Transitional Adaptation of Pretrained Models for Visual Storytelling

CVPR
VIRTUAL JUNE 19-25

Youngjae Yu*, Jiwan Chung*, Heeseung Yun, Jongseok Kim, and Gunhee Kim

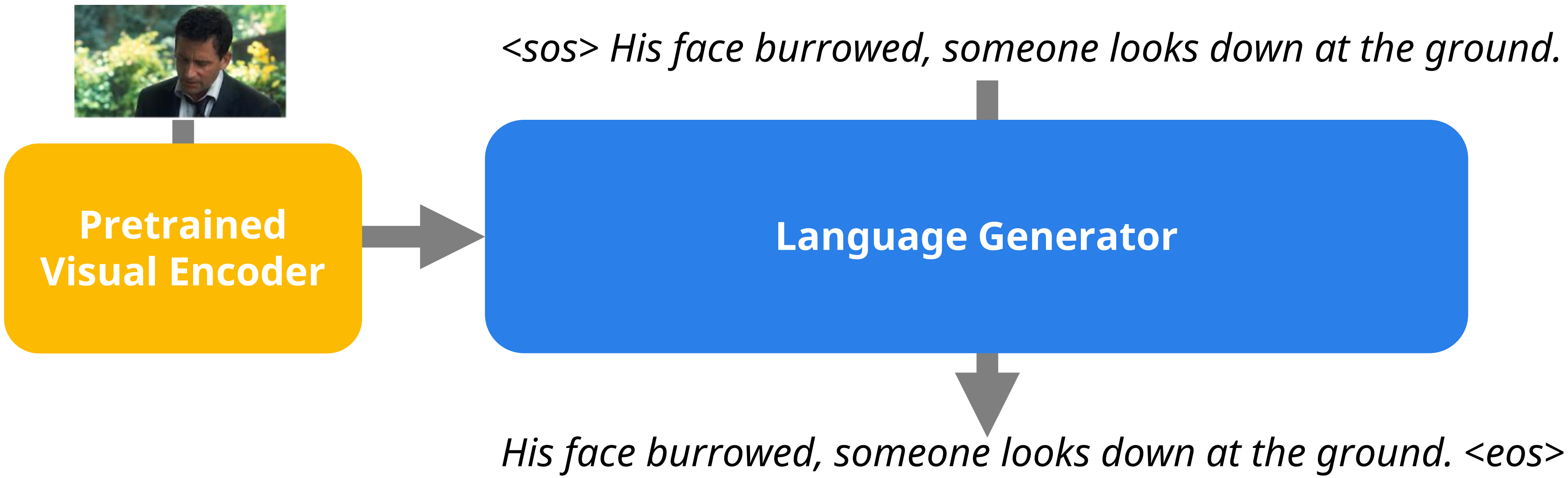
SEOUL NATIONAL UNIV.
VISION & LEARNING

*Equal Contribution

I. Transitional Adaptation

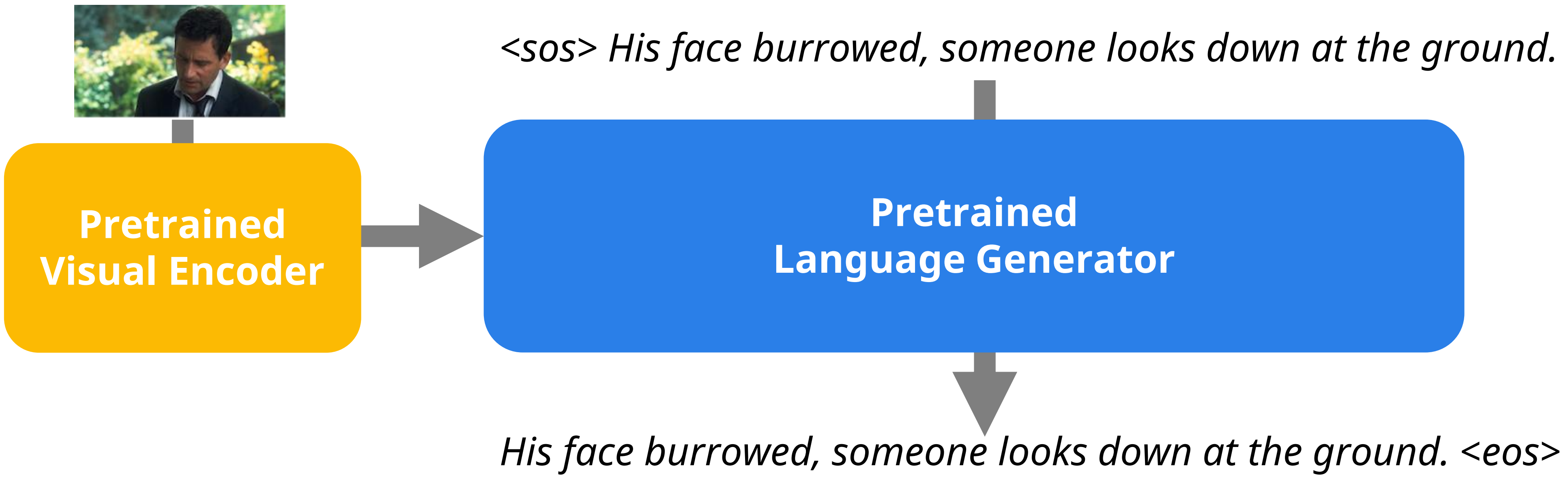
Caption Generation

- Conventional caption generation
  - Combining pretrained visual encoder with text generator
  - Example: ImageNet-pretrained ResNet + attention LSTM



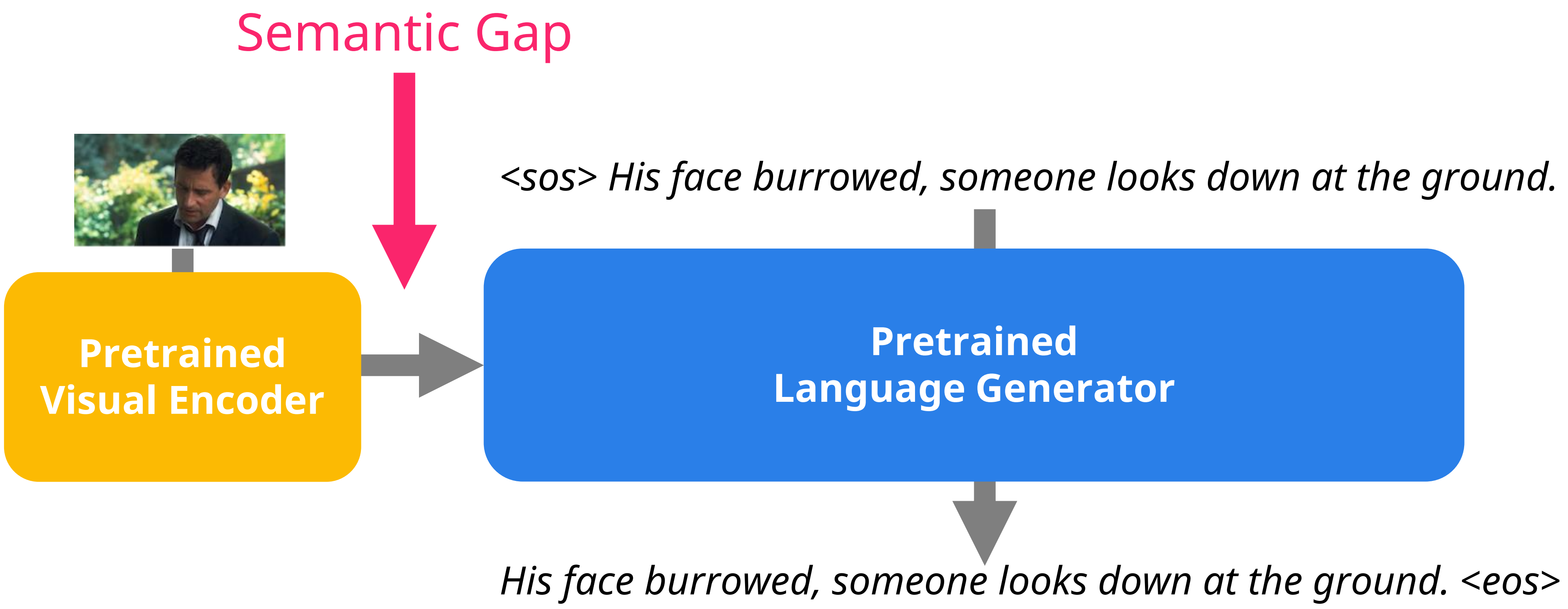
Caption Generation

- Conventional caption generation
  - Combining pretrained visual encoder with text generator
  - Example: ImageNet-pretrained ResNet + attention LSTM
- Replacing text generator with pretrained language models
  - Example: OpenAI GPT2



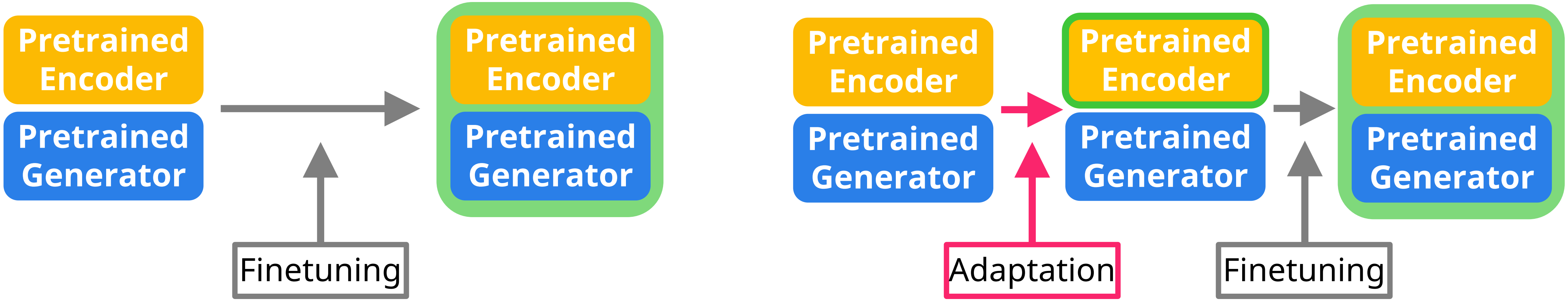
Domain Mismatch

- Exploiting pretrained language models
  - Naively utilizing pretrained language models does not improve caption quality
  - Gap between information stored in visual encoder and language model



Traditional Captioning vs. Adaptive Captioning

Traditional Captioning	Adaptive Captioning



- Directly finetune with the down-stream dataset

- Simple pretext task as an adaptation process
- Harmonizes two modules

Transitional Adaptation Process

Pretraining	Adaptation	Finetuning

Pretraining

Pretrained Encoder → *1: Goldfish*

<sos> The scientist named the … → Pretrained Generator → *The scientist named the population …*

Adaptation

→ Pretrained Encoder → Pretext Task

<sos> → Pretrained Generator → Pretext Task

Finetuning

→ Pretrained Encoder → *His face burrowed, someone …*

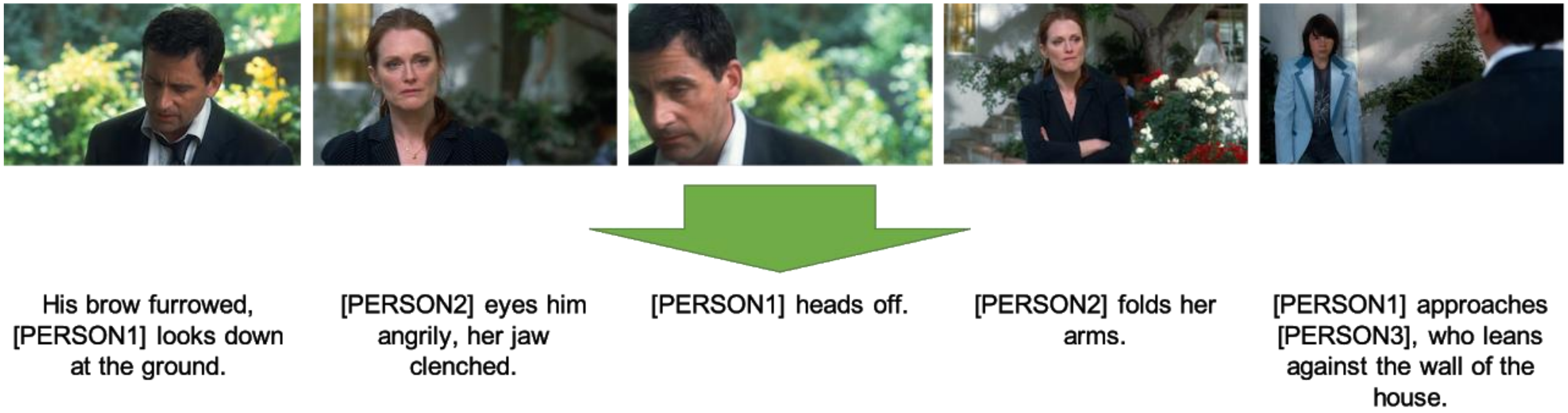
<sos> His face burrowed, … → Pretrained Generator → *His face burrowed, someone …*

Separate single-modal tasks	Adaptation cross-modal task	Target cross-modal task
Heavy computation	Light computation	Heavy computation
Unpaired data	Unpaired data	Paired data

II. Sequential Coherence Loss

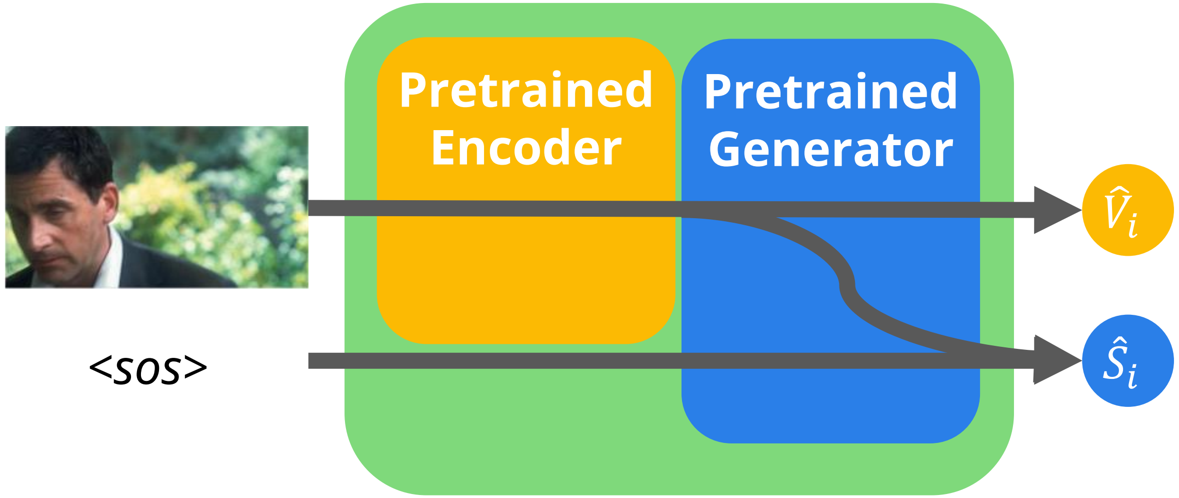
Visual Storytelling

- Visual Storytelling (Sequential Caption Generation)
  - Aims to generate a more consistent narrative for consecutive images or videos
  - Example: LSMDC 2019 task 1 (Videos), VIST (Images)



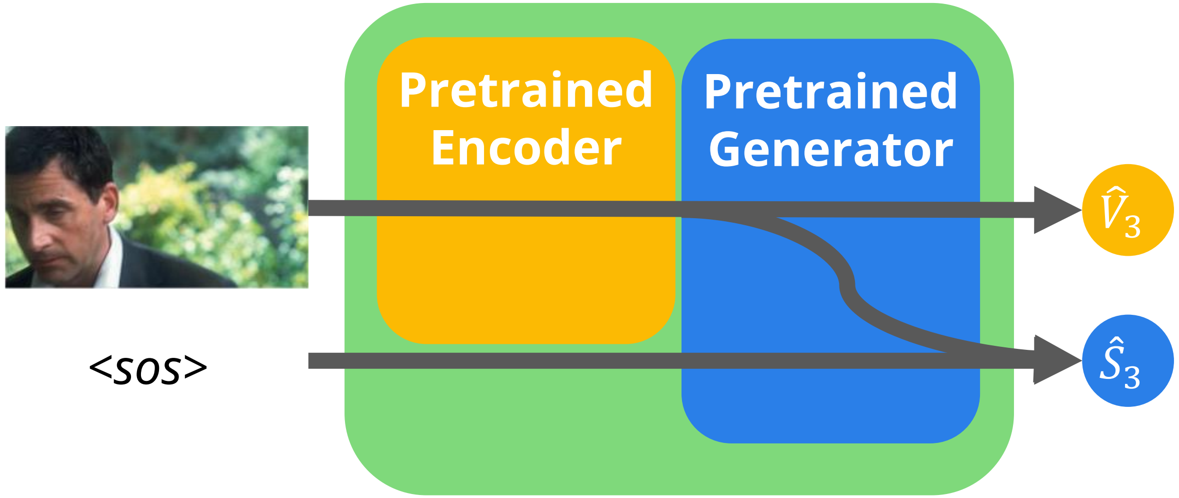
Sequential Coherence Loss

- The text representation predicts the neighbor visual representations



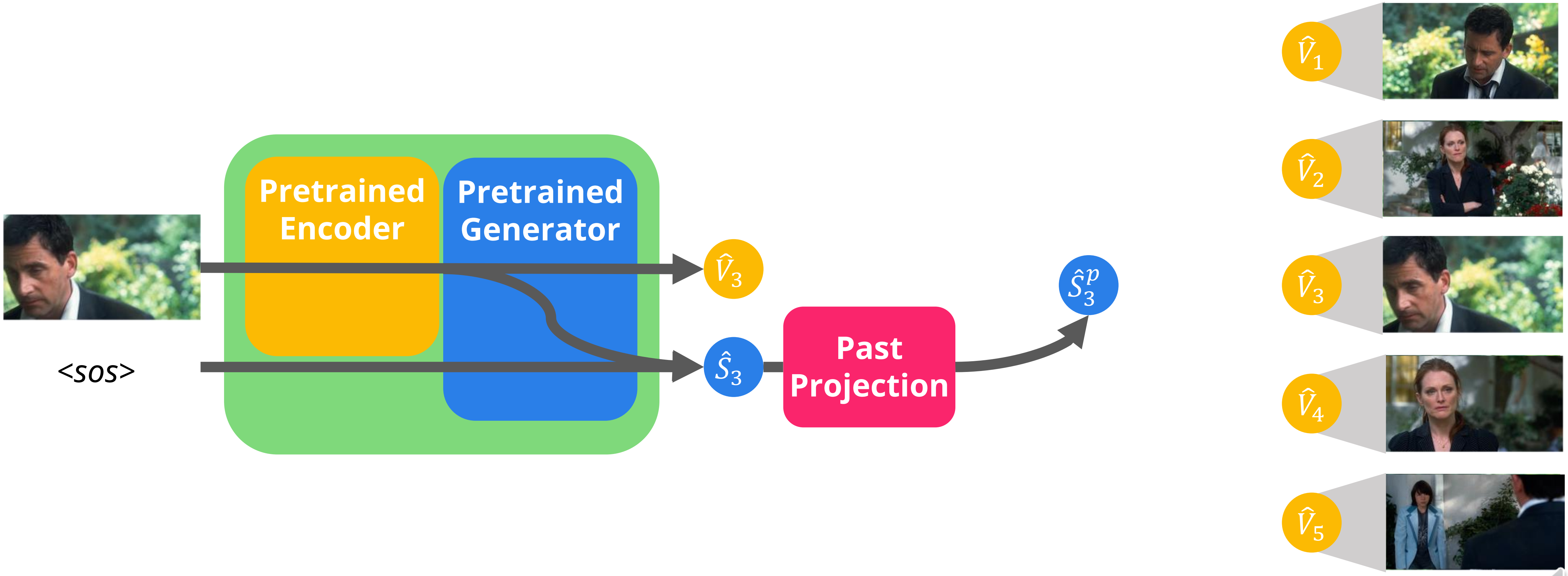
Sequential Coherence Loss

- The text representation predicts the neighbor visual representations



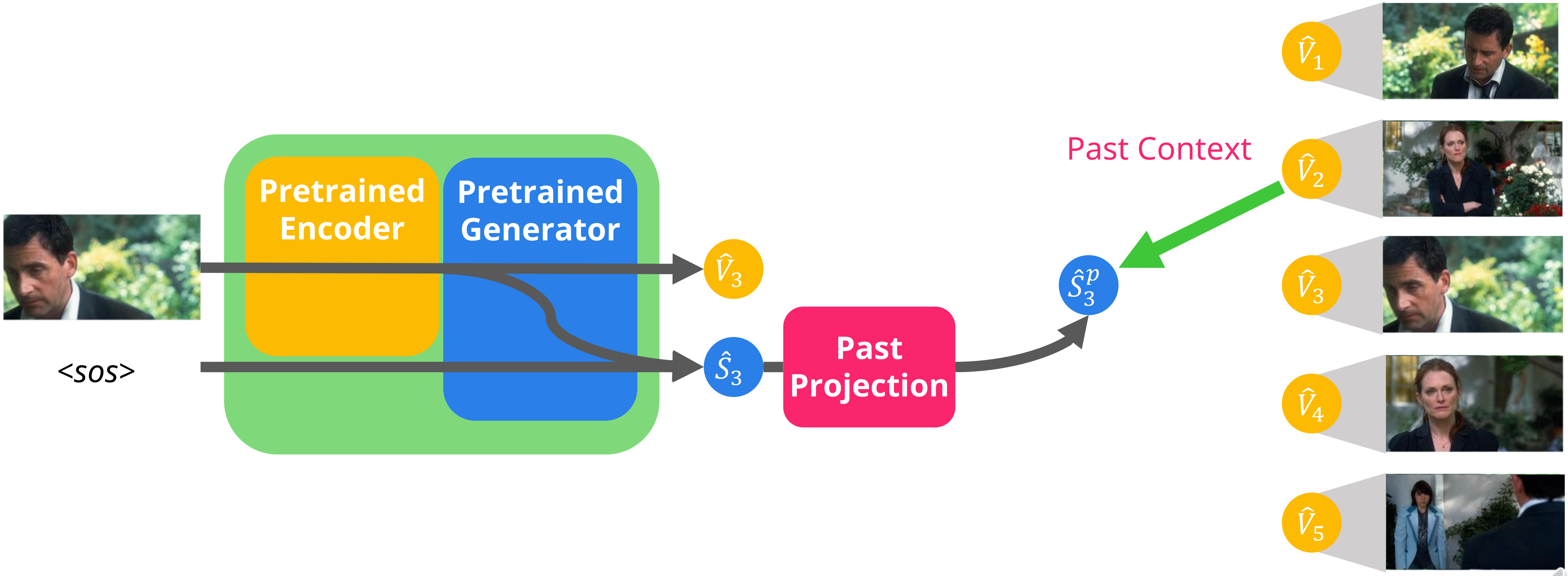
Sequential Coherence Loss

- The text representation predicts the neighbor visual representations



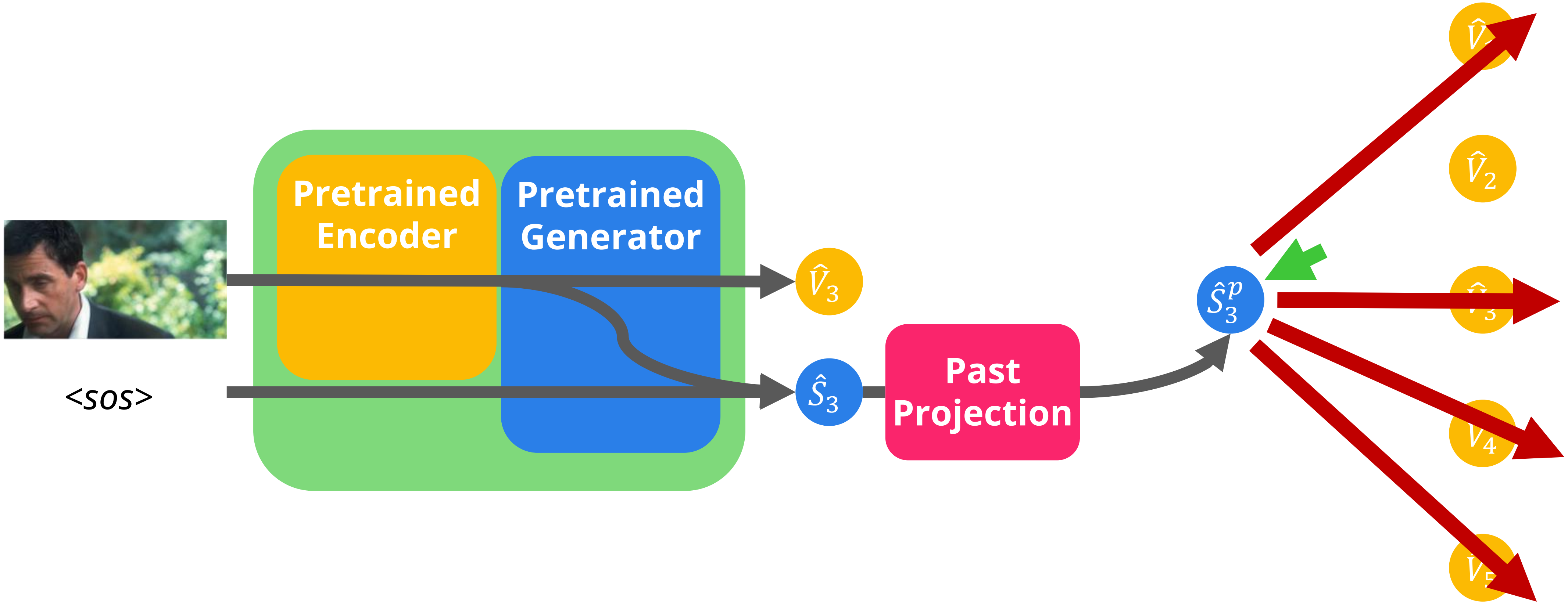
Sequential Coherence Loss

- The text representation predicts the neighbor visual representations



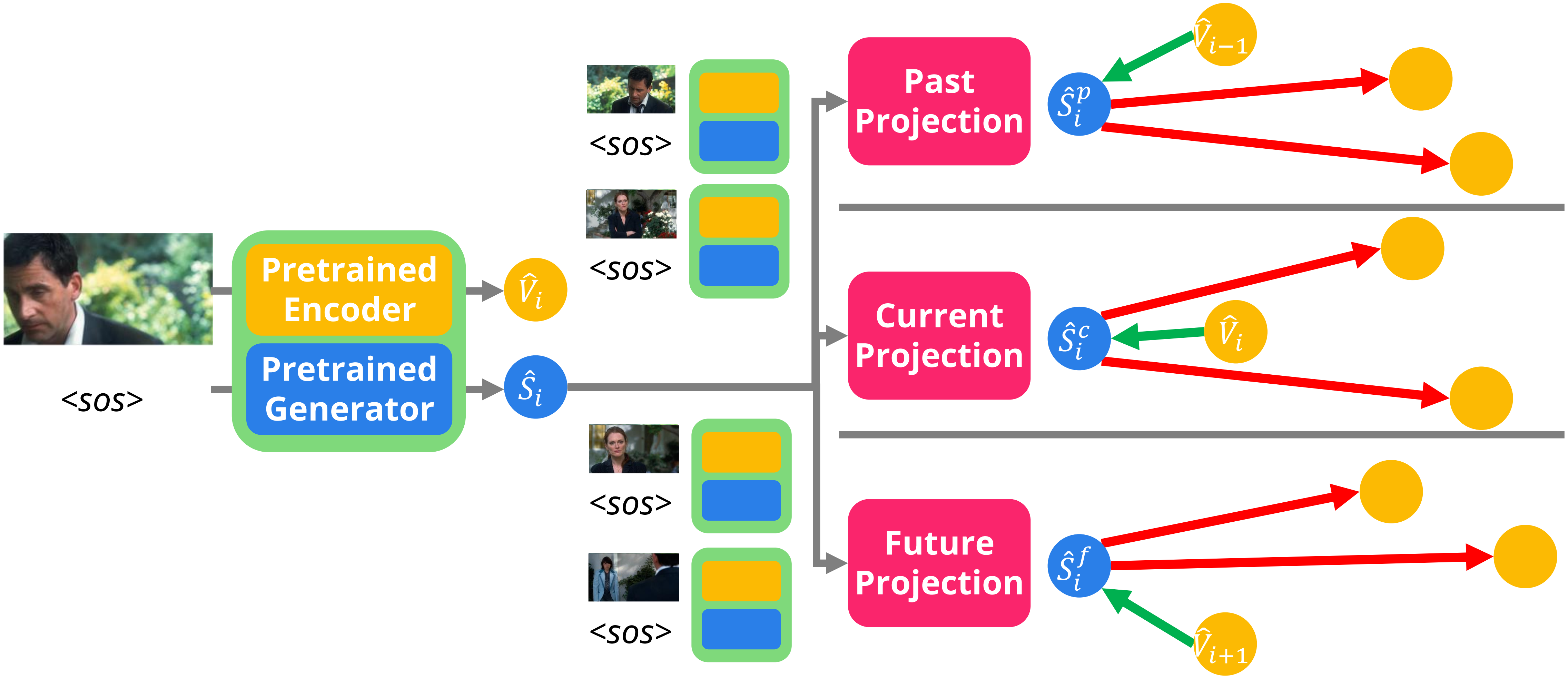
Sequential Coherence Loss

- We use contrastive loss between representations



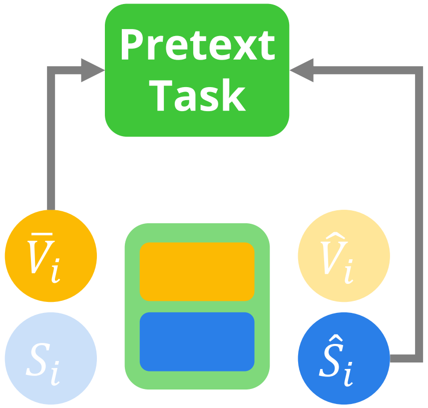
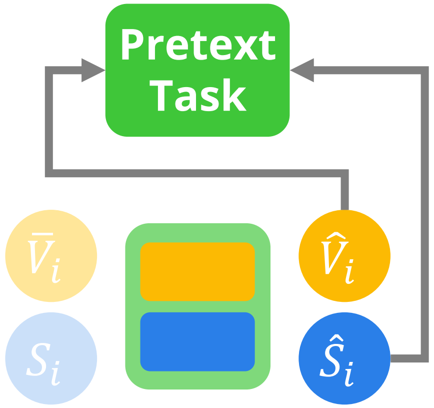
Sequential Coherence Loss

- Past, current and future losses yield both specific and coherent captions



Training with the Adaptation Loss

- Use visual representation processed with the language model for adaptation

Visual Encoder Outputs	Language Model Outputs
	
in isolation	in accordance with the language model

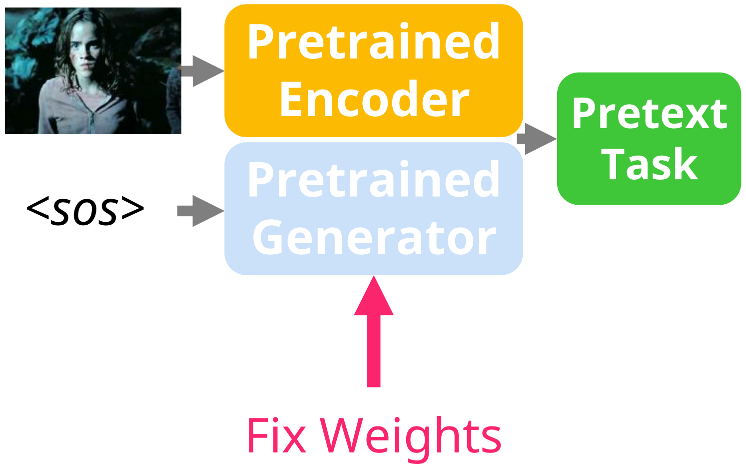
	LSMDC			VIST		
Models	C	M	R	C	M	R
Ours + language model outputs	15.37	8.41	20.21	8.3	34.1	30.2
Ours + visual encoder outputs	14.59	8.37	20.00	4.9	33.0	29.9

C: CIDEr, M: METEOR, R: ROUGE-L

Training with the Adaptation Loss

- Split-Training
  - Split the training process into adaptation and finetuning
  - Fix the generator weights during the adaptation

Adaptation



	LSMDC			VIST		
Models	C	M	R	C	M	R
Ours + split-training	15.37	8.41	20.21	8.3	34.1	30.2
Ours + joint training	14.28	8.34	19.71	4.5	32.8	29.8

C: CIDEr, M: METEOR, R: ROUGE-L

Experiment: Settings

- Datasets
  - LSMDC 2019 task 1: Video storytelling
  - VIST: Image storytelling
- Backbones
  - Visual encoder: Visual features + Simple encoder with FC layers and attention
  - Text Generator: GPT2-small
- Evaluation
  - Automatic text metrics: (CIDEr, METEOR, ROUGE-L)
  - Human evaluations
    - LSMDC 2019: how helpful they are for a blind person
    - VIST: pairwise comparison test following previous research

Experiment: Quantitative Results

- Superior performance in both LSMDC 2019 and VIST
- Greater performance gap in CIDEr score (human-likeness)

LSMDC

	Public Test		Blind Test	
Models	C	M	C	M
Official Baseline	7.0	12.0	6.9	11.9
XE	7.2	11.5	-	-
AREL	7.3	11.4	-	-
TAPM (Ours)	**10.0**	**12.3**	**8.8**	**12.4**

C: CIDEr, M: METEOR

VIST

Models	**C**	**M**	**R**
Huang et al	-	31.4	-
h-attn-rank	7.5	34.1	29.5
GLACNet	-	30.1	-
AREL	9.4	35.0	29.5
StoryAnchor	9.9	35.5	30.0
HSRL	10.7	35.2	30.8
TAPM (Ours)	**13.8**	**37.2**	**33.1**

C: CIDEr, M: METEOR, R: ROUGE-L

Experiment: Human Evaluation Results

- Automatic metrics often fail to capture expressiveness and coherence
- Human evaluators prefer TAPM in both LSMDC 2019 and VIST

LSMDC

Models	Scores
Human	1.085
Official Baseline	4.015
TAPM (Ours)	**3.670**

The lower the better.

VIST

	TAPM *vs* XE			TAPM *vs* AREL		
Choice (%)	TAPM	XE	Tie	TAPM	XE	Tie
Relevance	**59.9**	34.1	6.0	**61.3**	32.8	5.9
Expressiveness	**57.3**	32.3	10.4	**57.3**	34.0	8.7
Concreteness	**59.1**	30.3	10.7	**59.6**	30.4	10.0

The higher the better.

Summary

- The first attempt to use adaptation loss in welding a visual encoder with a pretrained language model
- Sequential coherence loss designed for adaptation in visual storytelling tasks
- Superior experimental results on both LSMDC 2019 and VIST

Thank You

http://vision.snu.ac.kr/projects/TAPM

SEOUL NATIONAL UNIV.
VISION & LEARNING