



Language-Image Multi-modal AI 기술 연구

이정우 장서윤 장성현

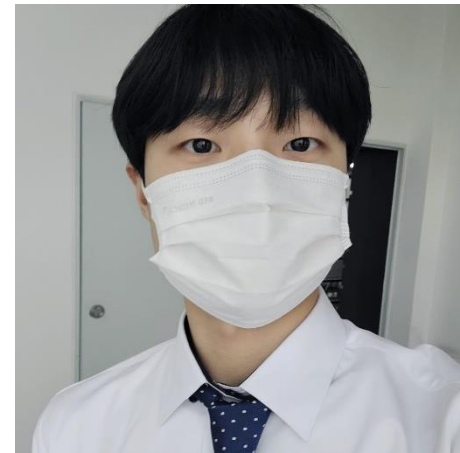
팀원 소개



이정우
성균관대 소프트융합대학
인공지능융합학과
석사과정



장서윤
AI빅데이터융합경영학과/
데이터사이언스연계융합전공



장성현
AI빅데이터융합경영학과/
데이터사이언스연계융합전공

CONTENTS

01 연구 목표

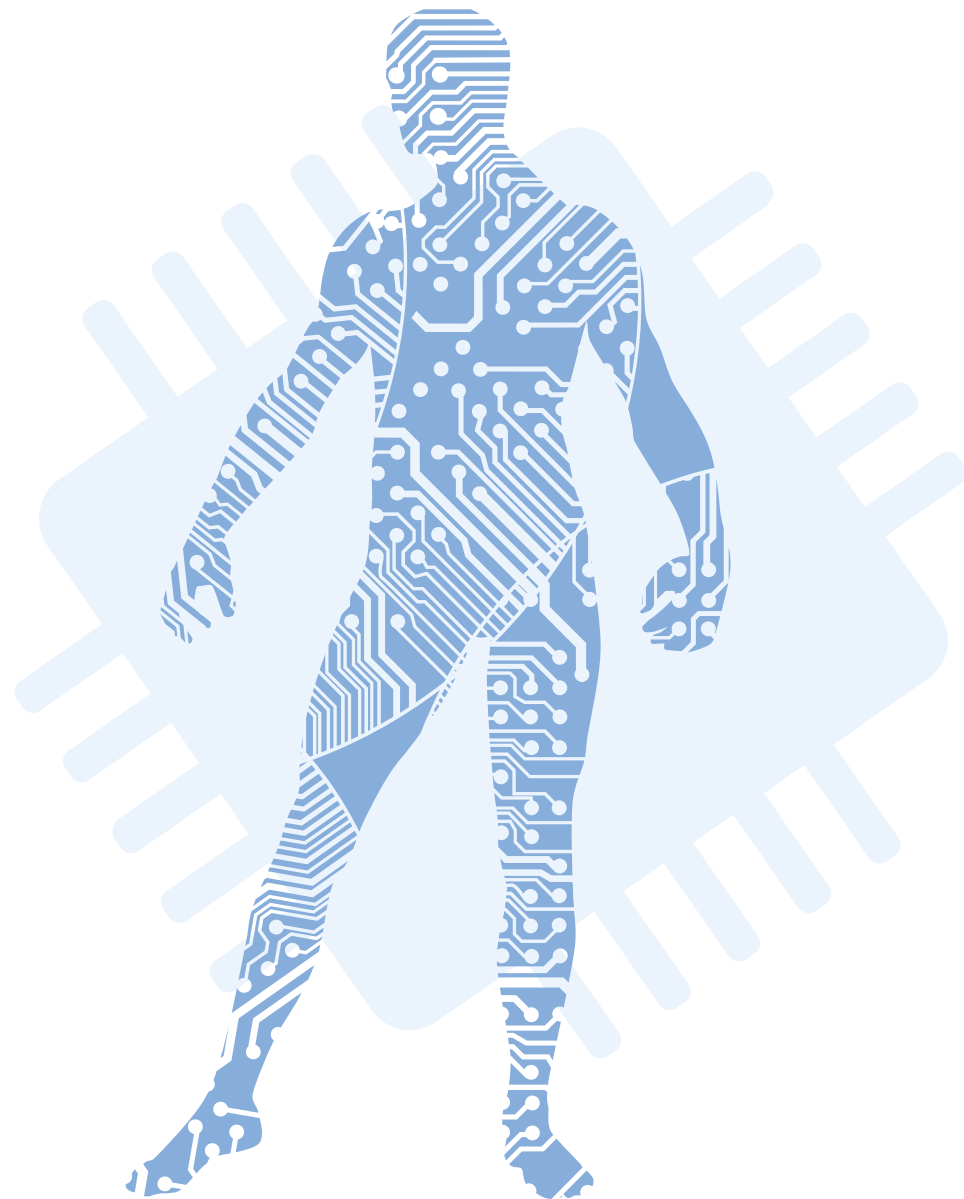
최신 연구, 새로운 접근, 기존 한계점 보완, 성능 향상

02 연구 배경

Vision Language 분야의 발전 방향
Flamingo(04,2022)

03 개선점 및 아이디어

Prompted learning
TCL
Data2vec



The background is a solid blue color. In the center, there is a faint, semi-transparent image of a person's head and hand. The head is shown in profile, facing right, and is composed of a network of white dots connected by thin white lines, resembling a digital or neural network structure. The hand is visible at the bottom, holding a tablet device. The overall theme is artificial intelligence and research.

AI

01. 연구 목표

01. 연구 목표

Vision-Language
Multi-modal AI



성능 향상
성능 개선 및 향상



기존 한계점 보완
기존 방법론과 VL 분야에 존재하는
한계점 개선



새로운 접근
최신 동향을 반영하면서도 새로운
접근법 제시



최신 연구
Vision-Language 분야의 최신
동향을 반영하는 연구

The background is a solid blue color. In the center, there is a faint, semi-transparent graphic of a human head in profile, facing right. The head is composed of a network of white dots connected by thin white lines, resembling a neural network or a globe. A hand is visible at the bottom, holding a tablet that displays a blue screen. The overall theme is artificial intelligence and research.

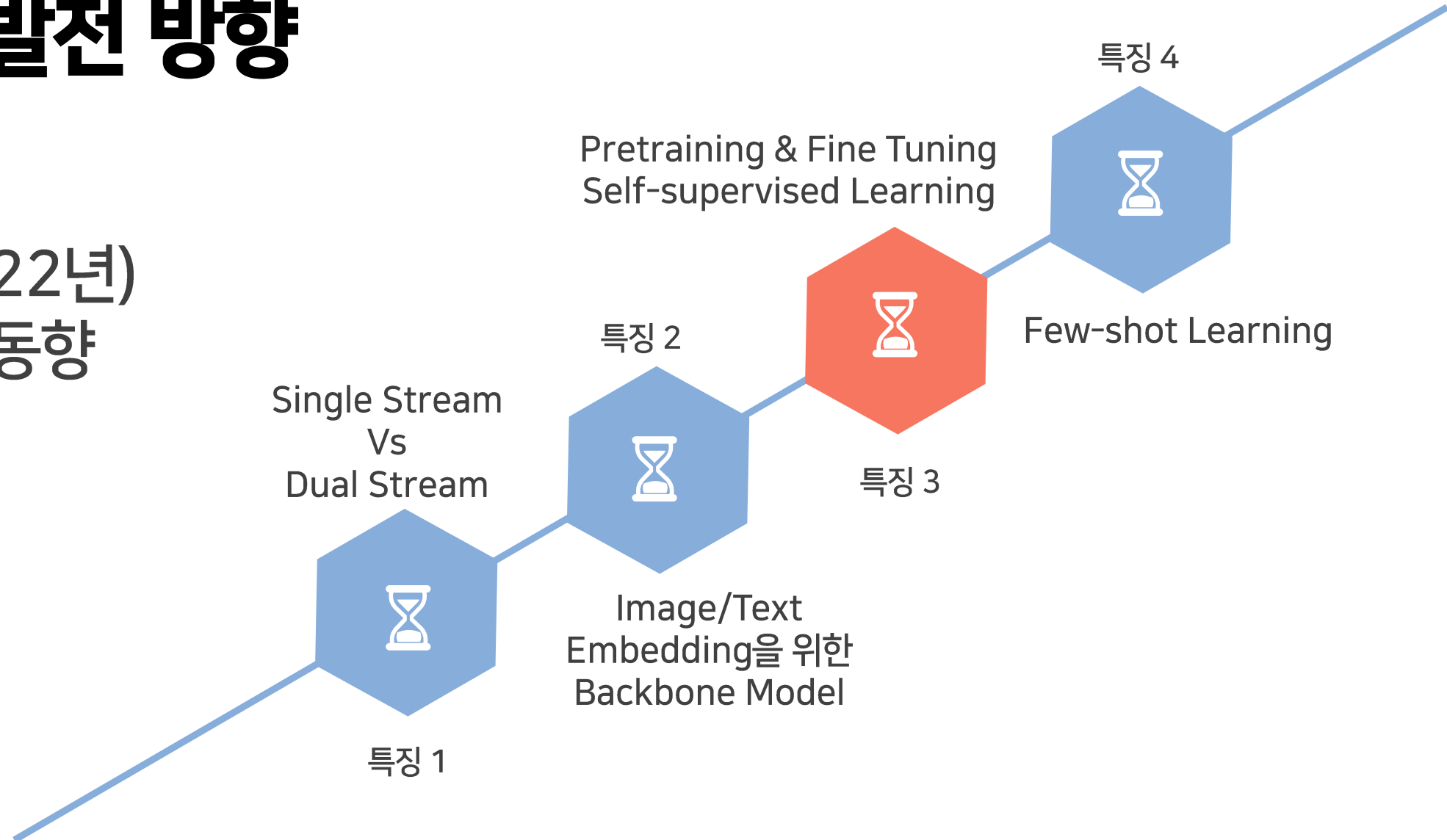
AI

02. 연구 배경

Vision-Language

분야의 발전 방향

(2020~2022년)
최근 연구 동향

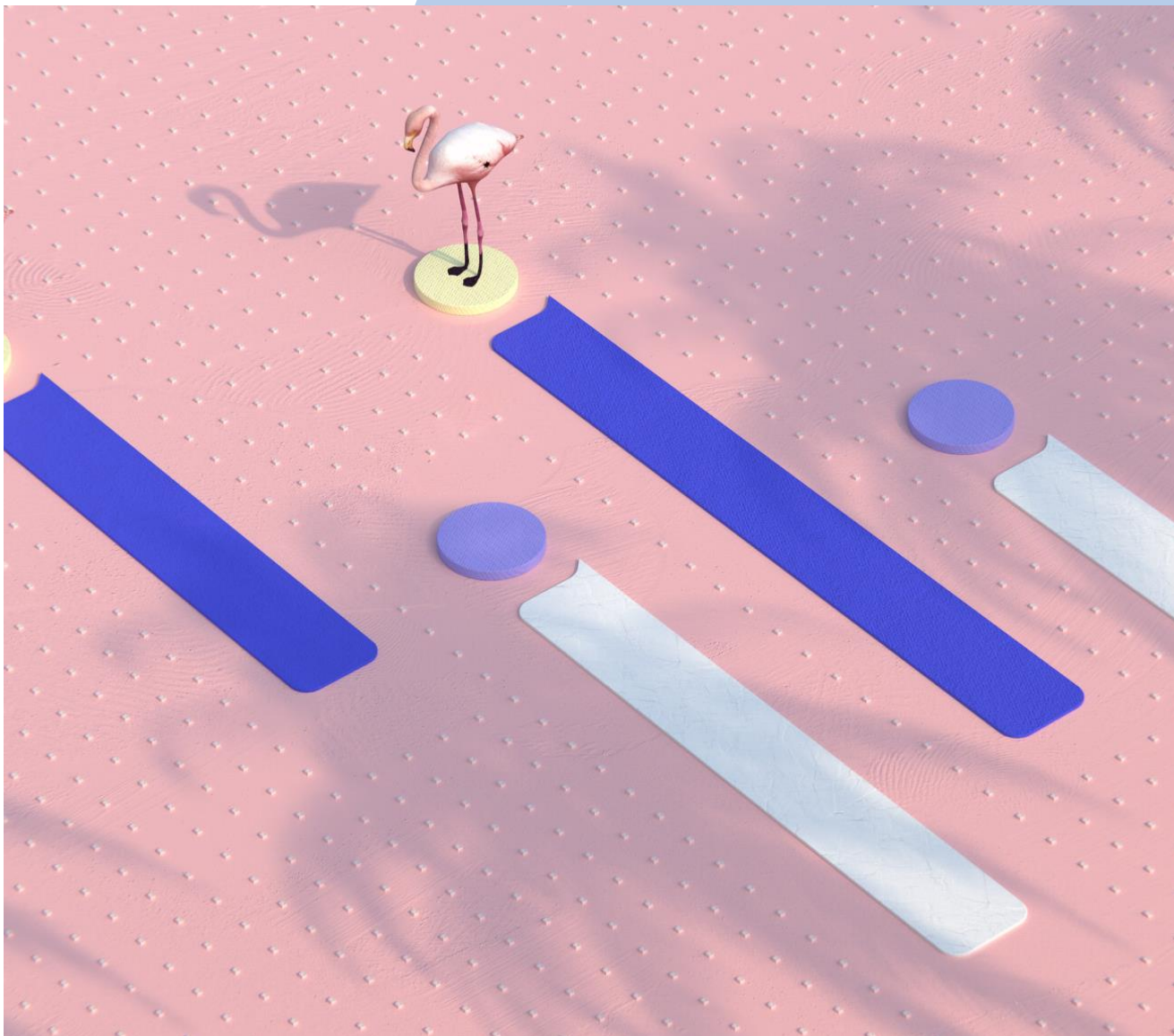


Flagminogo

A Visual Language for Few-Shot Learning



















Contributions

1. 임의로 **interleaved**된 이미지와 텍스트 데이터를 입력으로 받아들이고 **open-ended 방식**으로 출력 텍스트를 생성하기 위한 새로운 아키텍처
2. **Pretrained** 된 대규모 **vision-only models**와 **Language-only models**를 활용하여 초기 모델의 이점을 보존하면서도 효율적으로 융합하는 전략
3. 다양한 크기의 **visual** 입력에 효율적으로 적응하여 이미지 뿐만 아니라 **비디오**에도 적용 가능



Flagminogo

Input Prompt

	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.		What is the name of the city where this was painted? Answer:
	Output: "Underground"		Output: "Congress"		Output:
	2+1=3		5+6=11		
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.		Output:
	Les sanglots longs des violons de l'automne blessent mon cœur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?		

Completion

a flamingo. They are found in the Caribbean and South America.

Arles.

"Soulomes"

3x6=18

A portrait of Salvador Dali with a robot head.

Je suis un cœur qui bat pour vous.



This is a picture of a dog. It looks like it is a little bit naughty.

What happened here?

It looks like the dog is destroying the sofa.

How does the person thinking about it?

The person is angry.

What is the person holding?

It looks like a cup of coffee.

How is the weather outside?

It looks like a sunny day.



What do you think the robot is doing here?

The robot is picking up a ball.

What color is the ball?

It's red.

I want the ball to be hidden. Is there something the robot can use for that?

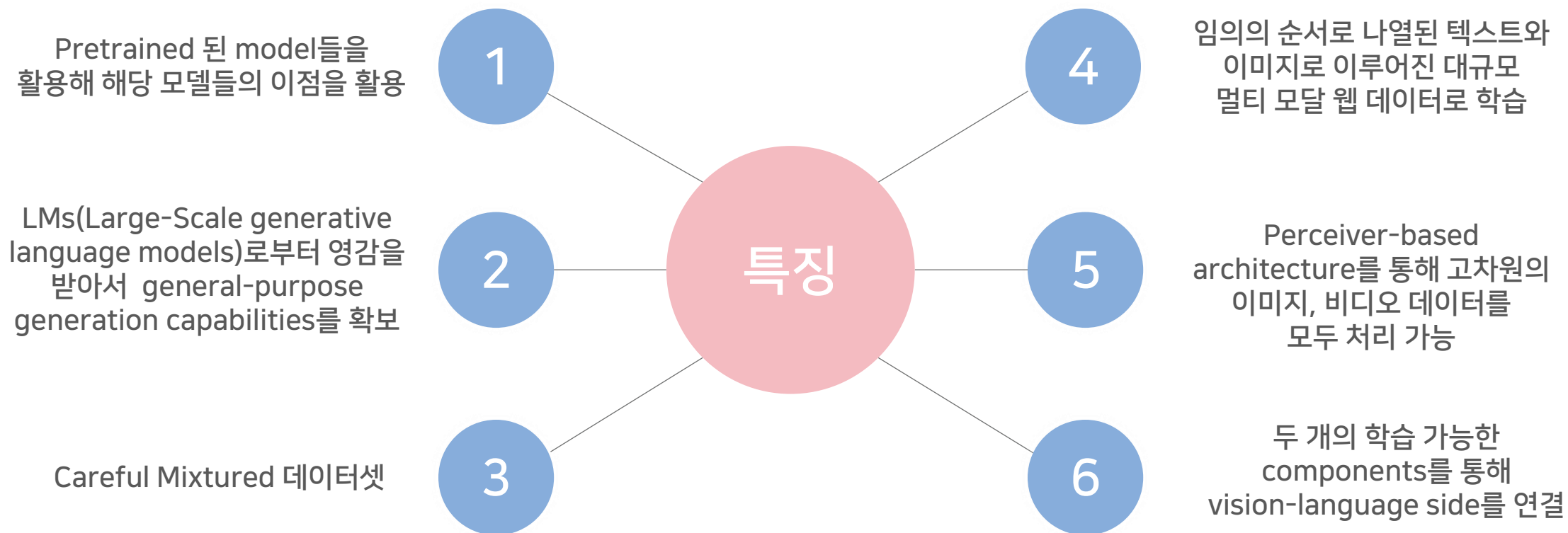
The robot can use a cloth to hide the ball.

What color is the cloth?

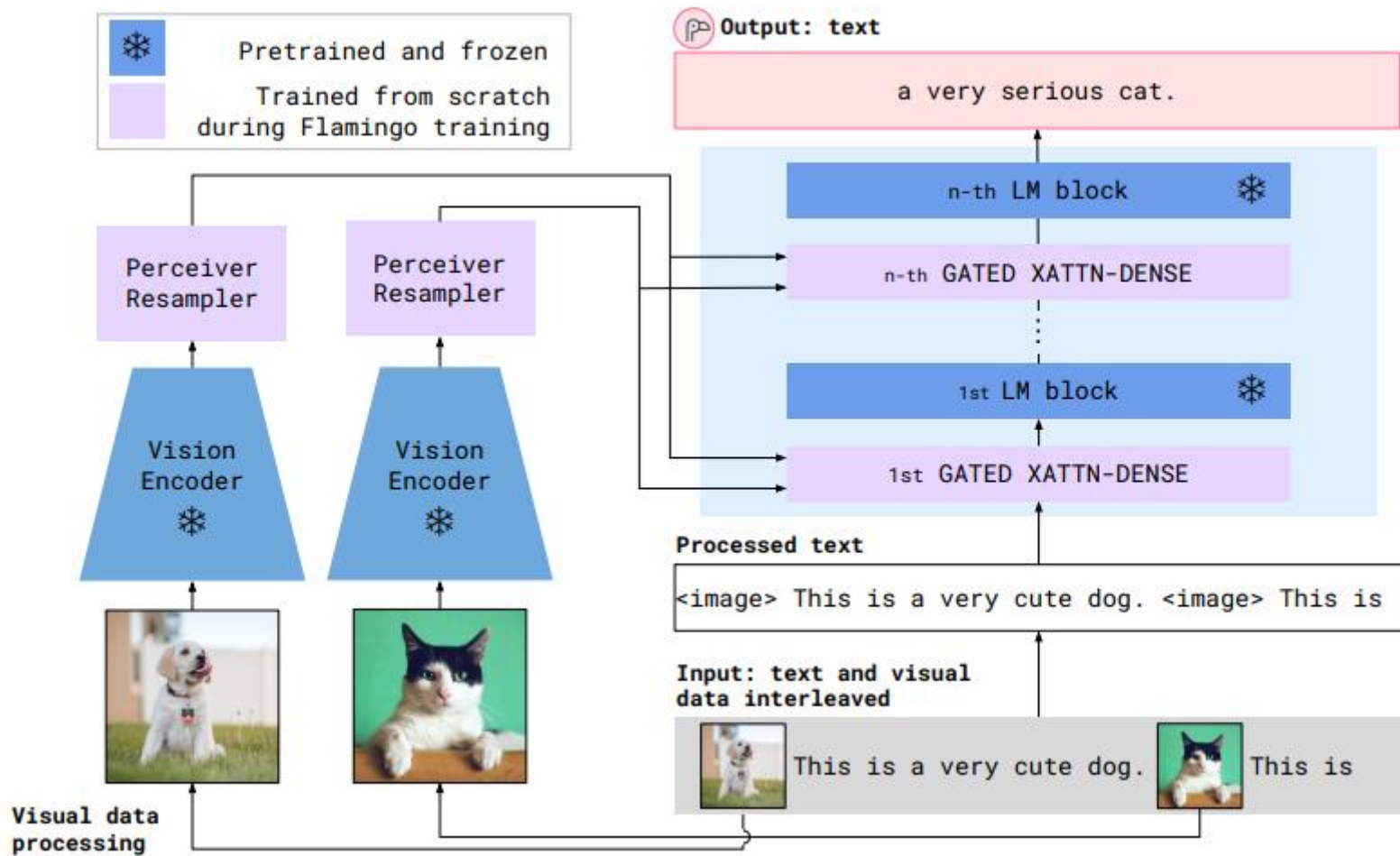
The cloth is blue.

▲ Selected dialogue samples

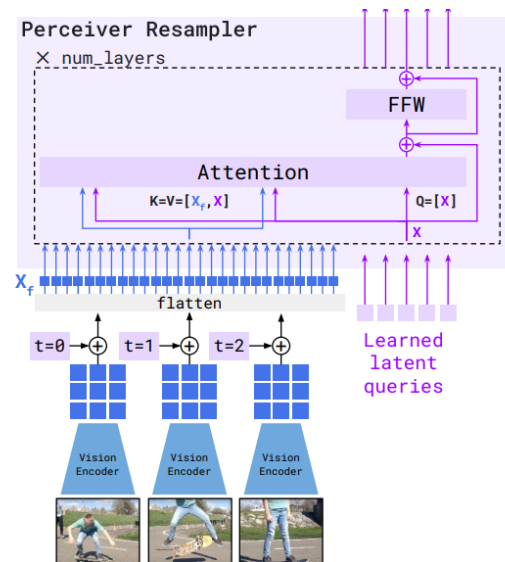
Flagminogo



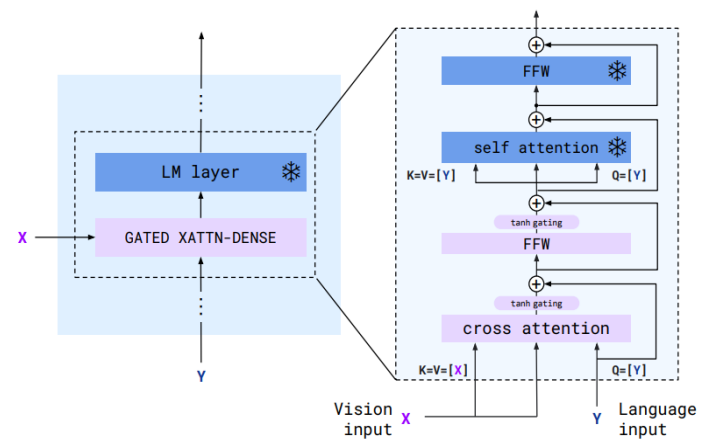
Architectures



▲ Overall Flow



▲ Vision Side Perceiver Resampler



▲ Language Side gated xattn -dense layers

The background of the slide features a person's head in profile, facing right. The head is composed of a wireframe mesh of white lines and dots, giving it a digital or AI-like appearance. The person is holding a tablet computer, which is visible at the bottom of the frame. The entire scene is set against a solid blue background.

AI

03. 개선점 및 아이디어

New Approach: P-tuning

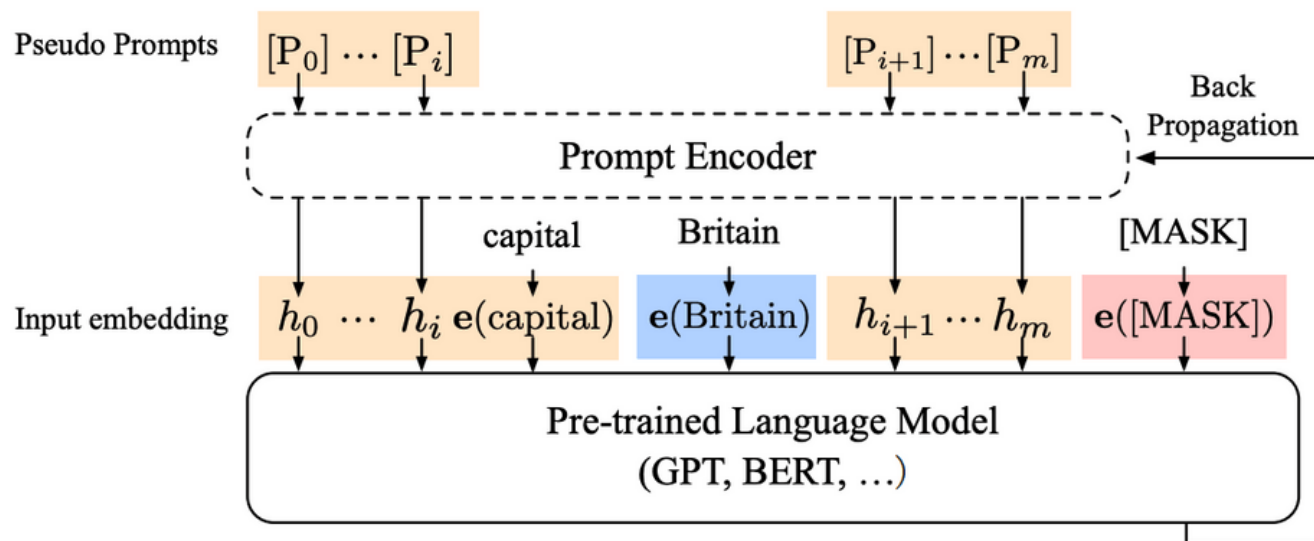
매번 좋은 프롬프트를
바로바로 찾거나 직접 입력하는 것은
현실적으로 어려움

Automatically
searching
prompts

Only tunes
continuous
prompts

연속적인 공간에서 prompt-based learning을 수행하여
prompt 설계에 따라 성능이 불안정한 것을 개선

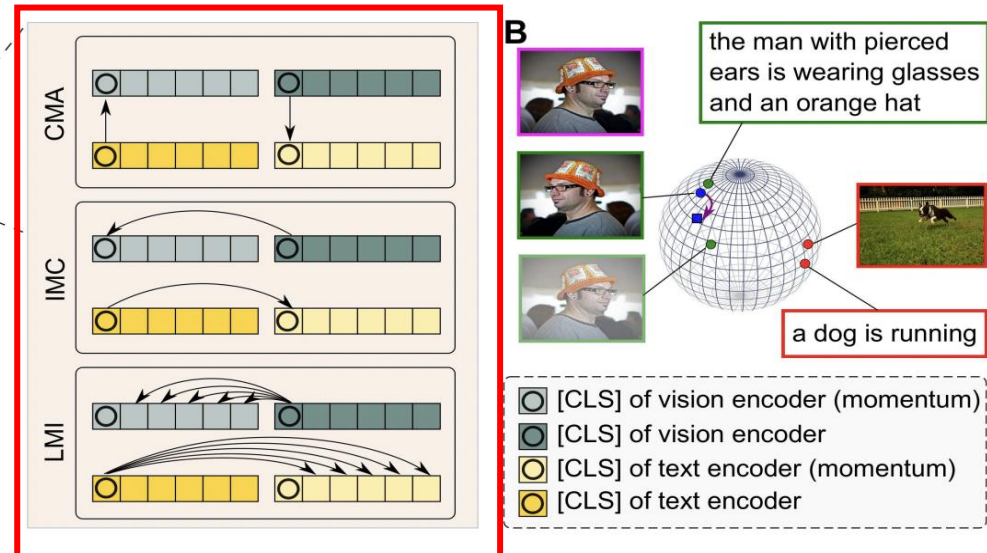
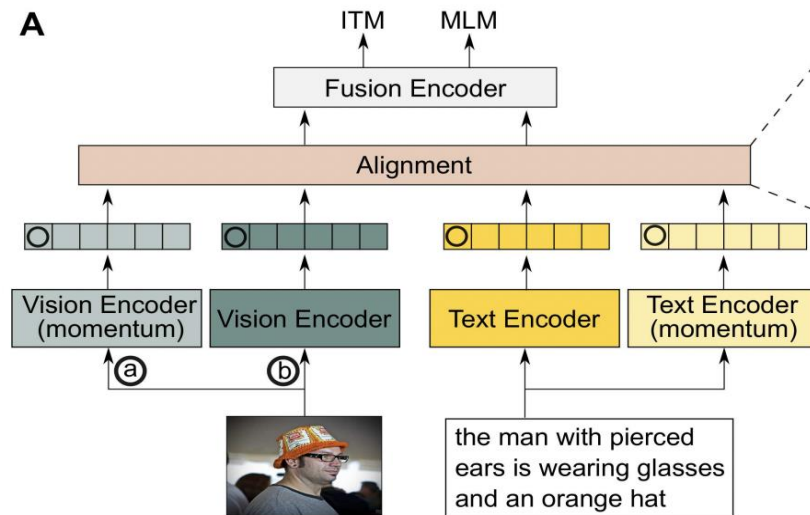
“ 소량의 프롬프트로 미리 학습시킨
Prompt Encoder를 추가해보자 ”



New Approach: TCL

Transformer 구조 사용으로
global한 feature들을 뽑는 데에만 유리
Local, structural한 정보를 고려하는 것을 실패

“ Local적인 정보에 대한
Loss function을 추가해보자 ”



New Approach: Data2vec

하나의 architecture를 통해서
여러 도메인에 일반화된 성능으로
적용하는 방법론

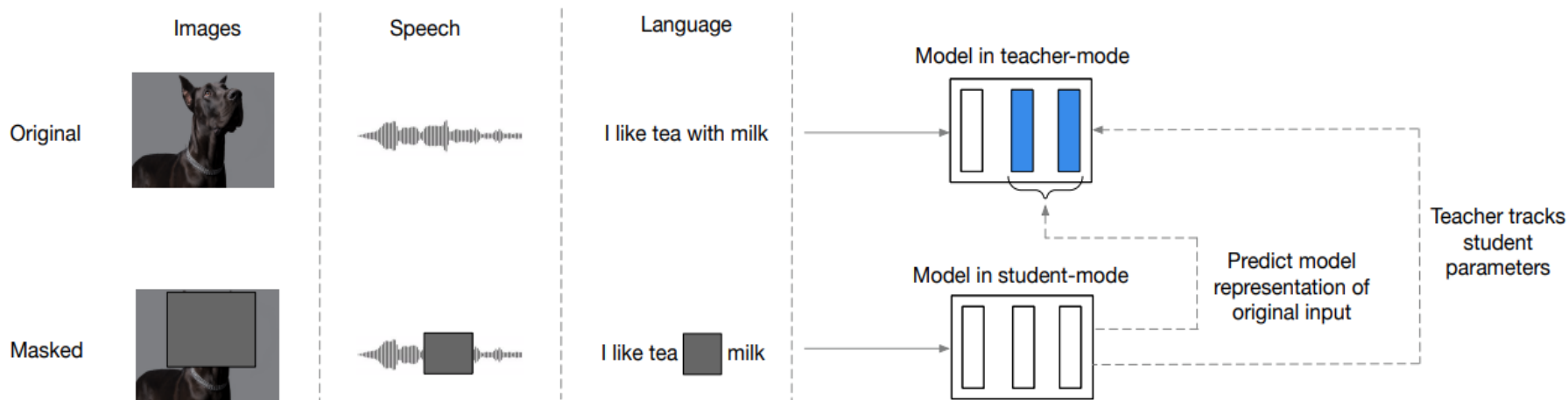
“ encoder-decoder 구조를
Teacher-Student 구조로 대체해보자 ”

A General Framework for Self-supervised Learning
in Speech, Vision and Language

Masked Input with
Teacher-Student
self-distillation

Contextualized
Representation
Learning

data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language



참고 자료

Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Vision-Language Pre-Training with Triple Contrastive Learning. 2022.


data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. 2022.

GPT Understands, Too. 2021.

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. 2022.

CoCa: Contrastive Captioners are Image-Text Foundation Models. 2022.

PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining. 2022.



METER: An Empirical Study of Training End-to-End Vision-and-Language Transformers. 2021.

VLMO: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. 2021.

MURAL: Multimodal, Multitask Retrieval Across Languages. 2021.

SimVLM: SIMPLE VISUAL LANGUAGE MODEL PRETRAINING WITH WEAK SUPERVISION. 2021.


Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. 2021.

CLIP-ViL: How Much Can CLIP Benefit Vision-and-Language Tasks?. 2021.

Probing Inter-modality: Visual Parsing with Self-Attention for Vision-Language Pre-training. 2021.

E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. 2021.

SOHO: End-to-End Pre-training for Vision-Language Representation Learning. 2021.



MDETR : Modulated Detection for End-to-End Multi-Modal Understanding. 2021.

Kaleido-BERT: Vision-Language Pre-training on Fashion Domain. 2021.

ALIGN: Scaling Up Vision-Language Representation Learning with Noisy Text Supervision. 2021.

ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. 2021.

VL-T5 : Unifying Vision-and-Language Tasks via Text Generation. 2021.

VinVL: Revisiting Visual Representations in Vision-Language Models. 2021.

RVL-BERT: Visual Relationship Detection with Visual-Linguistic Knowledge from Pre-trained Representations. 2021.

Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. 2020.

VILLA: A Generic Adversarial Training technique for Vision-and-Language. 2020.



Thank You