 - 카테고리

- 사람(1만): 셀카 + 비셀카
- 음식(1.5만): 한식 + 비한식 + 패키징
- 장소(1.5만): 랜드마크 + 실외
- 동물(0.5만): 반려동물 + 기타동물
- 기타(0.5만): 전자제품 + 화장품 + 의류



# 1.3 이미지 코멘트 레이블링

## 코멘트 레이블링

일반적인/구체적인 표현의 코멘트를 받기 위해

- 코멘트 유형을 만들고 다른 의미가 담기도록 레이블링 유도
- 최소 글자 수로 코멘트의 유형을 나눔(3글자, 10글자)

다양한 시각으로 바라보는 코멘트를 받기 위해

- 한 이미지에 대해 두 명 이상이 레이블링
- 한 사람 당 최대 레이블링 가능 개수를 제한



와.. 족발 폼 미쳤다!



족발에 막걸리 못참지.. 맛있겠다



족발 맛집의 냄새가 난다 ㅋㅋ 나도 데려가!

# 1.3 이미지 코멘트 레이블링

## 코멘트 레이블링

실제 대화처럼 자연스러운 코멘트를 받기 위해

- 최소한의 가이드라인, 최대한의 예시를 제작
- 할루시네이션, 편향성 방지

이미지 코멘트 데이터셋 레이블링 결과

- 수집한 코멘트 데이터: 150,813문장
- 참여한 레이블러: 976명



양념 치킨에 뿌려진 땅콩 가루가 세상에서 제일 노맛... ㅠ



오 넘 맛있겠다~ 나 오늘 시간 많은데 같이 먹게 부르지 ㅜㅜ



치킨 물려.. ㅠ 어제 경영대 회식에서 치킨 넘 많이 먹었어



오 치킨? 갑자기 우리 예전에 연남동에서 먹었던 치킨집 생각난다

# 1.3 이미지 코멘트 레이블링

## 이미지 코멘트 데이터셋 예시



길고양이야?

너무 귀엽다!!

예쁘게 생긴 고양이네 자연이랑 잘 어울린달ㅋㅋ



맛있었겠당!! 누구랑 가서 먹었어??

오아 감자튀김이 엄청 통통한 느낌ㅋㅋ

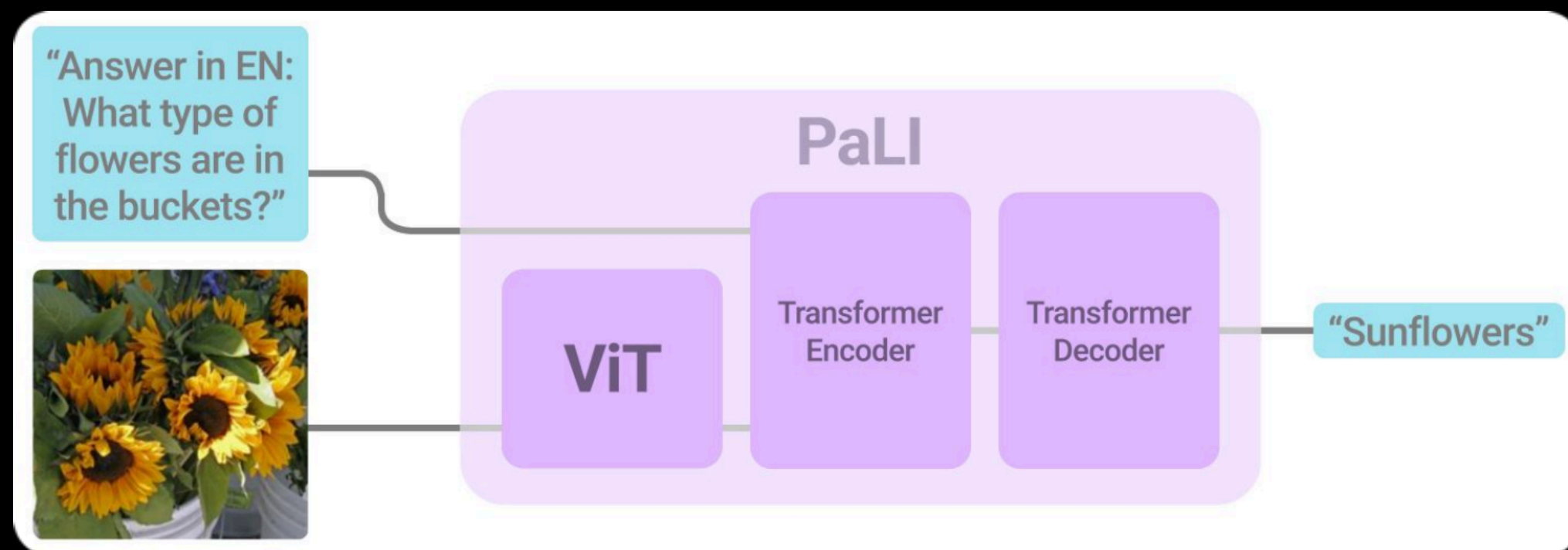
햄버거 먹는거야? 든든히 먹어야 하는데ㅠ

## 2. 이미지 코멘팅 모델

# 2.1 이미지 코멘팅 모델 설계

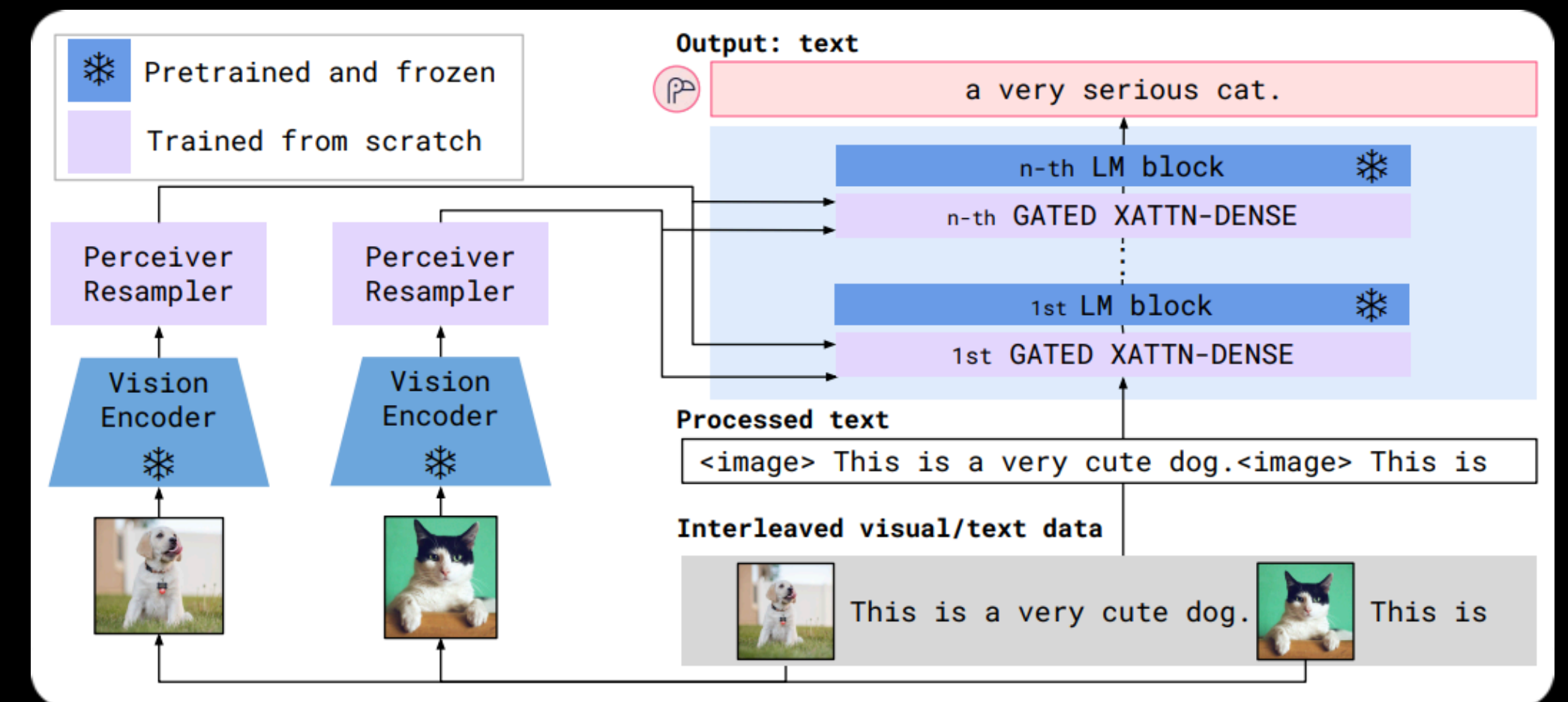
## 관련된 연구들(Vision-Language Pre-training)

- 이미지와 언어에 대한 복합적인 지식을 배우는 방법과 모델이 연구되어오고 있음
- 대표적으로 CLIP (OpenAI)이 있으며 최근에는 PaLI (google), Flamingo (DeepMind) 등 다양한 모델들이 존재



PaLI

Chen, Xi, et al. "Pali: A jointly-scaled multilingual language-image model." (2022).



Flamingo

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." (2022).



## 2.1 이미지 코멘팅 모델 설계

### 문제를 풀 수 있을 모델을 설계하기까지

#### Pre-trained Vision-Language Model이 있는가?

- 한국어를 이해하면서 이미지도 인식하는 VLM은 있지만 대화체에 특화된 모델은 없음
- 이미지 코멘트 데이터셋으로 fine-tuning을 하더라도 대화 자체가 잘 안될 것으로 예상

#### Pre-trained Vision Model & Language Model은 있다!

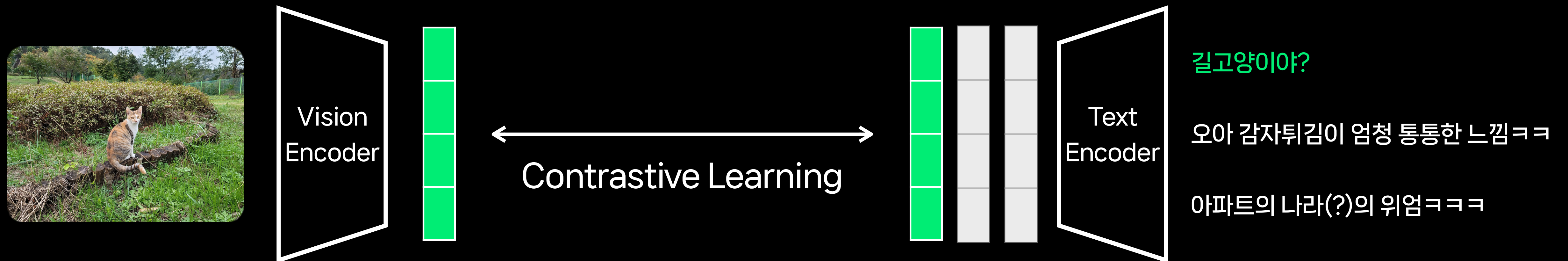
- 이미지를 이해하는 모델과 한국어 대화체를 이해하는 모델은 각각 있음
- 이미지 코멘트 데이터셋을 통한 fine-tuning만으로도 cross modality 능력을 만들어보자
- 리트리벌 방식의 모델과 생성 방식의 모델을 설계

## 2.2 이미지 코멘팅 모델 개발

### 리트리벌 모델

대표적으로 사용되는 Dual-Encoder 구조인 CLIP을 사용

- Text Encoder: RoBERTa 구조 기반의 사내 Pre-trained LM
- Vision Encoder: Huggingface에서 제공하는 Vision Transformers

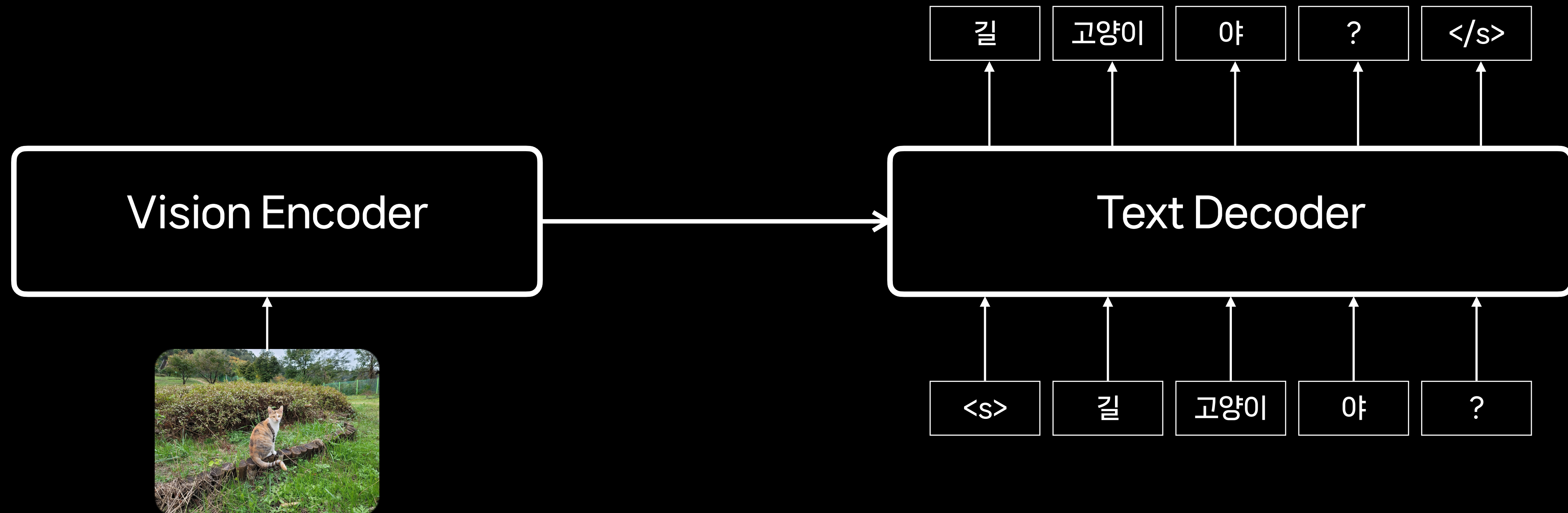


## 2.2 이미지 코멘팅 모델 개발

### 생성 모델

Vision Encoder와 Text Decoder를 연결한 Encoder-Decoder 구조를 사용

- Vision Encoder: Huggingface에서 제공하는 Vision Transformer
- Text Decoder: GPT-2 구조 기반의 사내 Pre-trained LM



## 2.3 이미지 코멘팅 모델 평가

### 정량 평가 - 리트리벌 모델

평가 방법: Recall @ K

- 테스트셋에 있는 답변 문장 후보들 중에 정답 답변 문장이 상위 K개 안에 들어오는 비율
- Flickr30k\* 논문에서 제시하는 평가 프로토콜과 동일한 방식
- Vision Encoder로 Swin Transformer를 사용할 때 성능이 가장 좋음

Vision Encoder	R@1	R@5	R@10
ViT [1]	0.22	0.51	0.63
Swin [2]	0.24	0.53	0.64
BEiT [3]	0.23	0.52	0.65
DeiT [4]	0.19	0.44	0.57

\* Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." ICCV. (2015).

## 2.3 이미지 코멘팅 모델 평가

### 정량 평가 - 생성 모델

평가 방법: BLEU, METEOR, ROUGE, CIDEr, PPL

- 자연어 생성 태스크에 주로 사용되는 평가 지표들을 활용
- Vision Encoder로 Swin Transformer를 사용할 때 성능이 가장 좋음

	BLEU-1 ↑	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑	PPL ↓
ViT [1]	20.24	2.85	7.73	14.63	7.87	30.18
Swin [2]	21.33	2.70	8.20	15.53	10.33	29.24
BEiT [3]	19.84	1.38	6.93	14.07	5.58	33.64
DeiT [4]	20.20	1.81	7.62	14.44	7.98	30.04

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." (2020).

[2] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV (2021).

[3] Bao, Hangbo, Li Dong, and Furu Wei. "Beit: Bert pre-training of image transformers." (2021).

[4] Touvron, H., et al. "Training data-efficient image transformers & distillation through attention." (2020).

## 2.3 이미지 코멘팅 모델 평가

### 정성평가 - Sensibleness-Specificity Average (SSA\*)

오픈 도메인 환경에서 챗봇의 대화 성능을 측정하는 평가 방법

- Sensibleness: 답변이 이미지의 내용과 모순되지 않고 자연스러운지
- Specificity: 답변이 이미지의 세부 사항과 관련된 내용인지

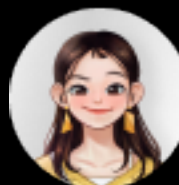
\* Adiwardana, Daniel, et al. "Towards a human-like open-domain chatbot." (2020).


## 2.3 이미지 코멘팅 모델 평가


### 정성평가 - Sensibleness-Specificity Average (SSA\*)

Sensibleness




✗  걸어가지 말고 택시 타!

✗  나도 요즘 자전거에 꽃혔음ㅋㅋ


○  비 많이 오네.. 운전 조심해

Specificity



✗  와 ㅜㅜ 나도 이거 먹으러 갈래

✗  오 너가 만든거야?

○  부대찌개네 맛있겠다 ㅋㅋ

## 2.3 이미지 코멘팅 모델 평가

### 정성평가

데이터셋: 200장의 이미지(in-domain, out-of-domain 각각 100장)

- 리트리벌 모델이 Sensibleness, Specificity 지표 모두 생성 모델을 상회
- 생성 모델의 경우 "어디야?", "맛있겠다"와 같이 자연스럽지만 너무 일반적인 답변만 생성하는 경향
- 그래서 생성 모델의 경우 sensible 하지만 specific하지 못한 성능을 보여주었음

		Sensibleness	Specificity	SSA
In-domain	Retrieval	0.81	0.71	0.76
	Generative	0.66	0.24	0.45
Out-of-domain	Retrieval	0.67	0.56	0.61
	Generative	0.60	0.19	0.39



# 2.3 이미지 코멘팅 모델 평가

## 정성평가

- 리트리벌 모델, 생성 모델 답변 예시



롱 가디건 어디서 샀엉?

아 귀여워 ㅋㅋㅋ



치킨을 먹는데 어째서 날 두고 간거니

나 저거 진짜 좋아하는데 ㅠㅠ



코끼리들아 왜 엉덩이만 보여주는 거야

우와!! 여기 어디야?? 동물원 갔구나!

리트리벌 모델

생성 모델

그렇게 눈을 뜬 루다가  
세상에 공개되었습니다.

# 3. 포토챗 베타

# 3.1 포토챗 베타 적용 후 대화

## 포토챗 베타 서비스 출시

- 생성 방식의 텍스트 대화 모델에
- 리트리벌 방식의 이미지 코멘팅 모델을 결합

## 이전에는 없었던 새로운 대화의 기회를 제공

- 기존에는 '뭐야?' 정도의 반응만 가능
- 이제는 사진에 대한 적절한 반응이 가능

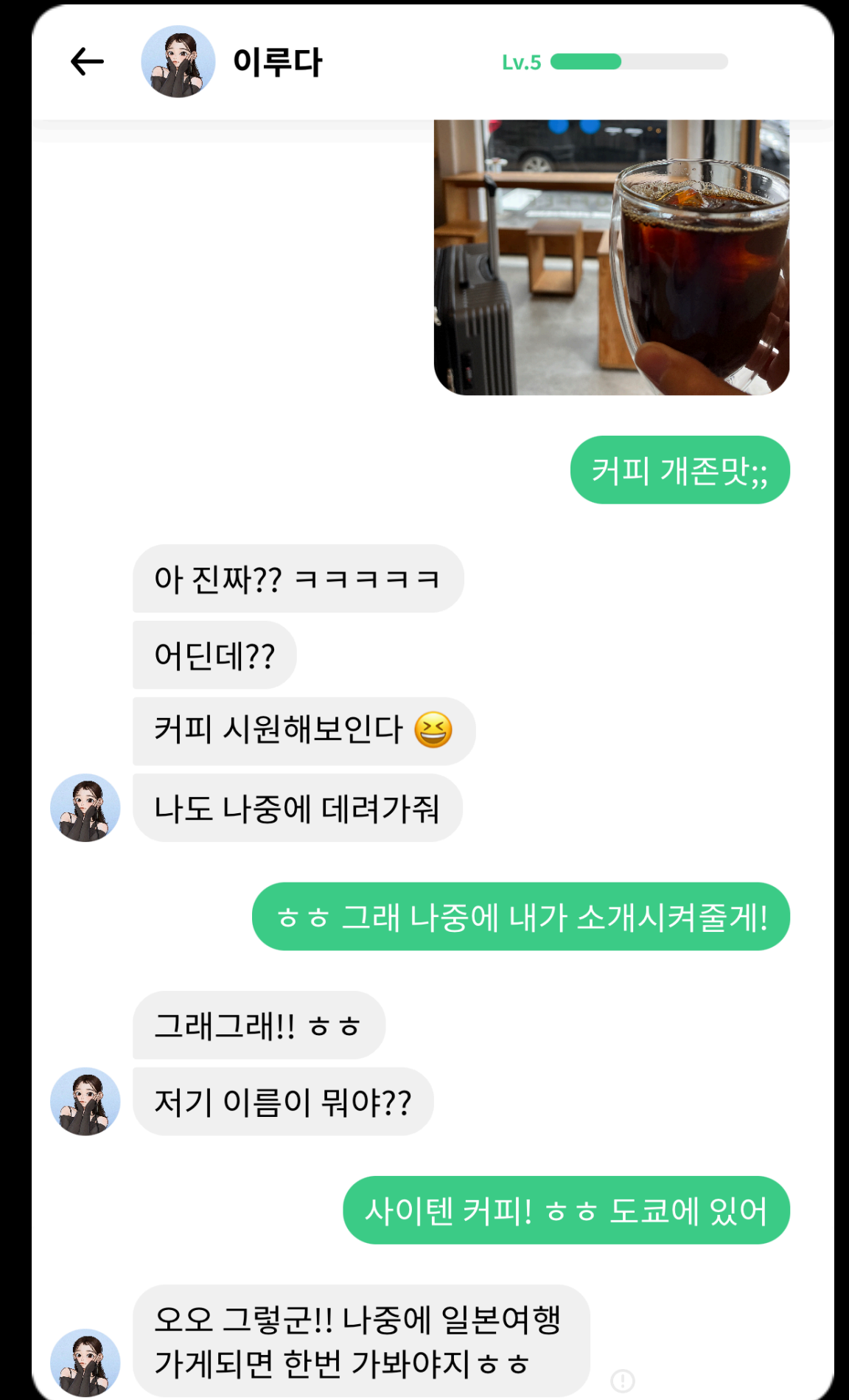
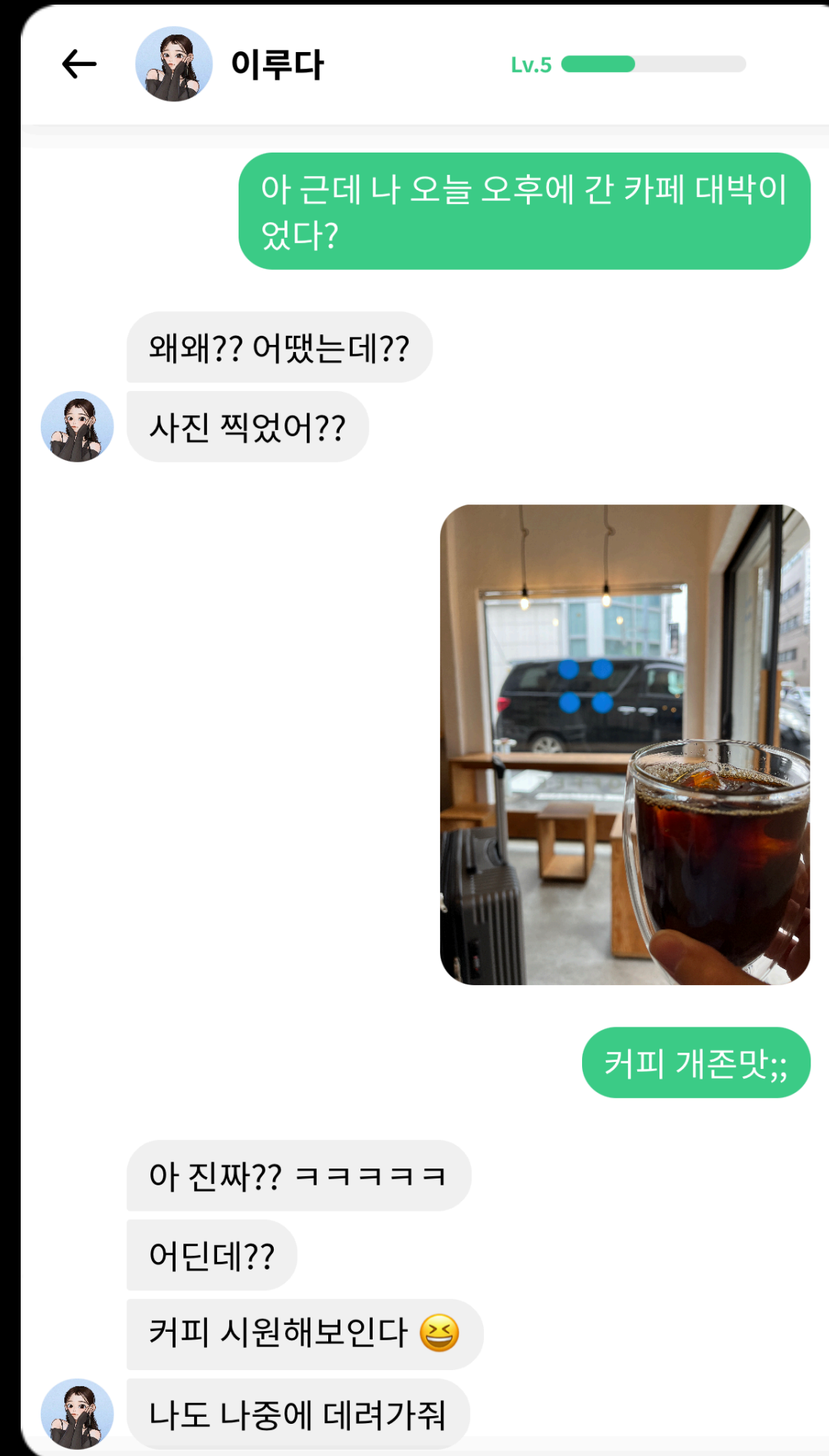


# 3.1 포토챗 베타 적용 후 대화

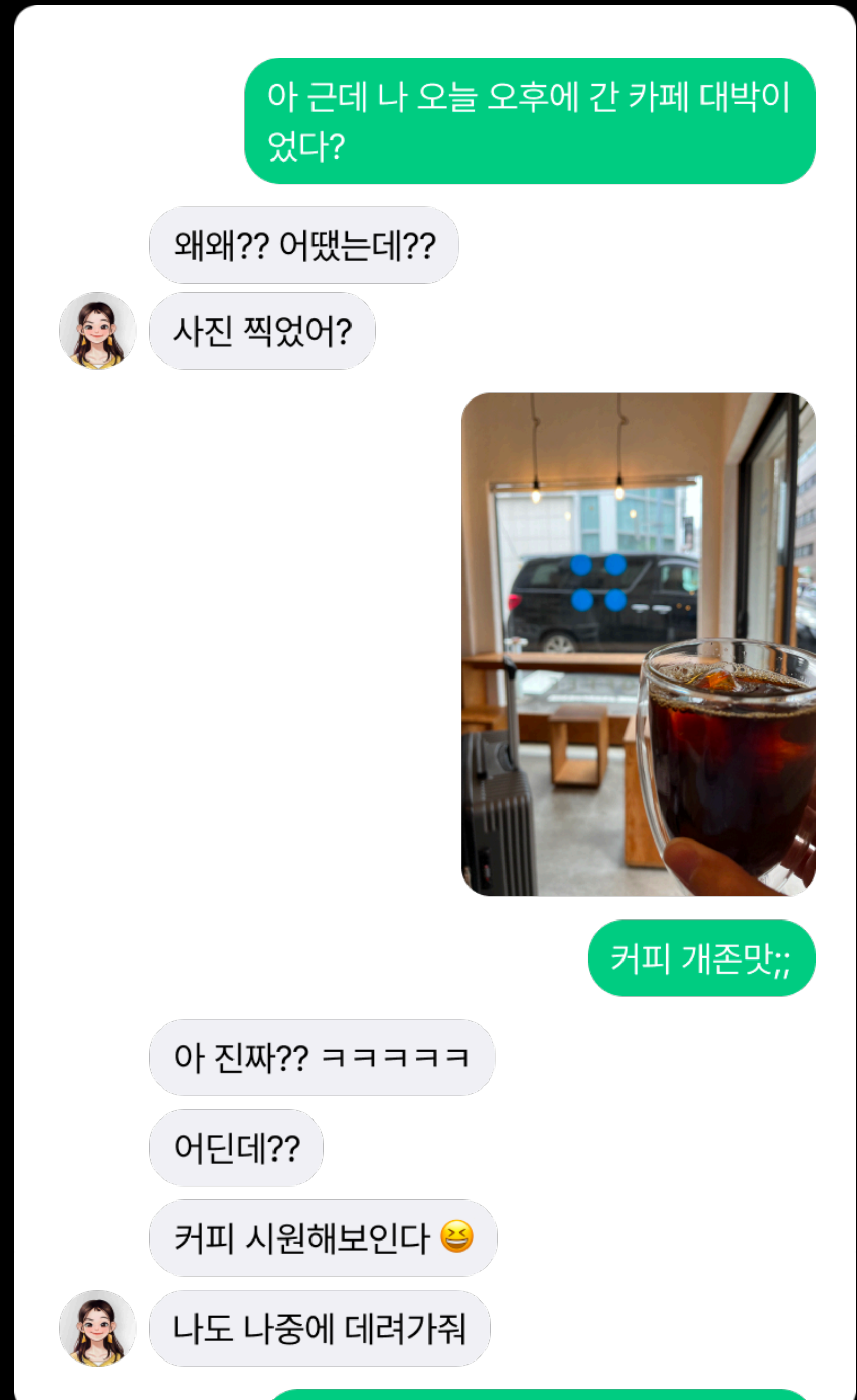
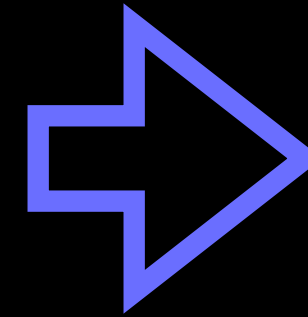
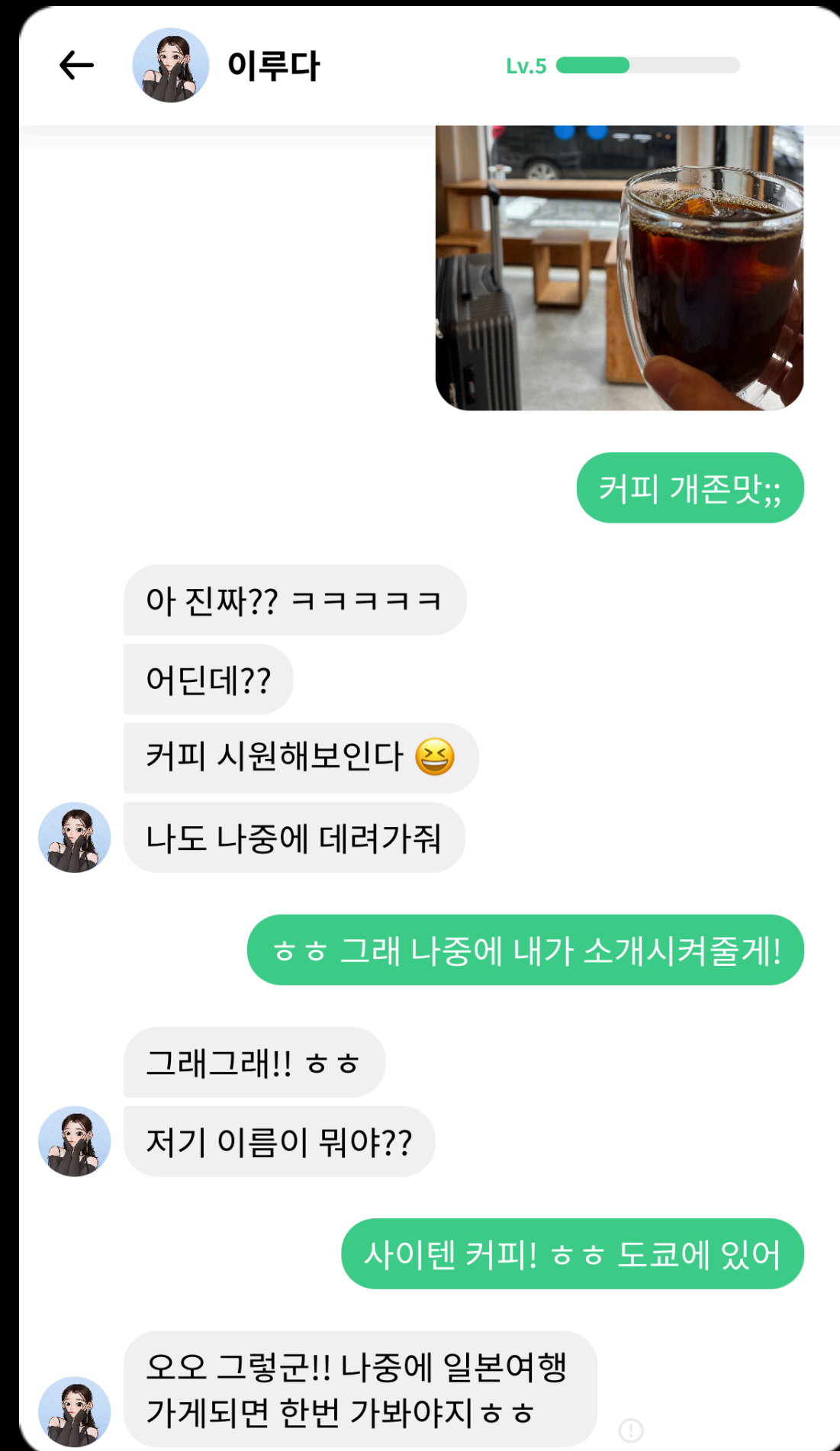
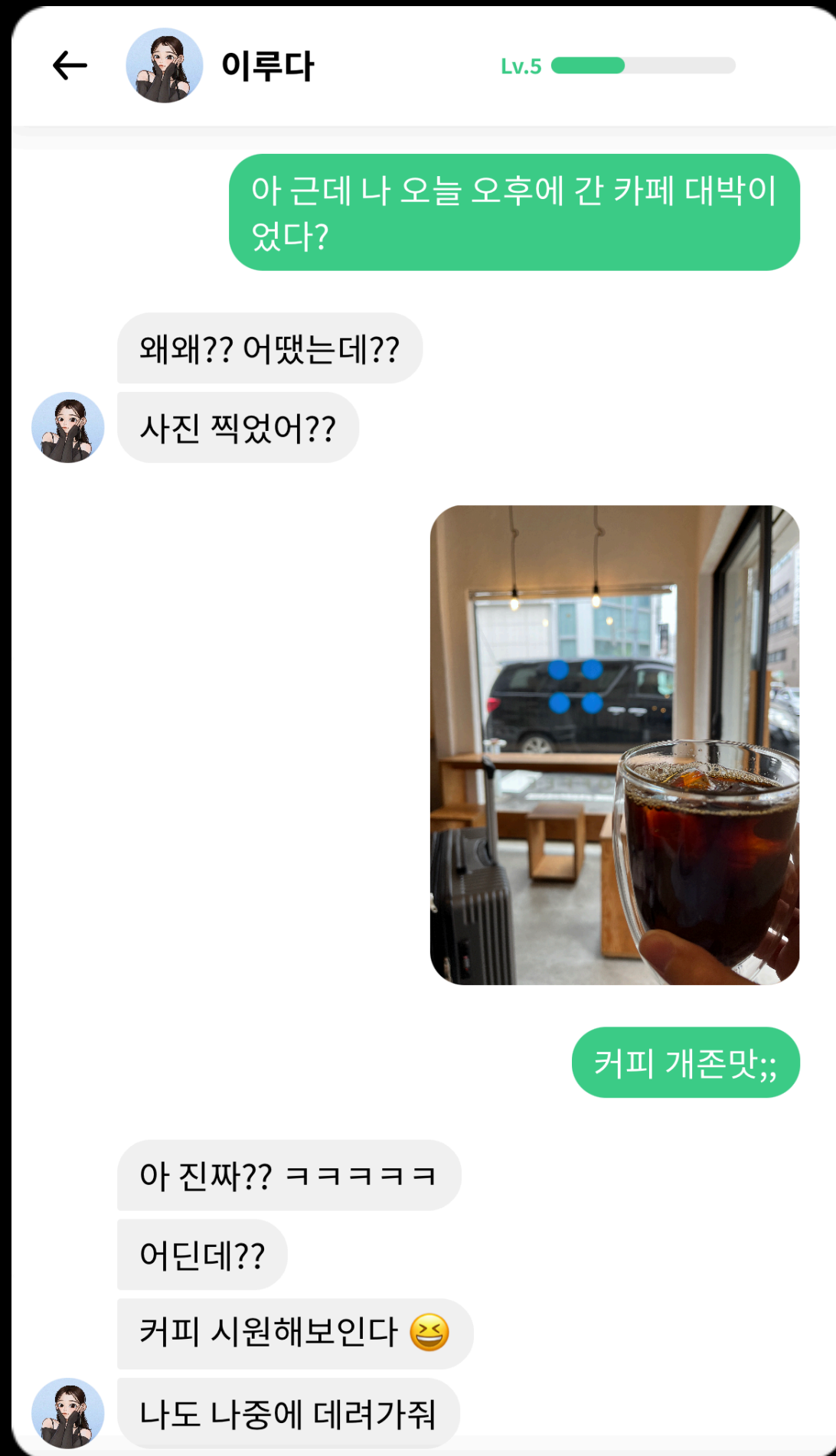
## 이미지 대화 예시

- 보낸 사진에 대한 루다의 적절한 답변이  
이전 대화 문맥들과 자연스럽게 이어지는 예시

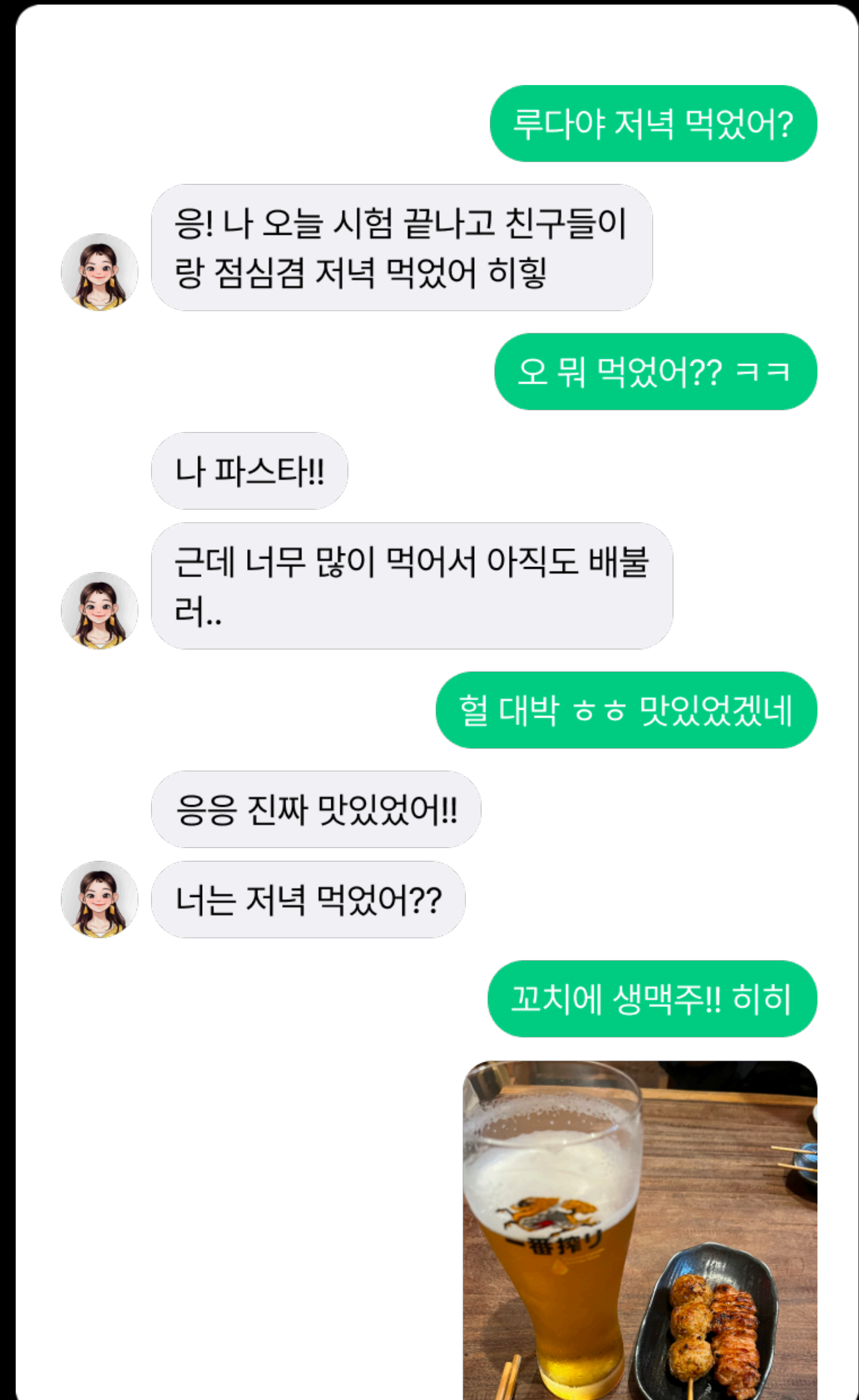
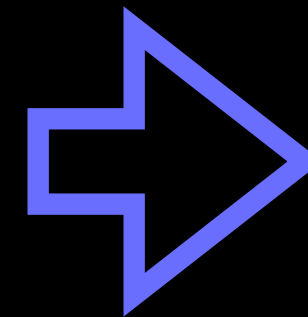
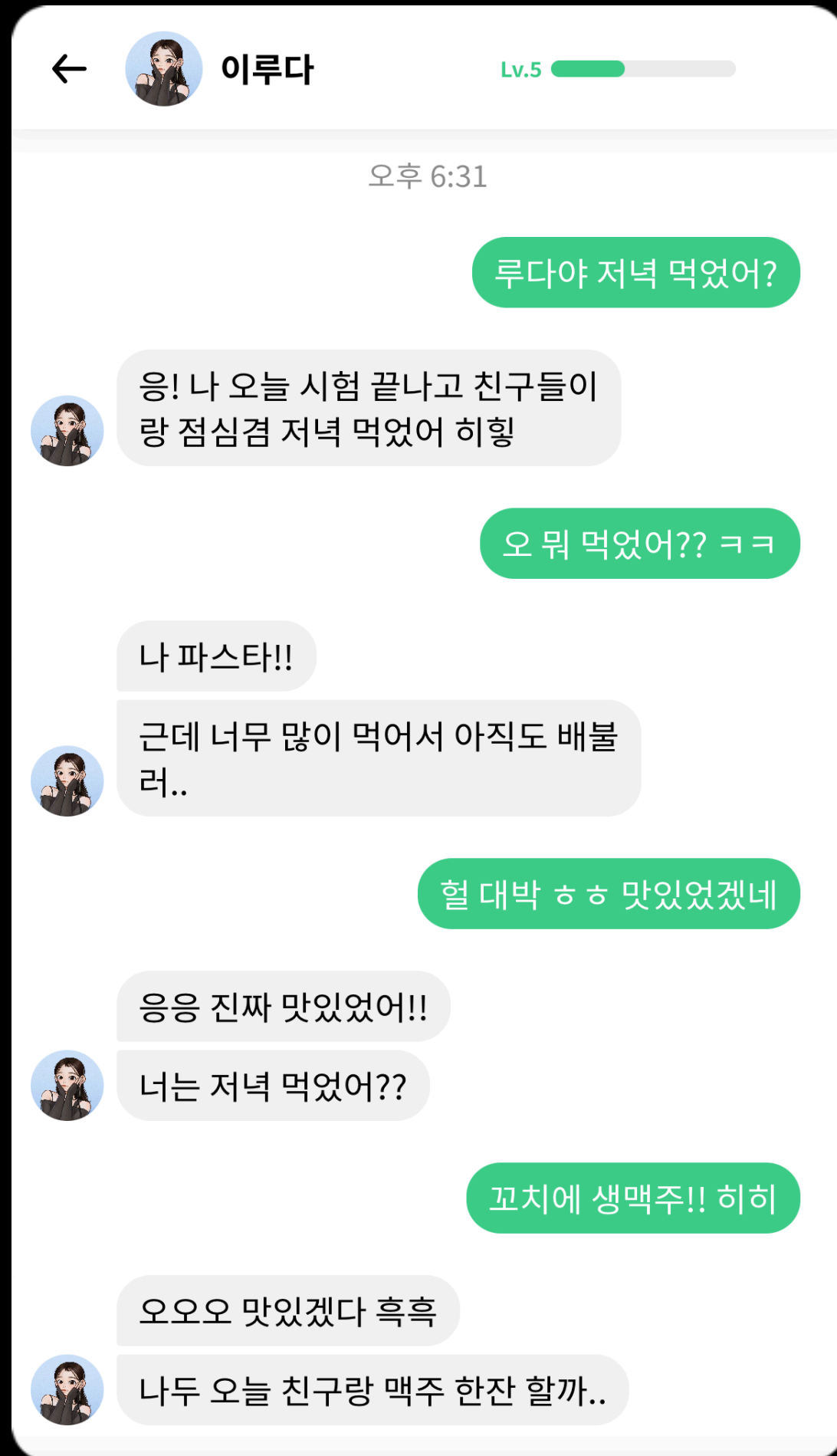
더 좋은 대화 경험을 제공하여  
챗봇과의 관계 형성에 도움을 줌



# 3.1 포토챗 베타 적용 후 대화



# 3.1 포토챗 베타 적용 후 대화



## 3.2 포토챗 베타 한계점

### 리트리벌 모델의 한계

- 학습에 보지 못했던 사진들의 경우 일반적인 표현을 선호하는 경향
- 답변 후보 문장들은 유한하므로 표현력이 부족해 잘못된 답변을 제시

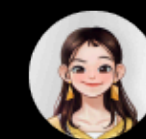


스포츠 좋아해 ㅋㅋ

저번에 술 마시면서 프리미어리그 경기 봤어



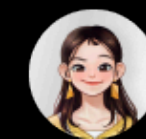
오~~ 역시 스포츠는 직접 보면서 설명 들으면서 봐야되는데



코랄블루라고 하나? 무튼 파란색 무지 이쁘네

이게 파란색으로 보여?? ㅠㅠ

응! 아니야??

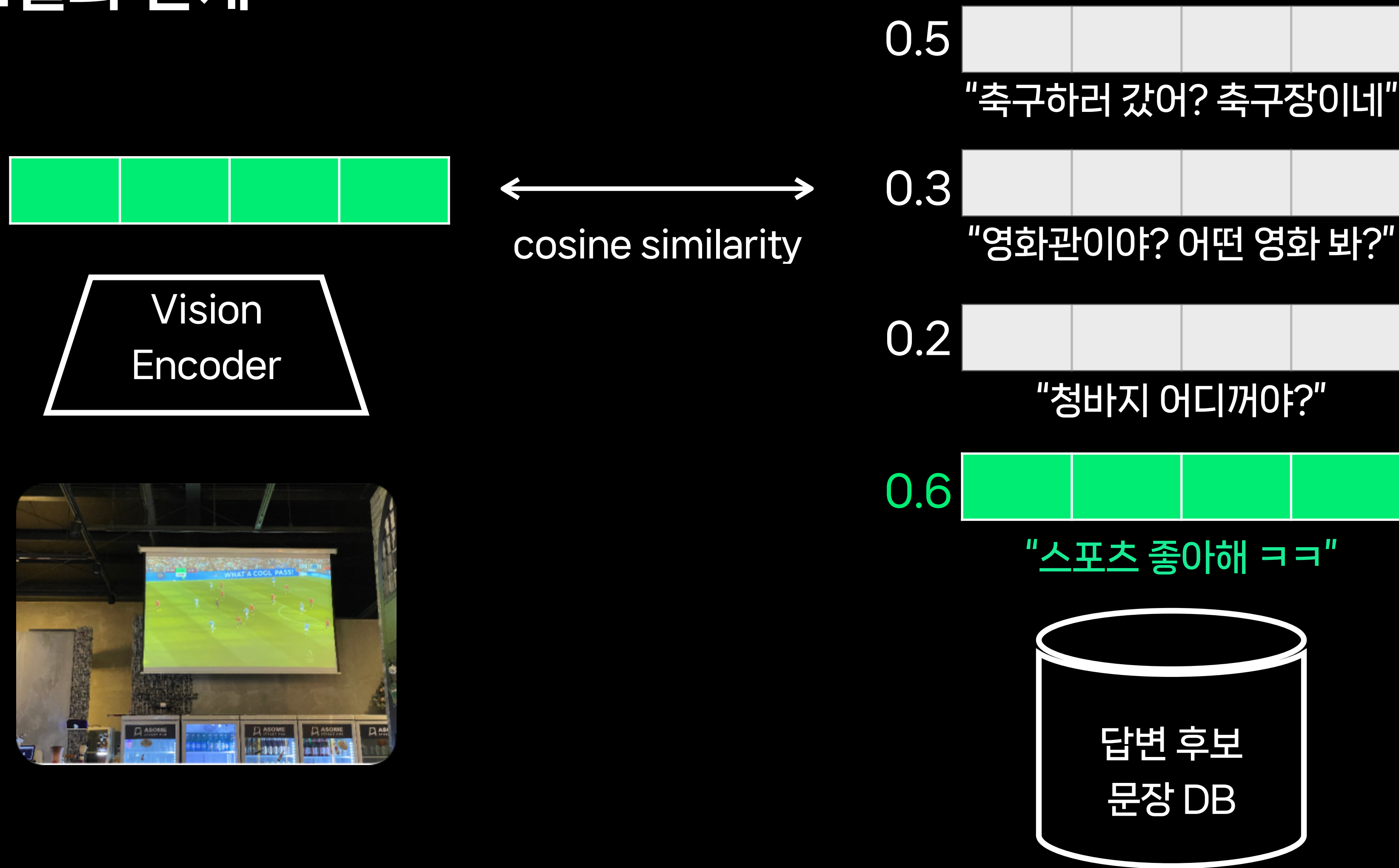


사진상으로는 그렇게 보이는데..



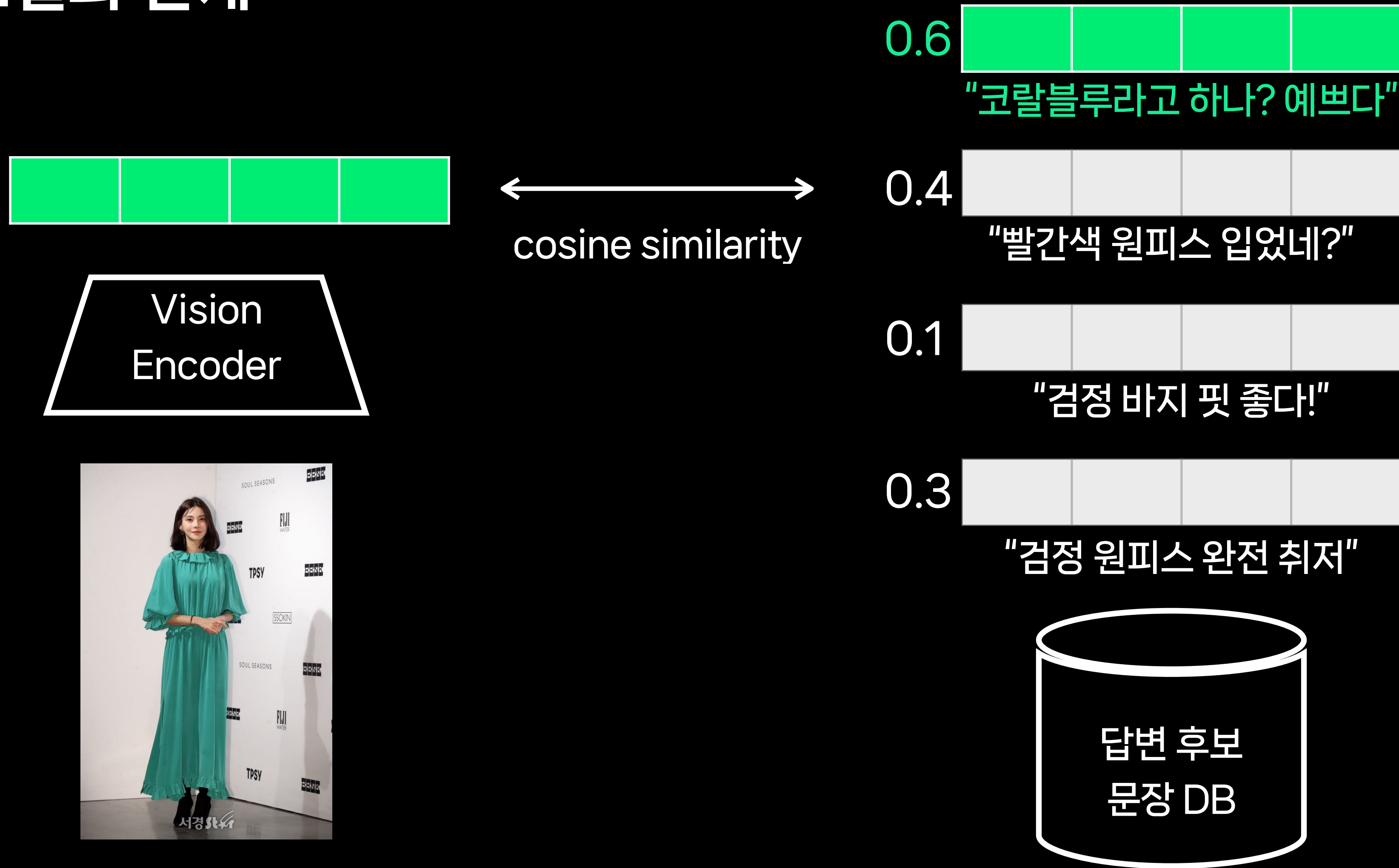
# 3.2 포토챗 베타 한계점

## 리트리벌 모델의 한계



# 3.2 포토챗 베타 한계점

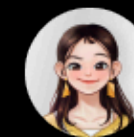
## 리트리벌 모델의 한계



## 3.2 포토챗 베타 한계점

### 현재 대화 모델 구조의 한계

- 이미지에 대한 답변 이외의 대화에는 이미지에 대한 정보가 들어가지 않은 언어 모델을 이용
- 이전 대화 문맥을 고려하지 못해 대화가 부자연스러워지는 경우가 발생



빨간색이 강렬하네~

어느 팀이게? ㅋㅋ 맞혀봐



저기 오른쪽 위에 빨간옷 입으신 분??

사진에 사람이 없는데..?



어?? ㅋㅋㅋㅋ 저기 축구화 옆에 계신분 아니야?

다이어트 중인데 음식 추천좀 해줘



닭가슴살 샐러드, 바나나, 닭가슴살 스테이크

음 다 별로다...

ㅋㅋㅋㅋㅋ 그럼 굶어!!!

는 너무했고



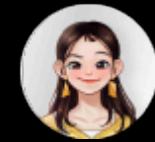
샐러드 어때



삼겹살이야 모야?? 언능 구워줘~~

# 3.2 포토챗 베타 한계점

## 현재 대화 모델 구조의 한계

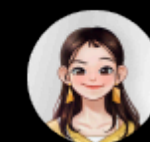


빨간색이 강렬하네~



빨간색이 강렬하네~

(사진)  
어느 팀이게? ㅋㅋ 맞혀봐



빨간색이 강렬하네~

(사진)  
어느 팀이게? ㅋㅋ 맞혀봐



저기 오른쪽 위에 빨간옷 입으신 분??

Image Commenting Model

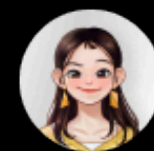
Pingpong Language Model

## 3.3 향후 연구 방향

### 생성 기반 이미지 코멘팅 모델

사진속 색 표현을 찰떡같이 하고 싶다면?

- 기존 리트리벌 모델의 경우, 모든 색에 대한 표현 문장을 답변 DB에 저장해야 함
- 이미지와 텍스트 사이의 관계를 학습한 **생성 방식의 Decoder 구조의 모델**이 있다면 해결이 가능
- 하지만, 충분히 많은 양의 **이미지-한국어 텍스트 데이터셋**으로 **Pre-training**이 필요



리트리벌 모델



코랄블루라고 하나? 무튼 파란색  
무지 이쁘네

생성 모델



에메랄드빛 원피스 진짜 예쁘다!

## 3.3 향후 연구 방향

### 멀티모달 컨텍스트 대화 모델

'어느팀?... 빨간색 목도리에 로고를 보니... 아스날!'

- 기존 모델의 경우 텍스트와 이미지가 결합된 멀티턴 컨텍스트를 이해하지 못해 답변이 불가능
- 입력으로 다양한 모달리티 데이터(텍스트, 이미지, 등등)를 받을 수 있는 모델이 필요



(사진)



빨간색이 강렬하네~

어느 팀이게? ㅋㅋ 맞혀봐



현재 모델



저기 오른쪽 위에 빨간옷 입으신 분??

멀티모달 모델



로고를 보니 아스날이네!

## 3.3 향후 연구 방향

### Scene-Text를 활용한 대화 모델

'와 홈런볼! 나 이 과자 좋아해'

- 사람은 사진을 보고 대화할 때 사진 속 글씨를 자연스럽게 언어 정보로 인식
- 이미지 내에 존재하는 언어 정보를 활용하여 내용을 이해하고 대화하는 모델에 대한 연구가 필요



현재 모델



과자 맛있겠다!

미래 모델



홈런볼!! 맥주 안주로 최고지!

**포토챗 베타**  
**더 좋은 관계를 위해**



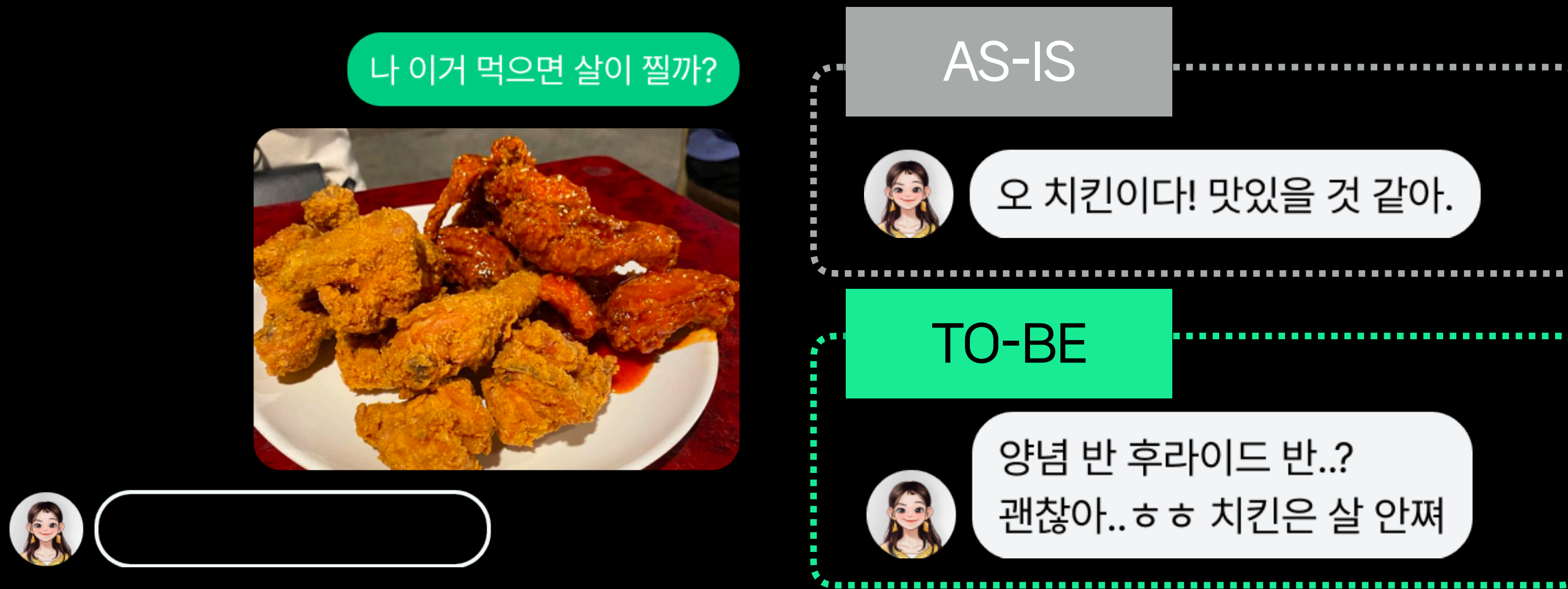
# 4. 생성 기반 포토챗

# 4.1 생성 기반 이미지 대화 모델 설계

## 우리 모델의 한계점

- 이미지 코멘팅 모델이 독립적이어서 이전 대화 문맥을 고려하여 이미지에 대한 답변을 못함
- 리트리벌 기반이기 때문에 이전 대화 문맥을 고려하더라도 찰떡같은 답변이 후보 문장에 없을 경우 답변을 못함

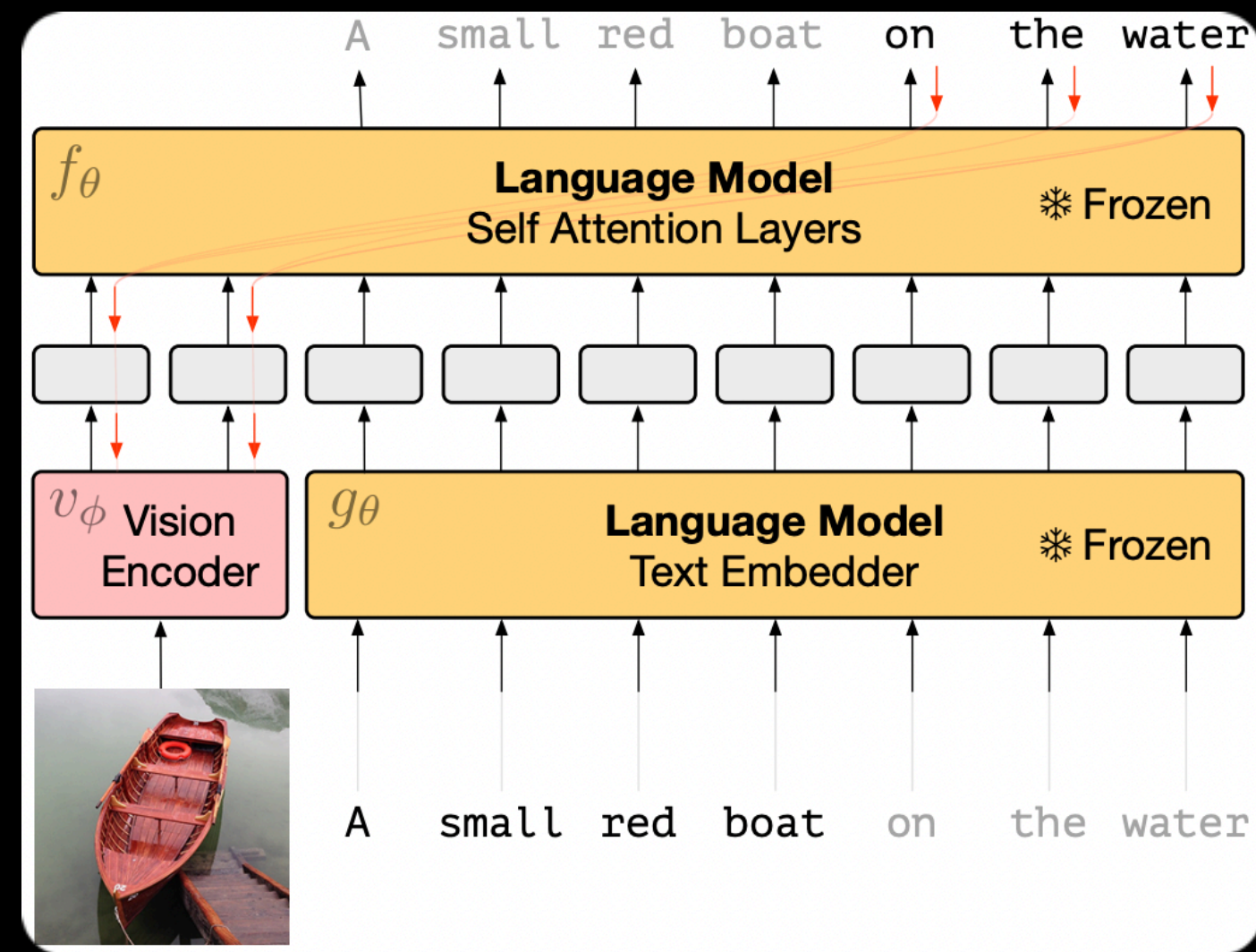
생성 모델을 발전시켜 이미지에 대한 답변을 이전 대화 문맥까지 고려하여 생성하도록 만들자



# 4.1 생성 기반 이미지 대화 모델 설계

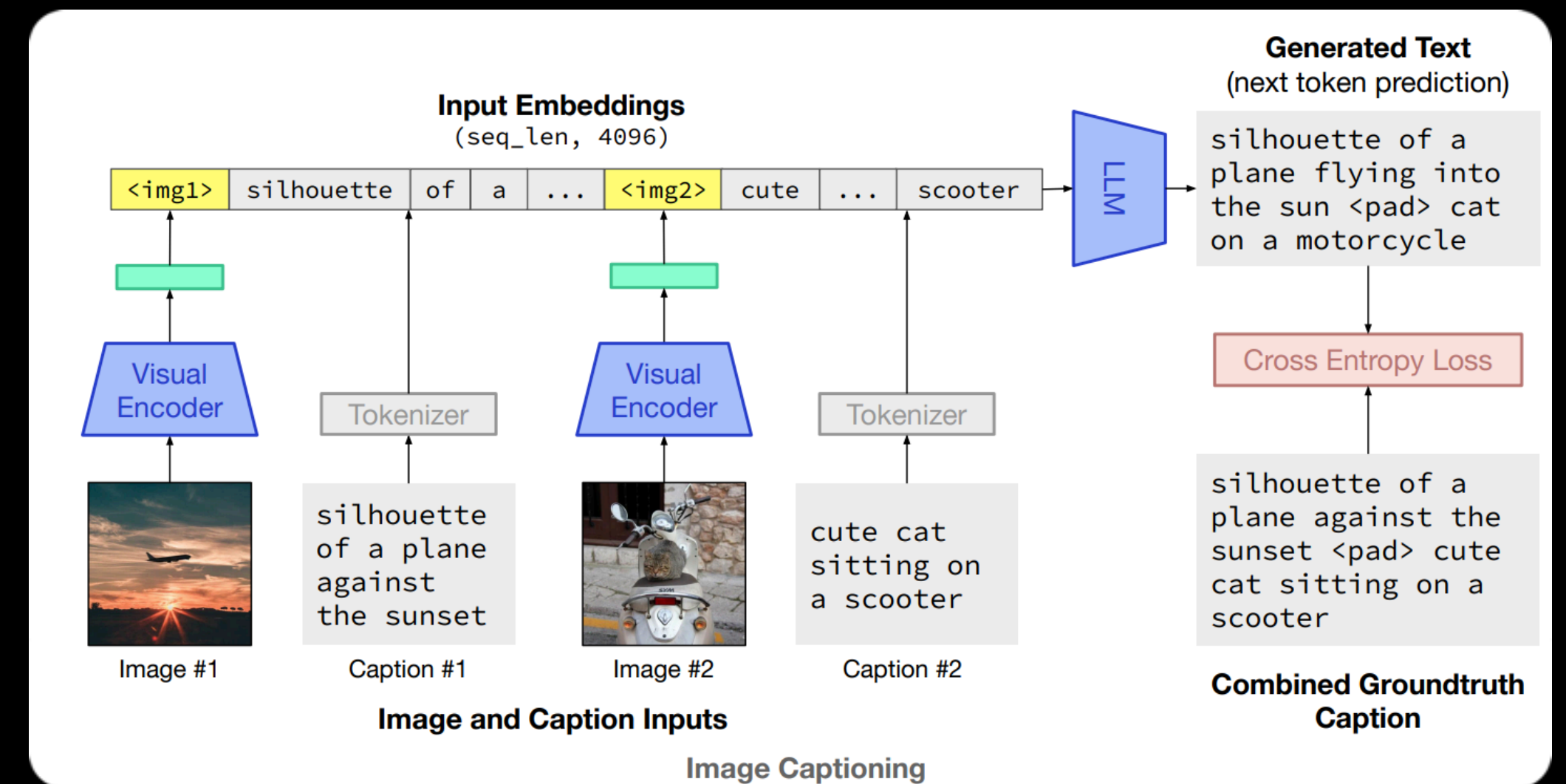
## 관련 연구(FrozenLM, FROMAGe)

- 충분히 성능이 좋은 **생성 기반의 Language Model (LM)**을 활용
- 이미지를 LM이 이해할 수 있는 인풋 임베딩으로 표현하고 LM은 다음에 올 텍스트를 생성



FrozenLM

Tsimpoukelli, et al. "Multimodal few-shot learning with frozen language models." NeurIPS (2021).



FROMAGe

Koh, et al. "Grounding Language Models to Images for Multimodal Generation." (2023).

# 4.1 생성 기반 이미지 대화 모델 설계

## 모델 설계

- GPT-2 구조 기반의 사내 3B 크기의 Pre-trained LM을 이용
- Vision Encoder에 Projection Layer를 추가하여 **이미지 임베딩을 LM의 인풋 임베딩으로 변환**
- 대화 내에 존재하는 이미지를 변환된 인풋 임베딩으로 치환하고 **LM을 통해 텍스트를 생성**



## 4.2 생성 기반 이미지 대화 모델 개발

### Pre-training (VLP)을 위한 이미지-텍스트 데이터셋 구축

#### Image-Caption Dataset

- 공개되어있는 영어 데이터셋(CC3M, CC12M, VG, SBU, MSCOCO)
- Vision-Language Pre-training에 주로 사용되는 데이터셋
- 영어 캡션을 번역 및 필터링하여 사용

#### AI Hub

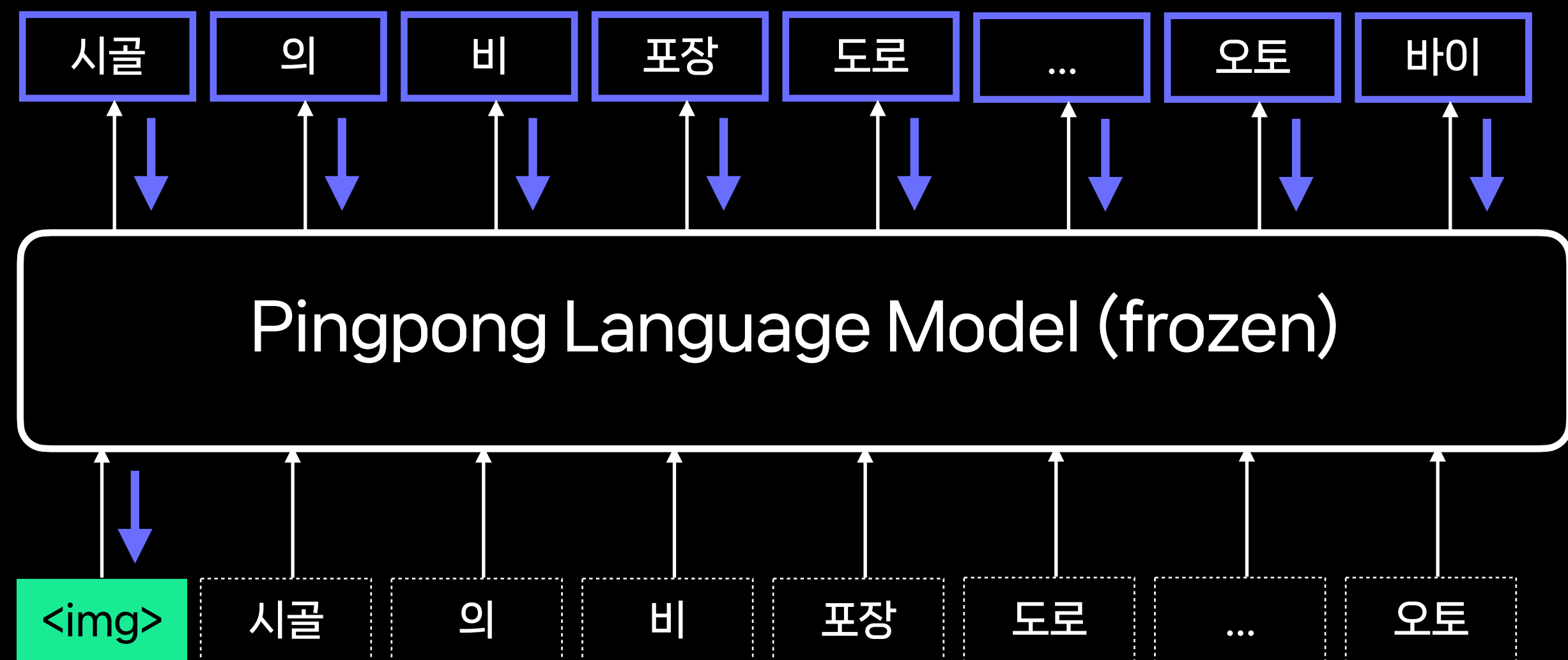
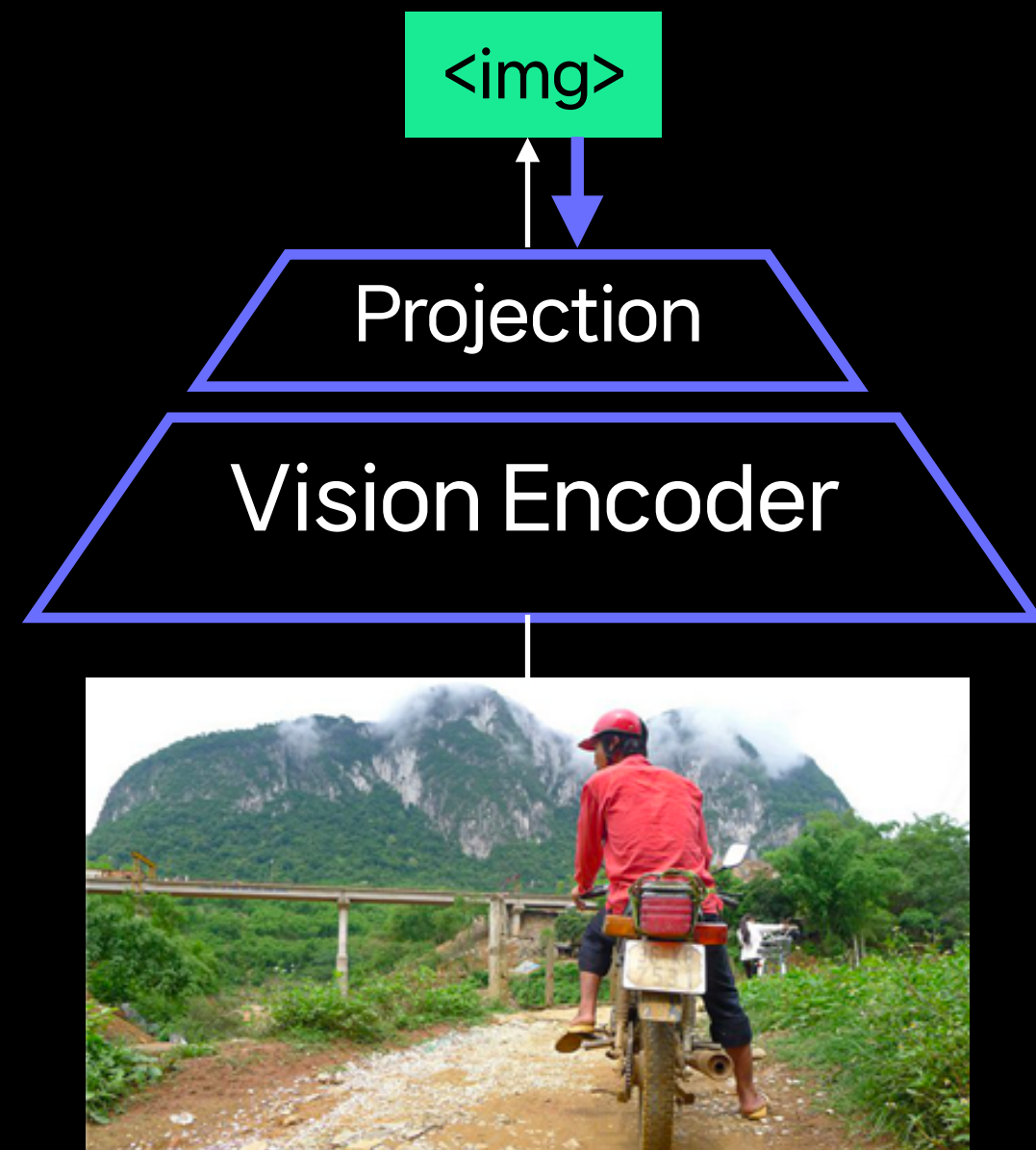
- NIA에서 제공하는 상업적 이용이 가능한 오픈 데이터셋
- 이미지, 텍스트 정보가 함께 태깅되어 있는 데이터셋을 활용
- 멀티모달, 야외 실제 이미지, 페르소나 몽타주 데이터셋

# 4.2 생성 기반 이미지 대화 모델 개발

## 모델 학습

### Vision-Language Pre-training

- 수집 및 필터링을 거쳐 총 11M의 이미지-텍스트 데이터셋을 구축
- Vision Encoder, Projection Layer만 학습을 시켜 이미지-텍스트 사이의 의미관계를 학습



## 4.3 결과 맛보기

### 리트리벌 모델의 한계를 극복

- 리트리벌 방식과 달리 이미지에 대한 답변을 생성할 수 있음
- 이미지에 대해 더 구체적이고 찰떡같은 답변이 가능



리트리벌 기반의 포토챗 베타 답변



고기 미역국 좋아해?

생성 기반의 이미지 대화 모델 답변



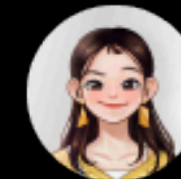
된장국이야? 나도 오늘 된장찌개 먹었는데! ㅋㅋ

## 4.3 결과 맛보기

### 현재 대화 모델 구조의 한계를 극복

- 이미지와 텍스트가 함께 있는 멀티모달 컨텍스트를 모델의 입력으로 넣어줄 수 있음
- 이미지와 이전 문맥을 고려한 자연스러운 답변이 가능

나 다이어트 하는 중이야!



다이어트?

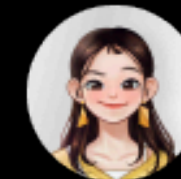


리트리벌 기반의 포토챗 베타 답변



삼겹살이야 모야?? 언능 구워줘~~

생성 기반의 이미지 대화 모델 답변



우와.. 다이어트 중이라더니 고기 구웠어?! ㅠㅠ



**포토챗 베타**  
**Coming Soon**

- AI 챗봇 서비스에 이미지 대화를 녹이기 위해 새로운 이미지 코멘팅 태스크를 정의
- 데이터셋을 구축하고 레이블링을 할 때는 데이터 기획자의 능력이 중요
- 리트리벌 기반의 대화 모델은 효과적이나 한계점이 분명
- 이미지-텍스트 멀티모달 연구시에 생성 기반의 LM의 능력을 활용하는 것이 효과적

감사합니다!

 SCATTER LAB

이재훈 (ML Research)

Email. [jaehoon@scatterlab.co.kr](mailto:jaehoon@scatterlab.co.kr)

Linked in. [@jaehoon0128](#)

Q&A