

착시 현상을 이용한 이미지 CAPTCHA 프레임워크 제시

목차 Table of contents



1 서론

5 실험 결과

2 이론적 배경

6 결론

3 연구 방법

4 실험 설정



Part 1

서론

연구배경

여론조작 도마 오른 '매크로'...서버장애 없으면 처벌도 어려워

입력 2023-10-05 15:48:46 수정 2023.10.05 15:48:46 이승령 기자

매크로로 마스크 싹쓸이, 한쪽에선 죽음의 배송

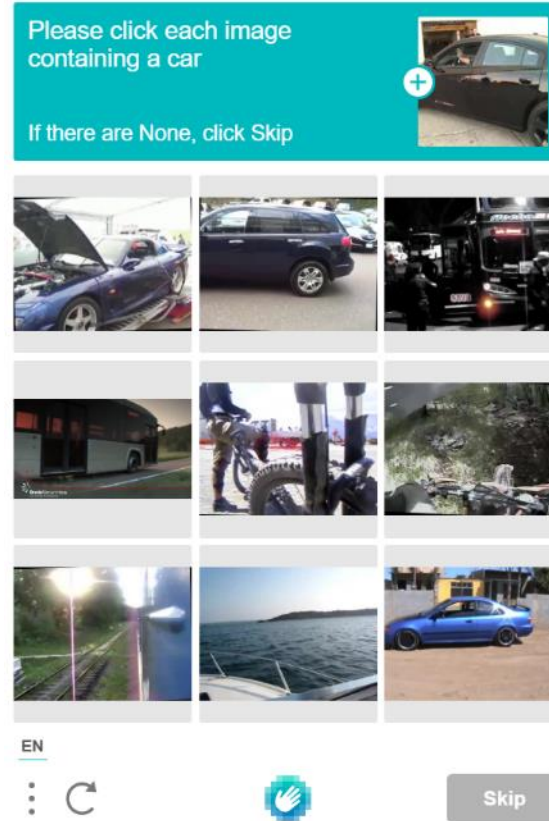
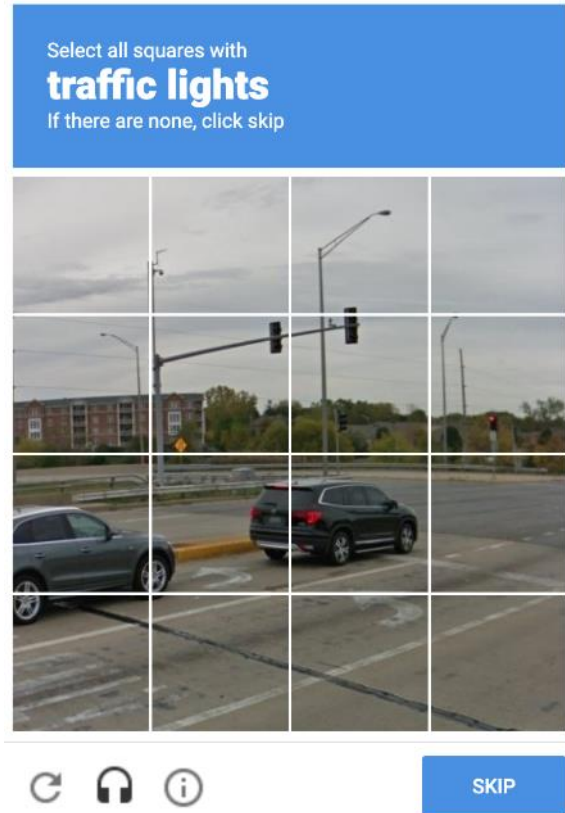
✎ 이정호 기자 | ⌚ 입력 2020.03.18 08:01 | ⌚ 수정 2020.03.18 08:29 | 💬 댓글 4

[홈 >](#) [사회 >](#)

[대한민국은 매크로와 전쟁중] '매크로 악용' 처벌 힘들어...기업이 직접 계정 적발해 제재

입력 2018-12-18 17:13:36 수정 2018.12.18 20:39:45 이지윤 기자

➔ 자동화 프로그램의 등장으로 인한 막대한 피해 발생



자동등록방지 영숫자를 순서대로 입력하세요.



웹상의 자동화 프로그램을 차단하기 위한
다양한 CAPTCHA 테스트의 개발

연구배경



OpenAI















Bard



YOLOv8

1st CAPTCHA
fastest & cheap

CaptchaAI

Supported Captcha Types & Pricing			
reCAPTCHA Token V2	reCAPTCHA Token V3	reCAPTCHA V2 Enterprise	reCAPTCHA V3 Enterprise
 reCAPTCHA	 reCAPTCHA	 reCAPTCHA	 reCAPTCHA
\$0.55 /1000 tokens	\$0.55 /1000 tokens	\$0.55 /1000 tokens	\$0.7 /1000 tokens
Speed Accuracy	Speed Accuracy	Speed Accuracy	Speed Accuracy
18 seconds 99%	2 seconds 99.8%	16 seconds 99%	4 seconds 99%
reCAPTCHA Recognition	FunCAPTCHA Token	FunCAPTCHA Outlook	FunCAPTCHA Twitter
 reCAPTCHA	 FunCAPTCHA	 FunCAPTCHA	 FunCAPTCHA
\$0.18 /1000 images	\$3 /1000 tokens	\$2.5 /1000 tokens	\$2.5 /1000 tokens
Speed Accuracy	Speed Accuracy	Speed Accuracy	Speed Accuracy
0.5 second 99%	10-20 seconds 100%	1 second 100%	1 second 100%
FunCAPTCHA Recognition	Image to Text	hCAPTCHA Token	hCAPTCHA Recognition
 FunCAPTCHA	 Image to Text	 hCAPTCHA	 hCAPTCHA
\$0.5 /1000 images	W6 8HP \$0.4 /1000 images	\$0.7 /1000 tokens	\$0.015 /1000 images
Speed Accuracy	Speed Accuracy	Speed Accuracy	Speed Accuracy
Coming soon 98%	1 second 95%	8 seconds 100%	50 ms 99%

성능이 뛰어난 Vision AI 및 Multimodal AI의 발전
& 이를 이용한 CAPTCHA 파훼 서비스의 개발

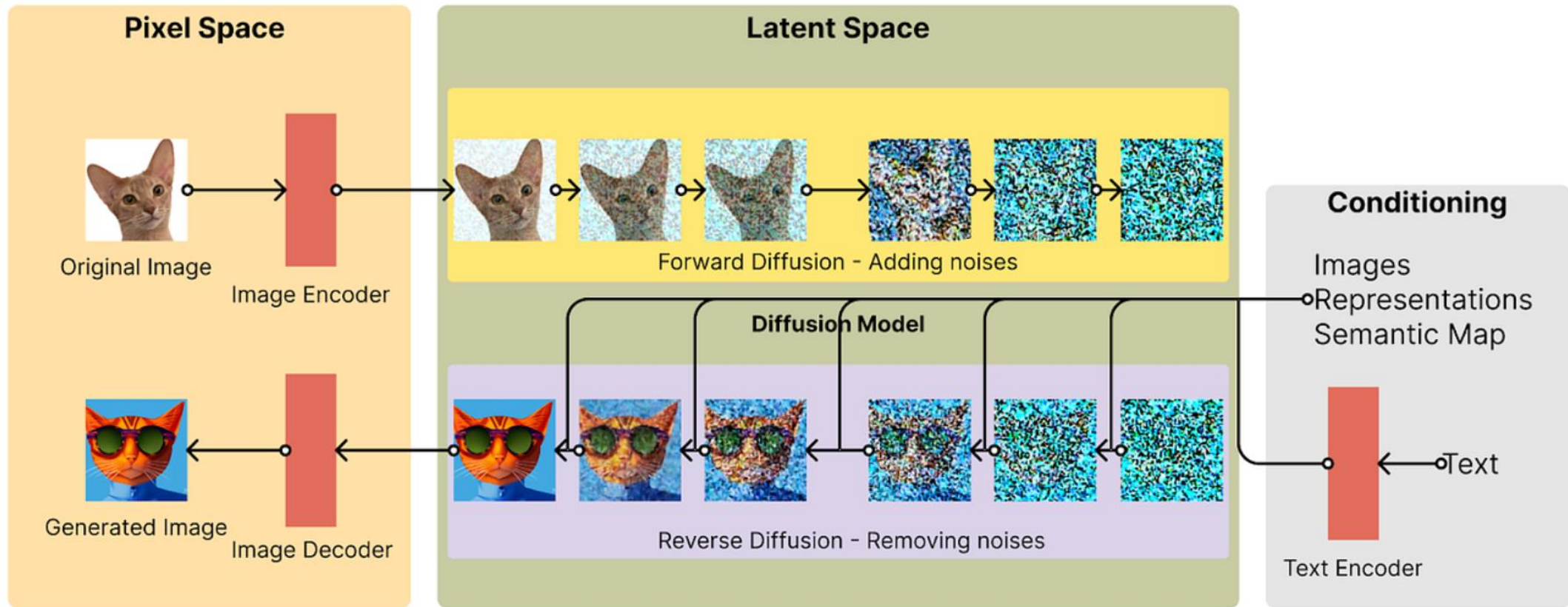
→ 기존 CAPTCHA 서비스의 효과 상실



Part 2

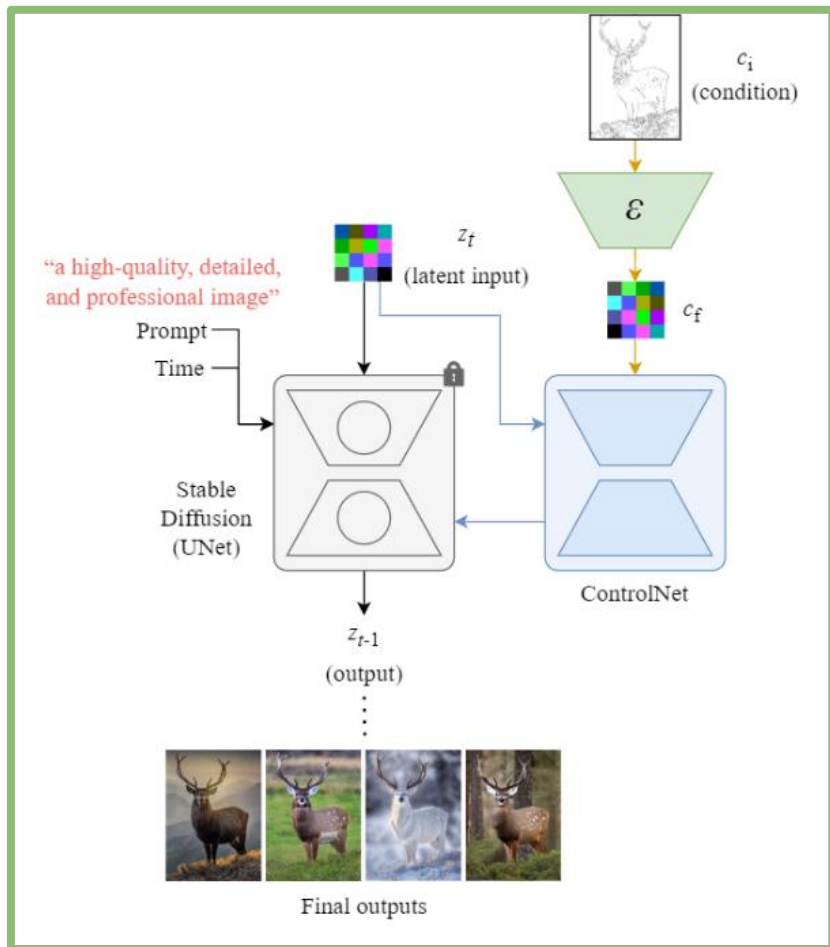
이론적 배경

Stable Diffusion



Latent Diffusion 모델

ControlNet



다양한 종류의 레퍼런스 이미지 사용 가능
 → 착시 현상 유발하는 ControlNet 모델 사용

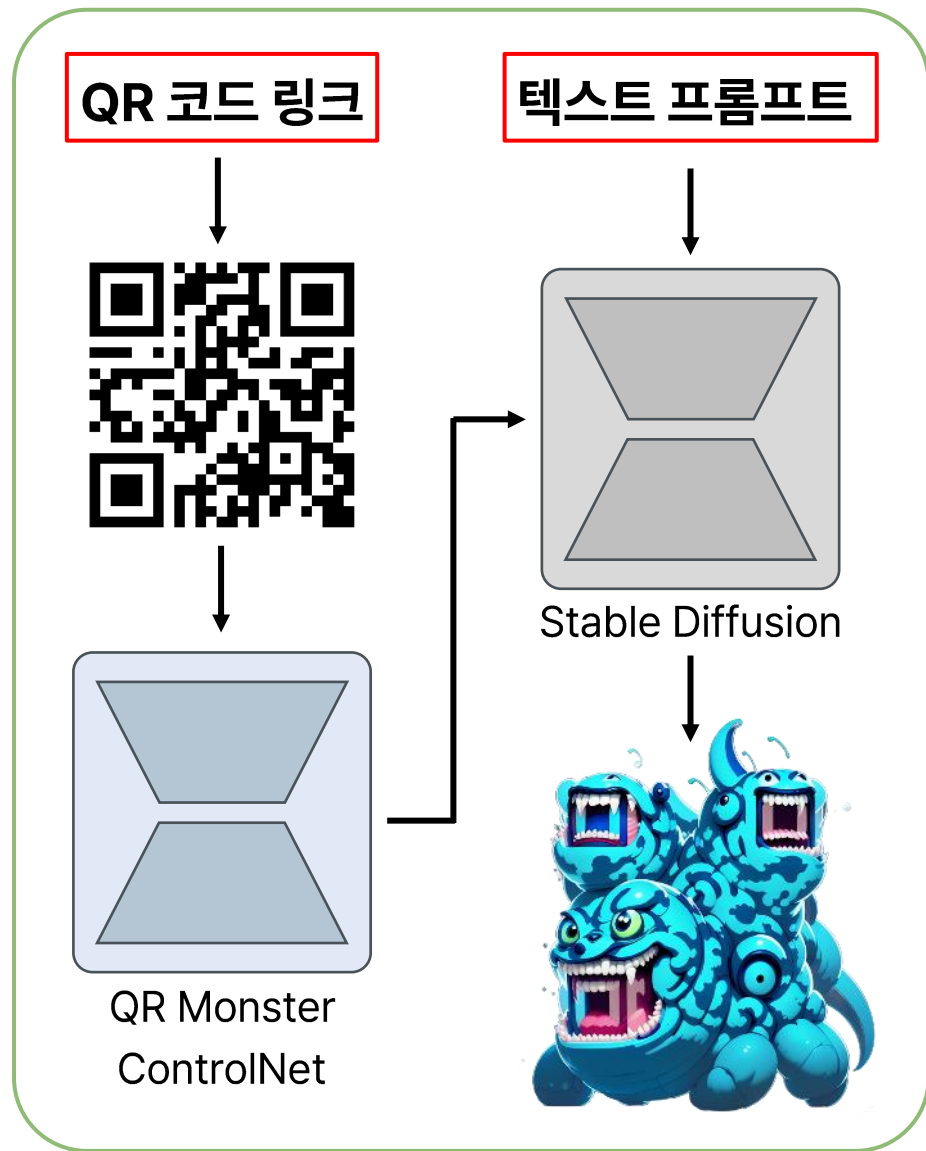


Part 3

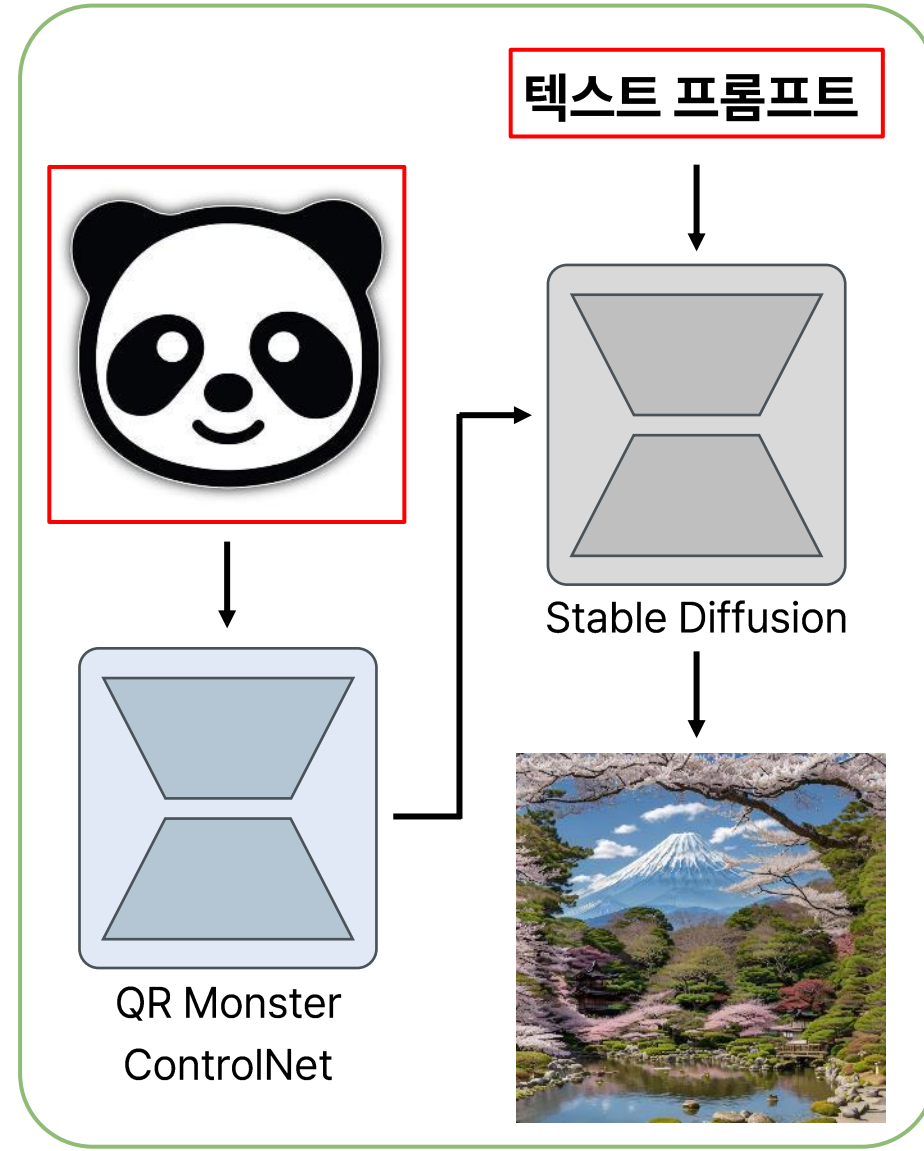
연구 방법

착시 현상 이미지 생성 방법

QR Monster ControlNet



Illusion Diffusion



레퍼런스 이미지 및 텍스트 프롬프트 생성 방법



→ 단색클립아트 형식의 이미지를 사용해
Stable Diffusion 모델 Fine-Tuning 진행

텍스트 프롬프트 생성

"Generate a random sentence describing a scenery containing several of the following elements:

- Background Scene
- Objects in the scene (people, trees, etc)
- Time period
- Weather
- Overall structure"

레퍼런스 이미지의 주제

"Suggest a random object that doesn't have a complicated outline. (ex. Panda, Tree, House, etc)
Only output the object without any explanation.

(Repeat 10 times)"

gpt-3-turbo의 API를 사용하여 텍스트 생성

생성 이미지 검증



What is in front of the building?

Compute

Computation time on cpu: 0.163 s

flowers

0.665

fountain

0.087

bench

0.078

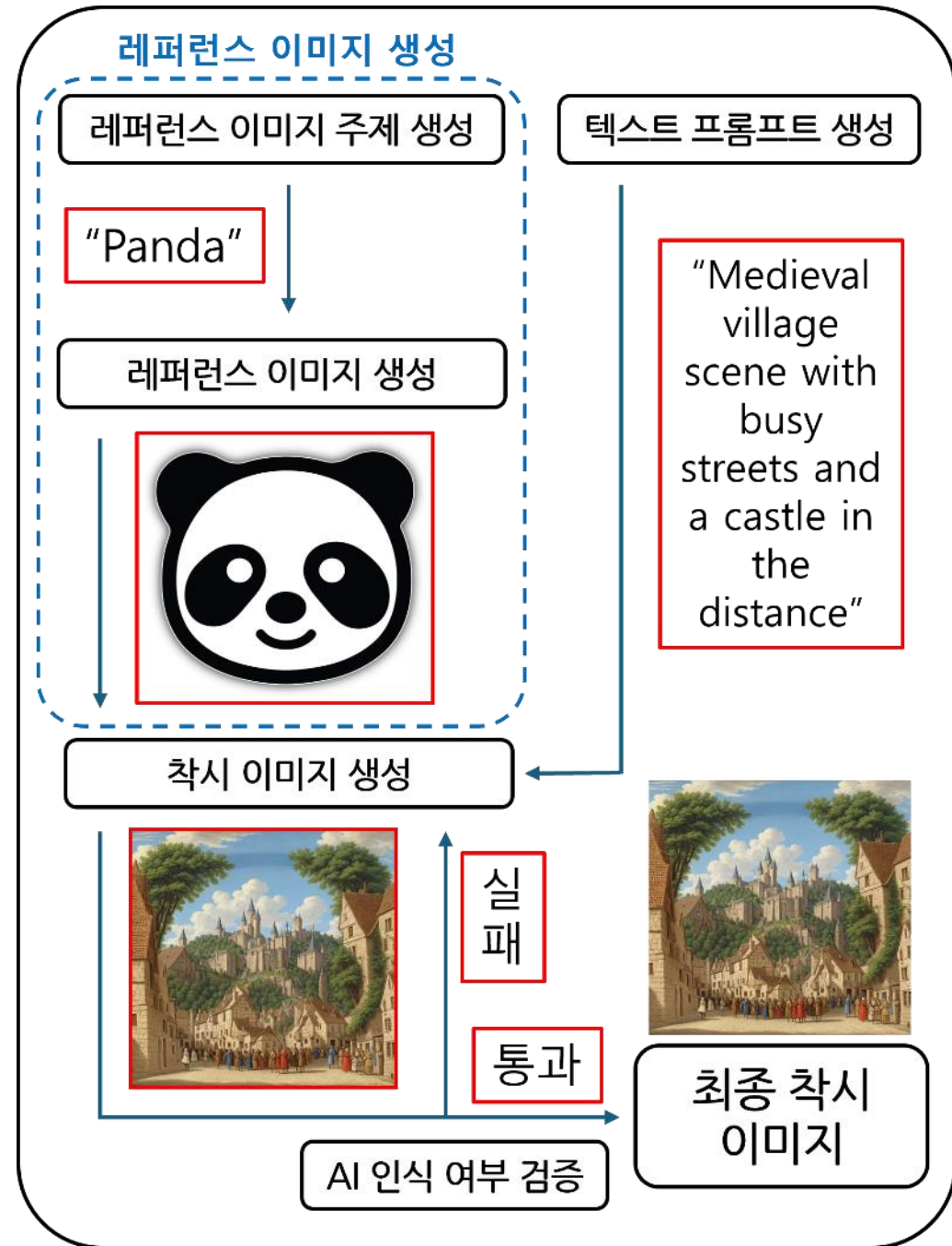
park

0.068

bushes

0.059

**VQA v2.0 데이터셋으로 학습된
ViLT-B/32 모델 사용**



캡차 챌린지 구성 및 성능 평가

Which of the following can be seen in the image?



- ☐ Apple
- ☐ Panda
- ☐ Key
- ☐ Hat
- ☐ Car

None

착시캡차챌린지의 구성

$$F_1 \text{ Score} = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	False Positive (FP) Type I Error
	Positive +	False Negative (FN) Type II Error	True Positive (TP)

AI의 착시캡차파훼 성능 평가



Part 4

실험 설정

Illusion Diffusion 파라미터 설정

Negative Prompt

worst quality, normal quality, low quality, low res, blurry, text, watermark, logo, banner, jpeg artifacts, signature, username, error, monochrome, horror, mutation, disgusting

Stable Diffusion, ControlNet 파라미터

- Realistic_Vision_V5.1_noVAE
- 이미지 크기 512*512
- Guidance, Conditioning scale: 7.5, 1

실험 데이터

- 50개의 착시 캡차 챌린지 데이터
- 제외된 데이터 없음

실험 평가

- ChatGPT-4 (GPT-4) 및 Google Bard (PaLM) 모델 사용하여 검증
- 정량적 평가 위한 F_1 Score 척도 도입

실험 데이터



착시캡차챌린지데이터셋일부

착시 유발 요소



착시캡차챌린지데이터셋의 일부



Part 5

실험 결과

AI의 착시 캡차 파훼 실험 결과



(A,B)



(B)

ChatGPT-4 (A) 및 Bard (B)의 착시 캡차 파훼 성공 사례

검증 실패
이미지의
재생성 과정

ViLT 검증 결과

```
Answer: no, Confidence: 0.6882
Answer: yes, Confidence: 0.3099
Answer: unknown, Confidence: 0.0004
Answer: not sure, Confidence: 0.0002
Answer: can't tell, Confidence: 0.0002
```



AI의 착시 캡차 파훼 성능에 대한 정량적 평가



텍스트 프롬프트,
틀린 답안의 중복

- 틀린 경우, 대부분 정답 없음 옵션 선택
→ 정답을 맞힌 경우 우연에 의한 것이 아닌, 확신을 가지고 판단하였다는 것을 확인함
- AI의 오판은 예외적인 데이터를 제외하고 발견되지 않음

실험결과분석

ChatGPT-4 및 Bard의 착시 캡차 파훼 성능

	ChatGPT-4 (GPT-4)	Bard (PaLM)
F1 Score	0.4	0.2857
Success Rate	4%	2%

→ 파훼가 불가능한 수준



Part 6

결론

- 발전하는 AI를 활용한 캡차의 파훼가 심각한 보안 문제임을 파악
- 착시 현상을 이용한 캡차의 개발 방법을 제시
- Stable Diffusion 및 DreamBooth, ControlNet 등을 사용해 착시 캡차 챌린지 개발
- 현존하는 Multimodal AI 모델 중 가장 성능이 좋은 GPT-4 및 PaLM을 사용하여 실험을 진행
- AI는 착시 캡차 챌린지를 파훼 하지 못한다는 사실 입증
- 사람은 큰 문제 없이 해결 가능함을 확인함
- 추후 Universal Adversarial Perturbation 등을 통해 착시 캡차 챌린지의 보안 강화 가능
- 착시 캡차 API를 제공하여 자동화 프로그램에 의한 피해 최소화 도모