VLM 끄적끄적 3

Pretraining Task (2): Generative Objectives

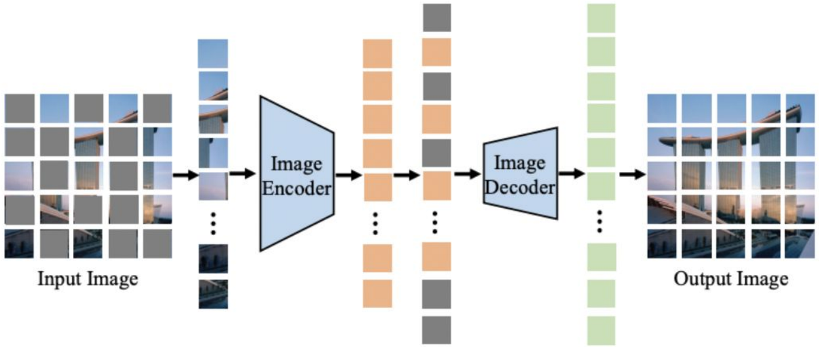
Generative Objectives의 종류

1. Masked Image Modeling (MIM)
2. Masked Language Modeling (MLM)
3. Masked Cross-modal Modeling (MCM)
4. Image-to-text Generation

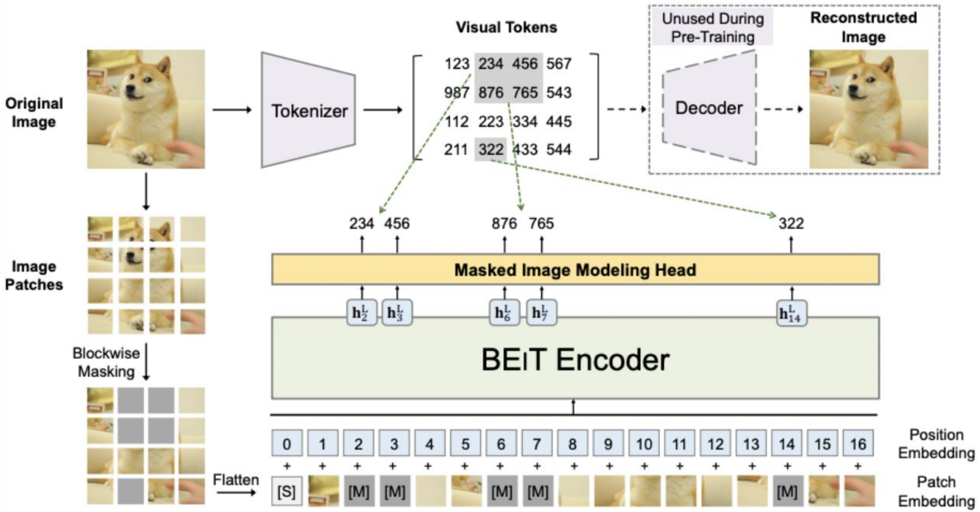
Generative Objectives의 종류

1. Masked Image Modeling (MIM)
   - FLAVA (CVPR 2022), KELIP (arxiv 2022), SegCLIP (ICML 2023)
2. Masked Language Modeling (MLM)
   - FLAVA (CVPR 2022), FIBER (NeurIPS 2022)
3. Masked Cross-modal Modeling (MCM)
   - FLAVA (CVPR 2022)
4. Image-to-text Generation
   - CoCa (arxiv 2022), NLIP (AAAI 2023), PaLI (ICLR 2023)

1. Masked Image Modeling (MIM)



MAE

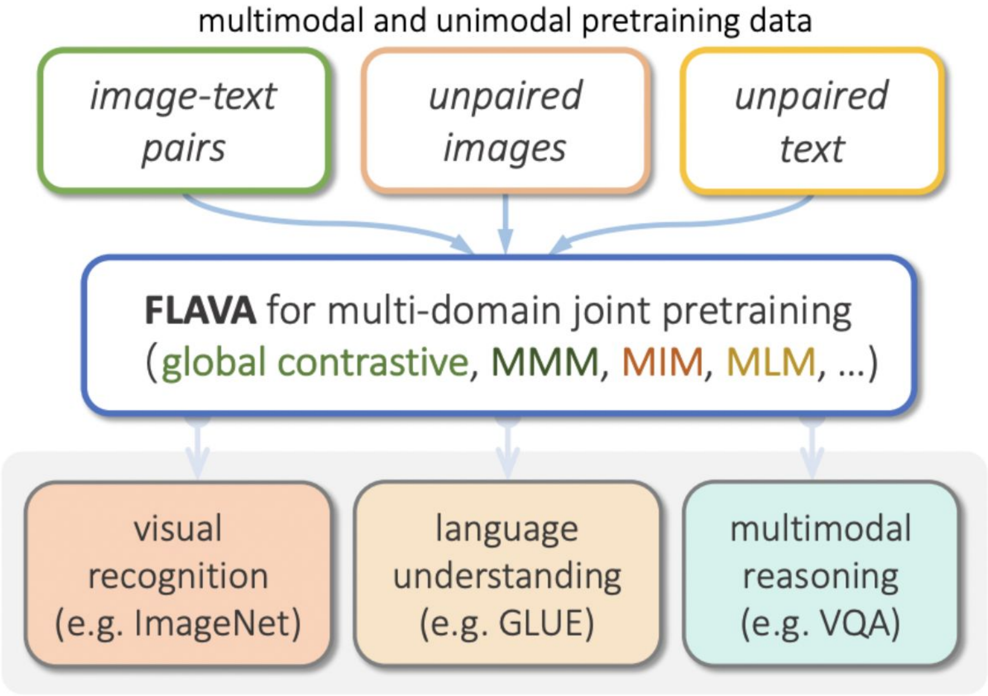


BEiT

1. Masked Image Modeling (MIM)

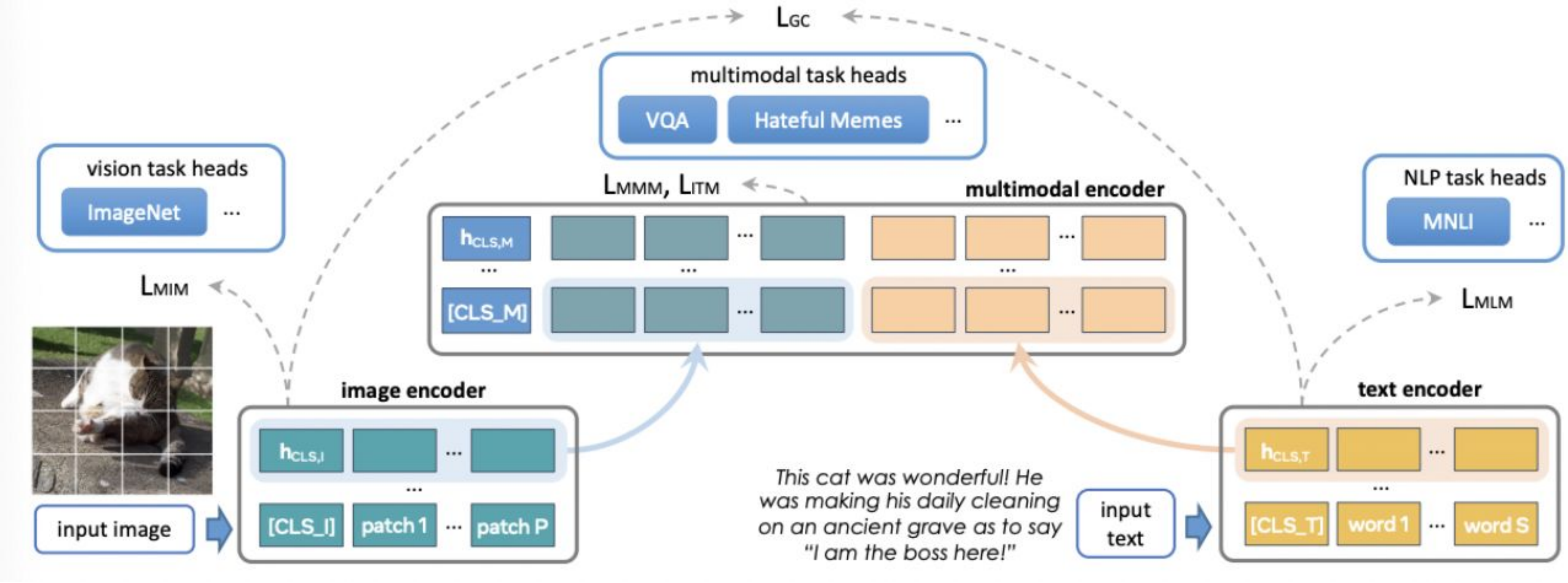
(1) FLAVA (CVPR 2022)

- Previous works: 둘 중 하나
  - Cross-modal (e.g., CL)
  - Multi-modal (e.g., fusion)
- Solution: FLAVA
  - 둘 다 하자!



1. Masked Image Modeling (MIM)

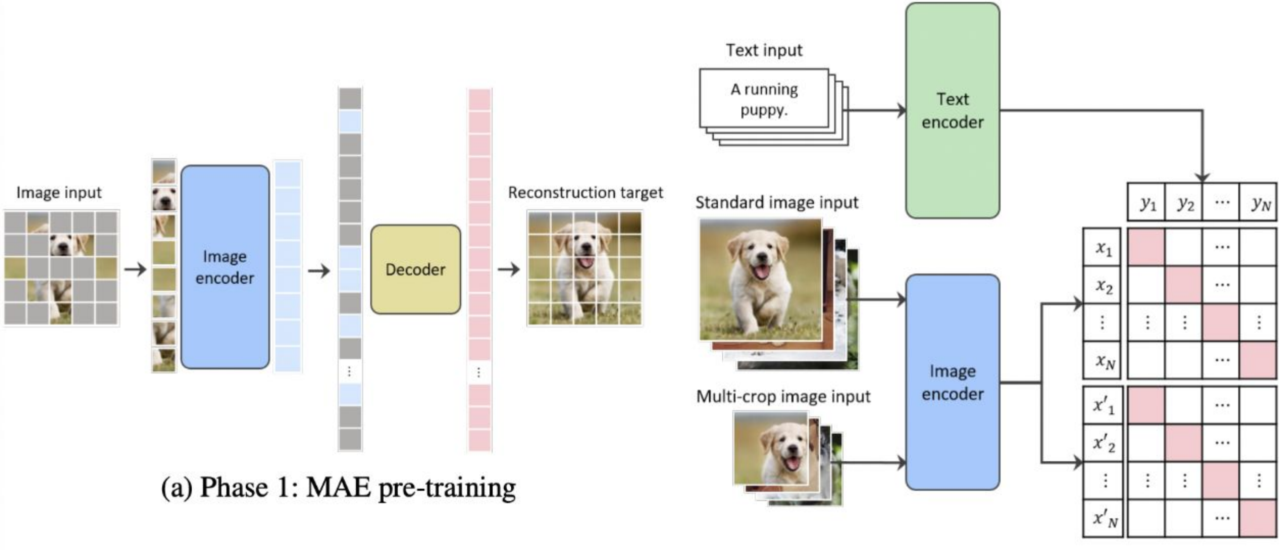
(1) FLAVA (CVPR 2022)



1. Masked Image Modeling (MIM)

(2) KELIP (arxiv 2022)

- K & E = Korean & English
- 두 가지 전략
  - MAE
  - Multi-crop



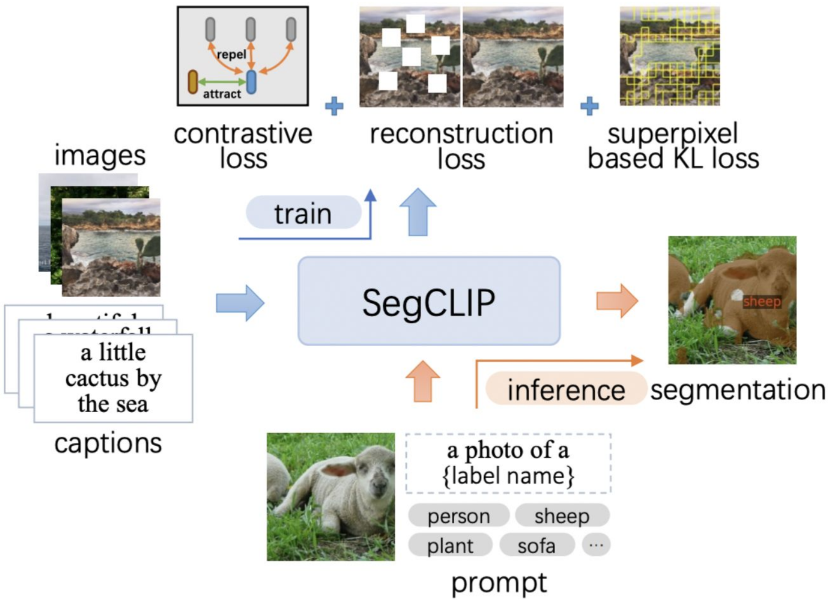
(a) Phase 1: MAE pre-training

(b) Phase 2: Multi-modal fine-tuning

1. Masked Image Modeling (MIM)

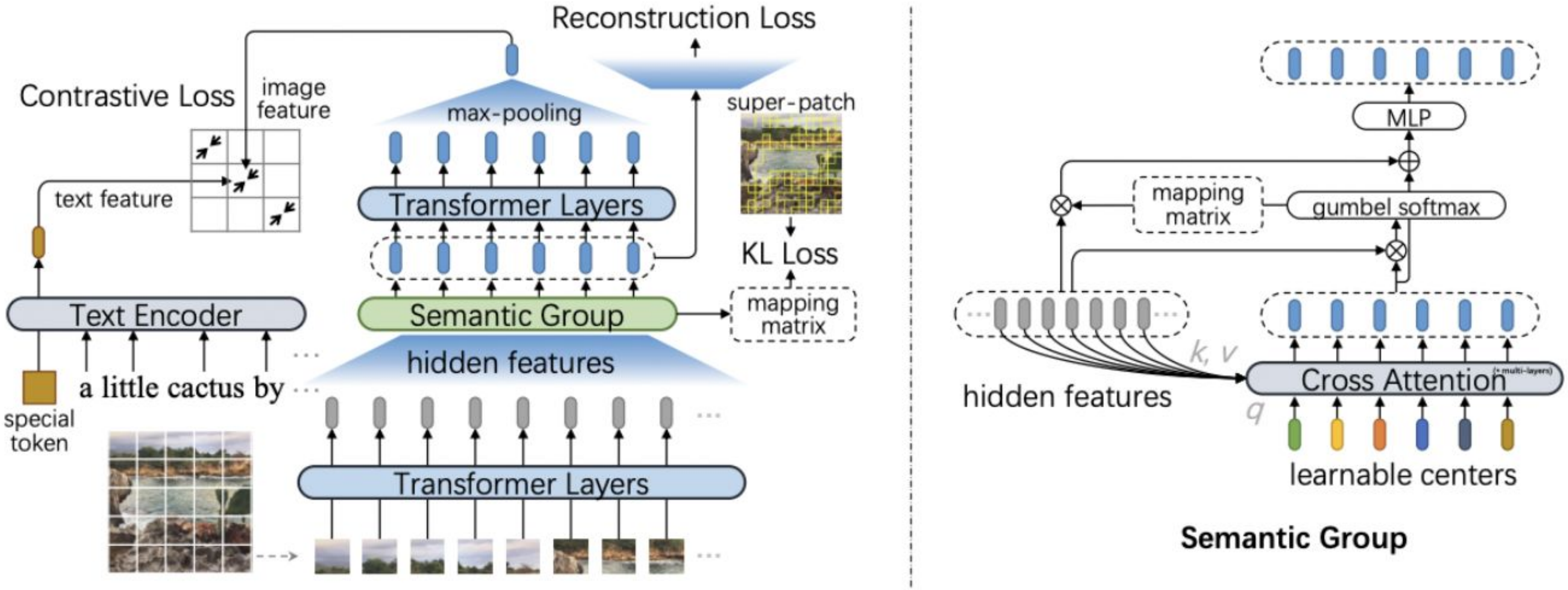
(3) SegCLIP (ICML 2023)

- Limitation of previous works?
  Open-Vocab Semantic Segmentation
  으로의 transfer는 잘 연구되지 않음!
- Proposal: Segmentation + CLIP
- 세 가지의 Loss
  - Contrastive Loss (Coarse)
  - Reconstruction Loss (Fine-grained
  - Superpixel-based KL loss



1. Masked Image Modeling (MIM)

(3) SegCLIP (ICML 2023)



1. Masked Image Modeling (MIM)

(3) SegCLIP (ICML 2023)

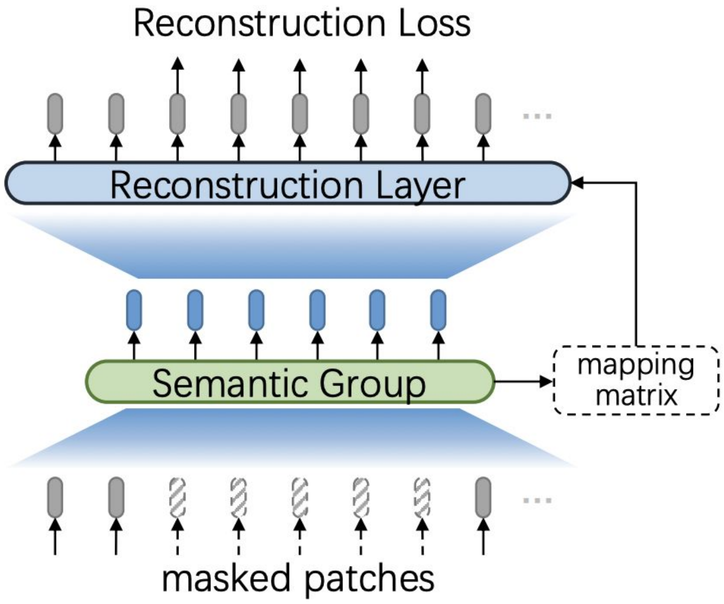


Figure 3. **Reconstruction Loss.**

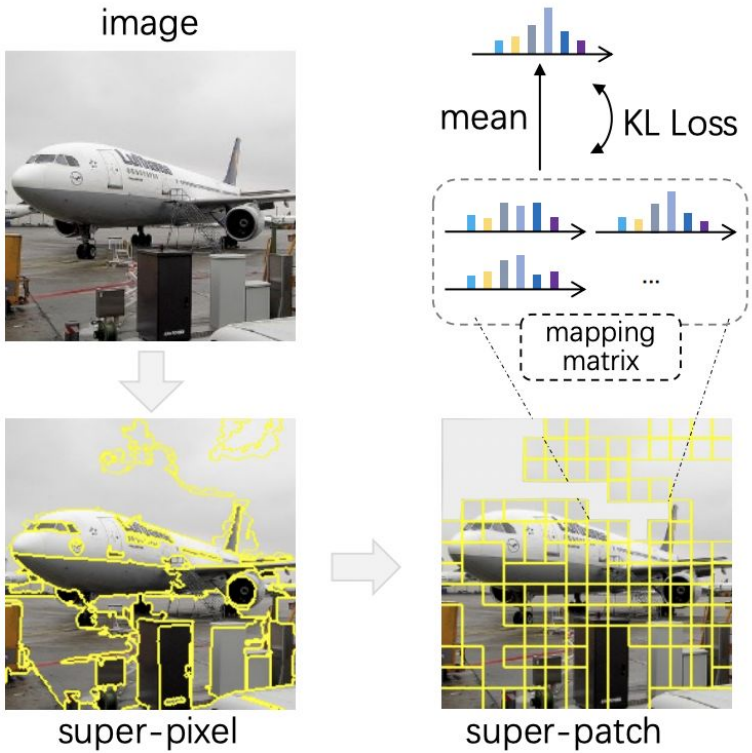


Figure 4. **Superpixel based KL Loss.**

2. Masked Language Modeling (MLM)

(1) FLAVA (CVPR 2022)

- 앞서 MIM에서도 설명

2. Masked Language Modeling (MLM)

(2) FIBER (NeurIPS 2022)

- FIBER = Fusion-In-the-Backbone-based Transformer
  - Previous: Unimodal 이후 fusion
  - FIBER: Multimodal fusion
- How? “Cross-attention” (btw image & text)
- Two-stage pretraining:
  - Step 1) Coarse (image-text)
  - Step 2) Fine-grained (image-text-box)

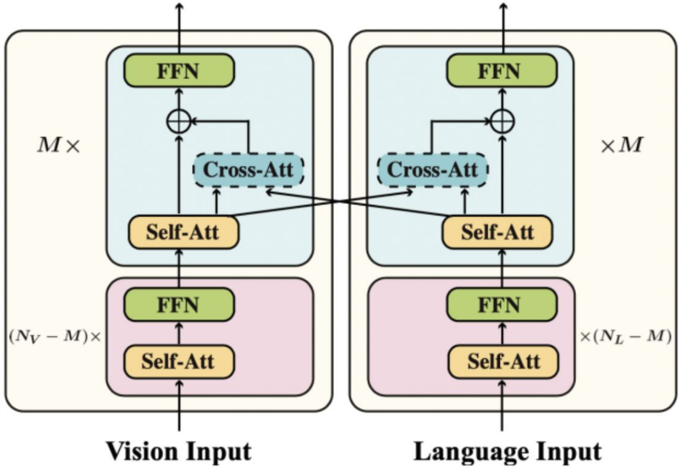
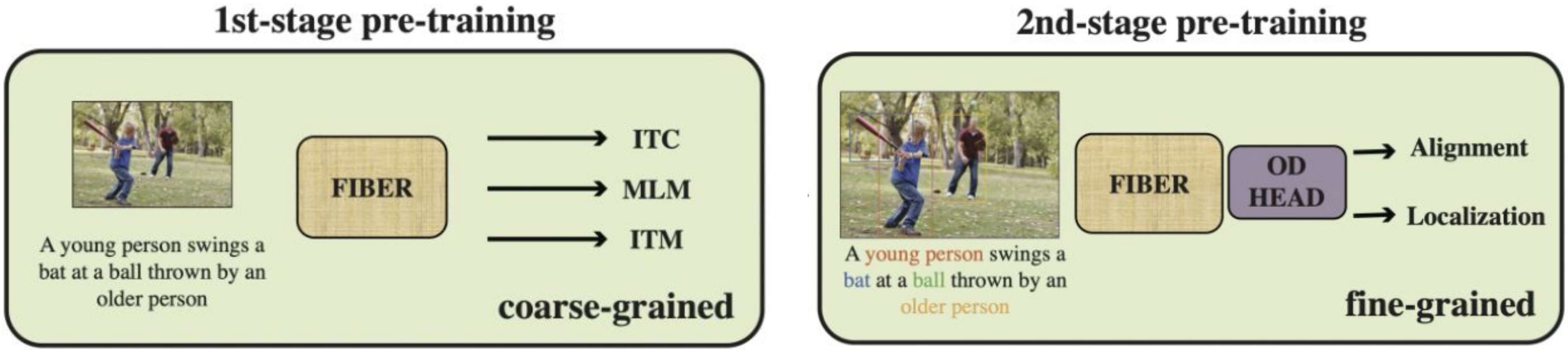


Figure 2: Model architecture for FIBER. Swin transformer is used as the image backbone, simplified here for illustration purposes.

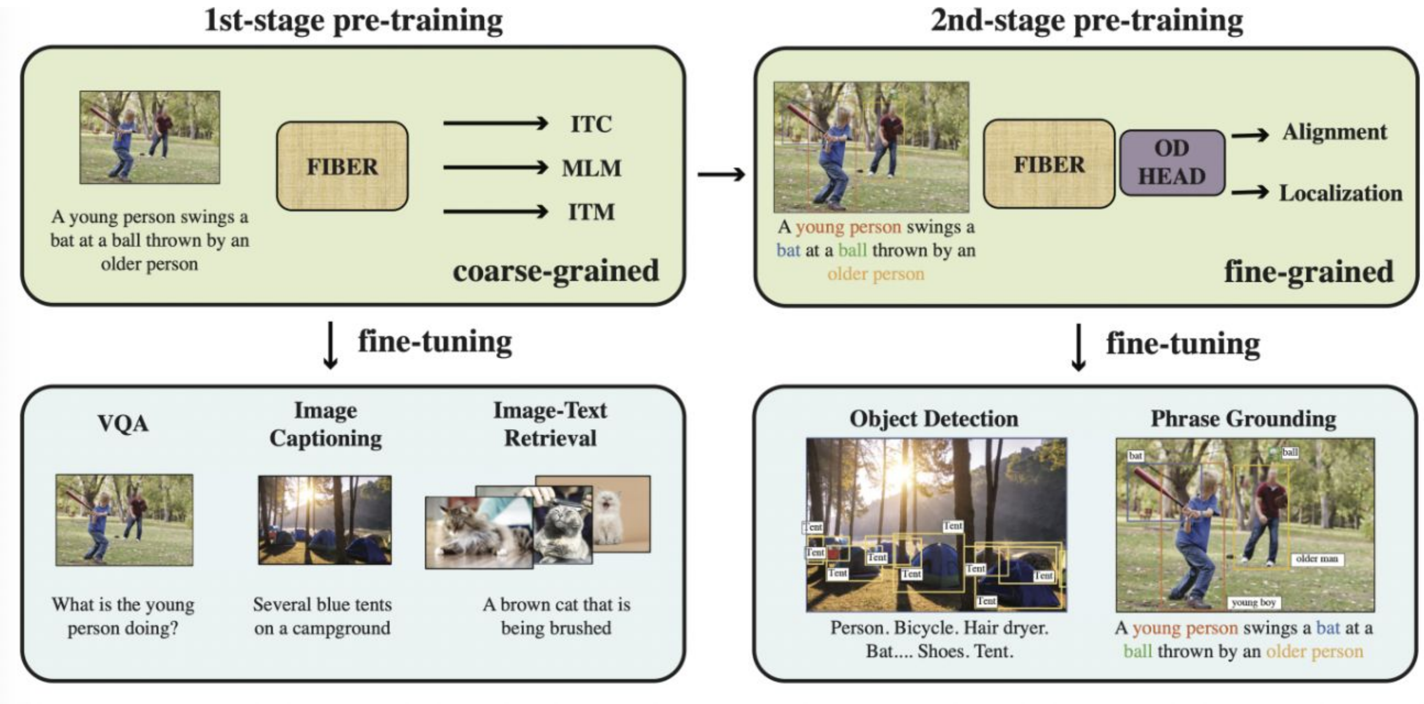
2. Masked Language Modeling (MLM)

(2) FIBER (NeurIPS 2022)



2. Masked Language Modeling (MLM)

(2) FIBER (NeurIPS 2022)



3. Masked Cross-modal Modeling (MCM)

핵심: Image & Text를 모두 masking & reconstruction

(1) FLAVA (CVPR 2022)

- 앞서 MIM에서도 설명

4. Image-to-Text Generation

Procedure

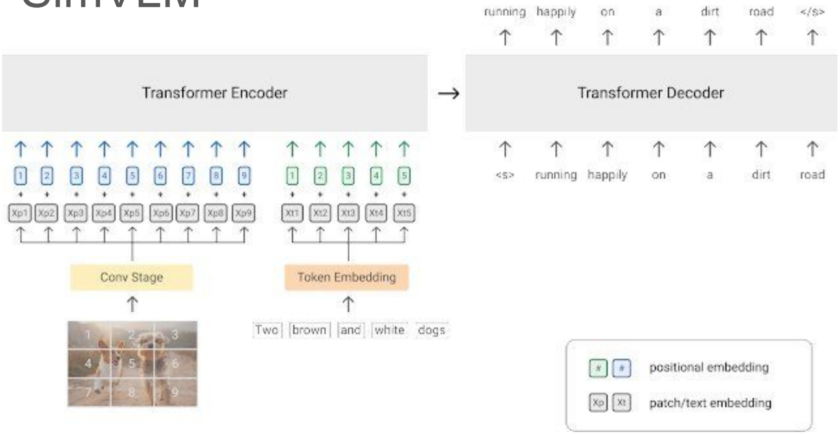
- Step 1) $z = f(\text{image})$
- Step 2) $\text{text} = g(z)$

4. Image-to-Text Generation

(1) CoCa (arxiv 2022)

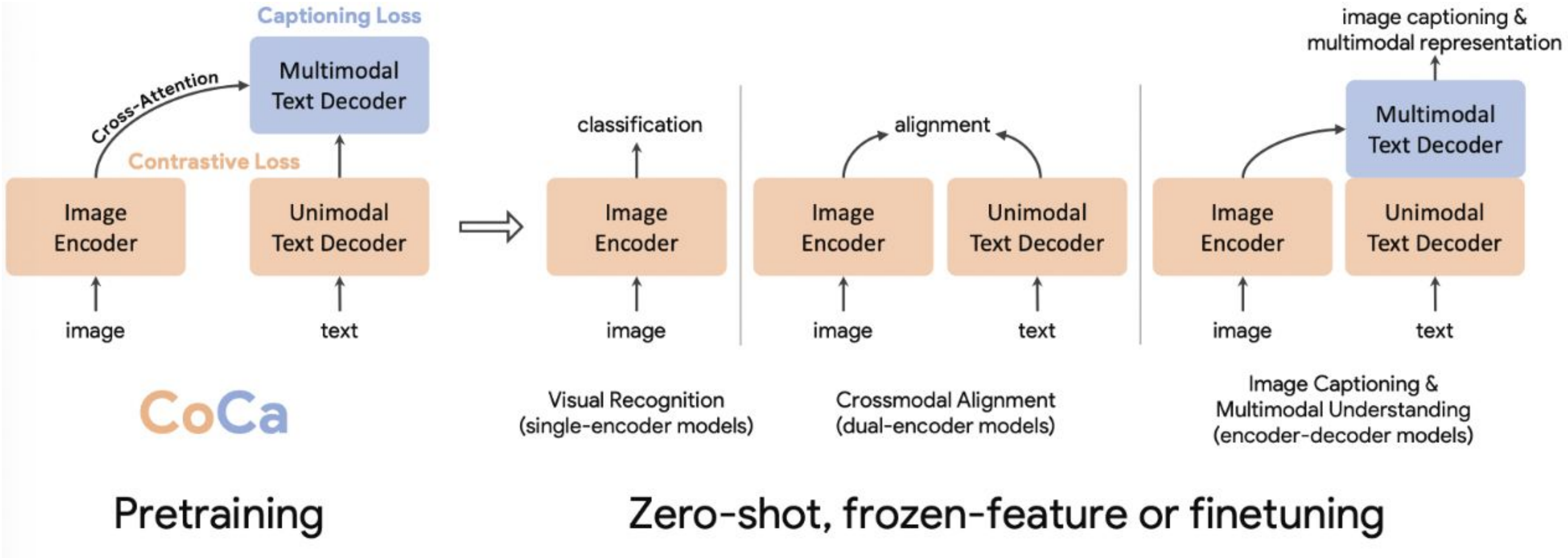
- CoCa = Contrastive Captioner
- Arch: Image-text encoder-decoder
- Loss: a) + b)
    - a) Contrastive Loss (feat. CLIP)
    - b) Captioning Loss (feat. SimVLM)

SimVLM



4. Image-to-Text Generation

(1) CoCa (arxiv 2022)



4. Image-to-Text Generation

(2) NLIP (AAAI 2023)

- CLIP 한계점: “Noisy” image-text pair
- Previous works (to reduce noise)
  - Manual rules to clean data
  - Pseudo-targets as auxiliary signals
- Proposal: Noise-robust CLIP
  - Automatically mitigate the impact of noise!

4. Image-to-Text Generation

(2) NLIP (AAAI 2023)

- Pretraining via 2 schemes
  - a) Noise harmonization: Noise 확률 예측
  - b) Noise completion: Noise 정도 감안해서 auxiliary input 넣어주기

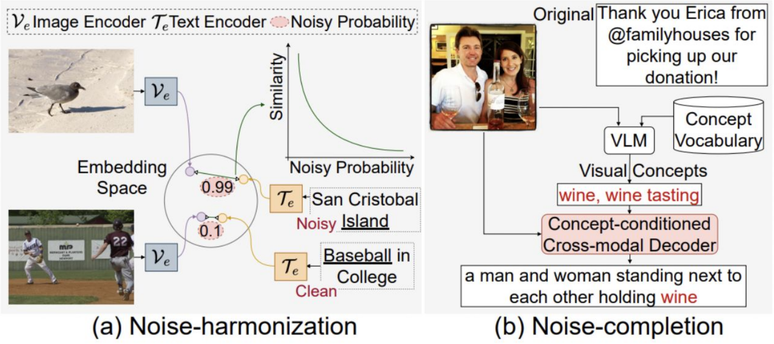


Figure 1: Illustration of two proposed schemes. (a) *Noise-harmonization*: NLIP estimates the noise probability of each image-text pair and enforces the pairs with larger noise probability to have fewer similarities in embedding space. (b) *Noise-completion*: NLIP generates enriched descriptions via a concept-conditioned captioner by taking visual concepts retrieved from a vocabulary as auxiliary inputs.

4. Image-to-Text Generation

(3) PaLI (ICLR 2023)

- PaLI = Pathways Language and Image model
- Joint modeling (Language & Vision)

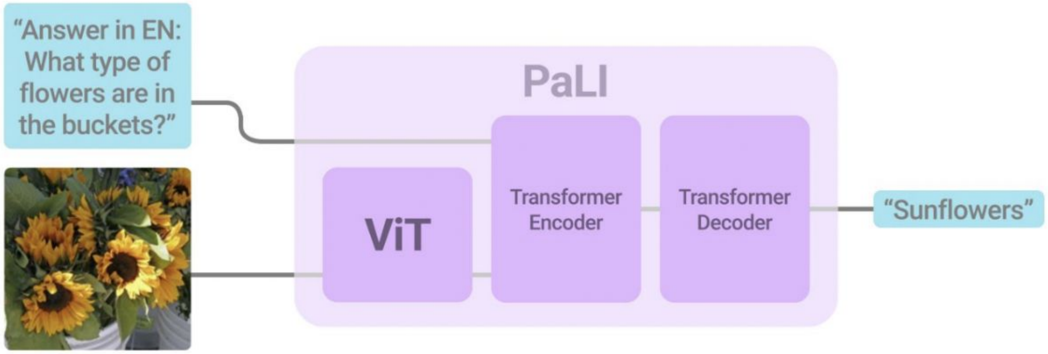


Figure 1: The PaLI main architecture is simple and scalable. It uses an encoder-decoder Transformer model, with a large-capacity ViT component for image processing.