
EMNLP 2021

Multimodal Pretraining Unmasked:

A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs



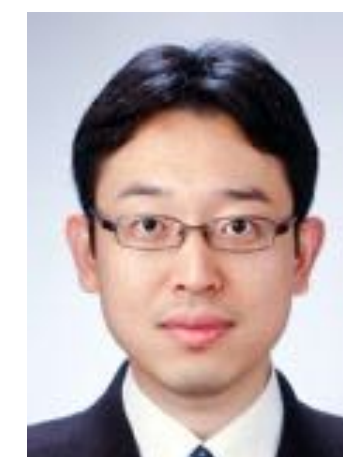
Emanuele Bugliarello

University of Copenhagen
emanuele@di.ku.dk



Ryan Cotterell

ETH Zürich
University of Cambridge
rcotterell@inf.ethz.ch



Naoaki Okazaki

Tokyo Institute of Technology
okazaki@c.titech.ac.jp



Desmond Elliott

University of Copenhagen
de@di.ku.dk

TL;DW

TL;DW

- Unified theoretical and software framework for current V&L BERTs

Code, models & data online: github.com/e-bug/volta

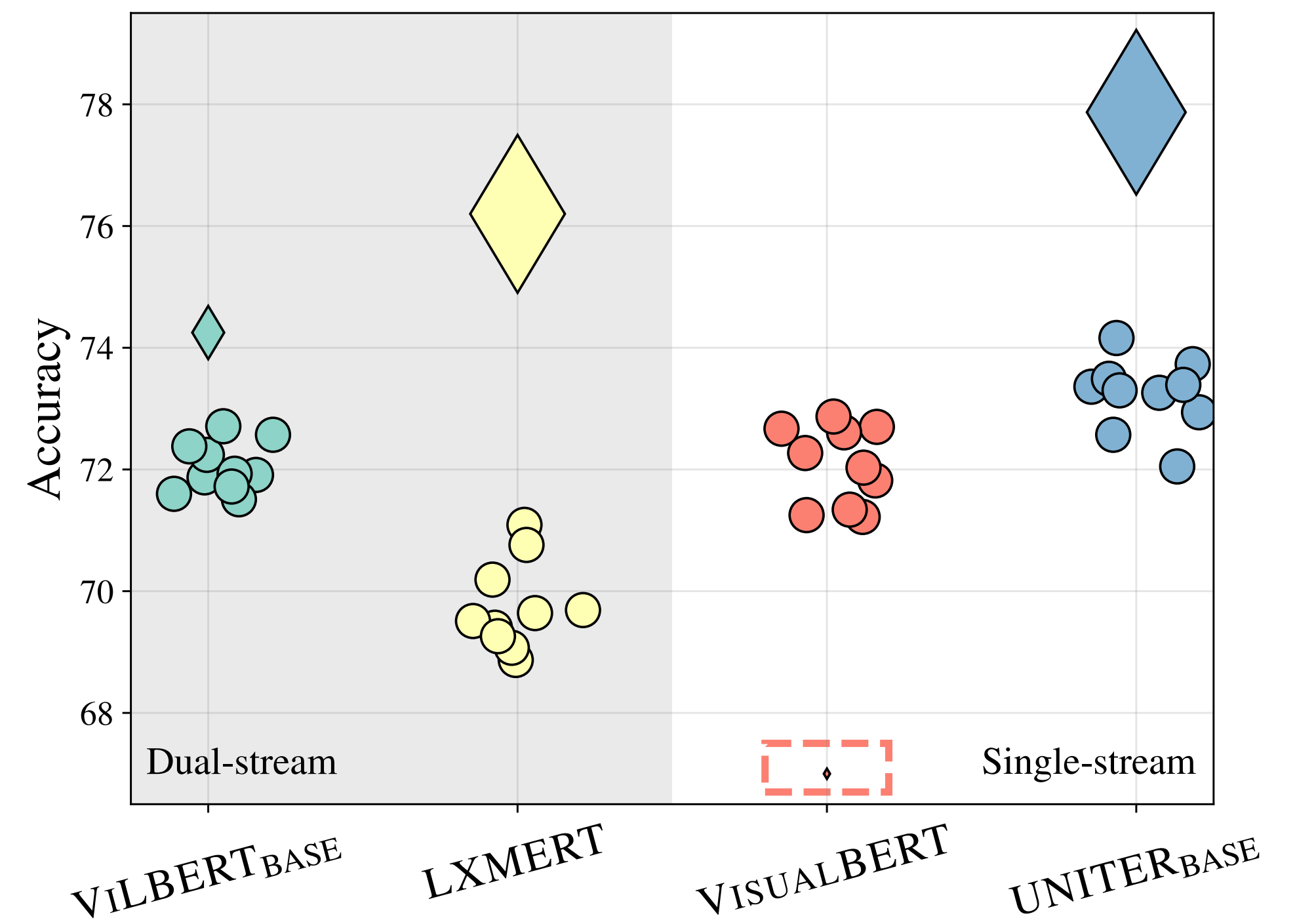
TL;DW

- Unified theoretical and software framework for current V&L BERTs
- Controlled experiments

Code, models & data online: github.com/e-bug/volta

TL;DW

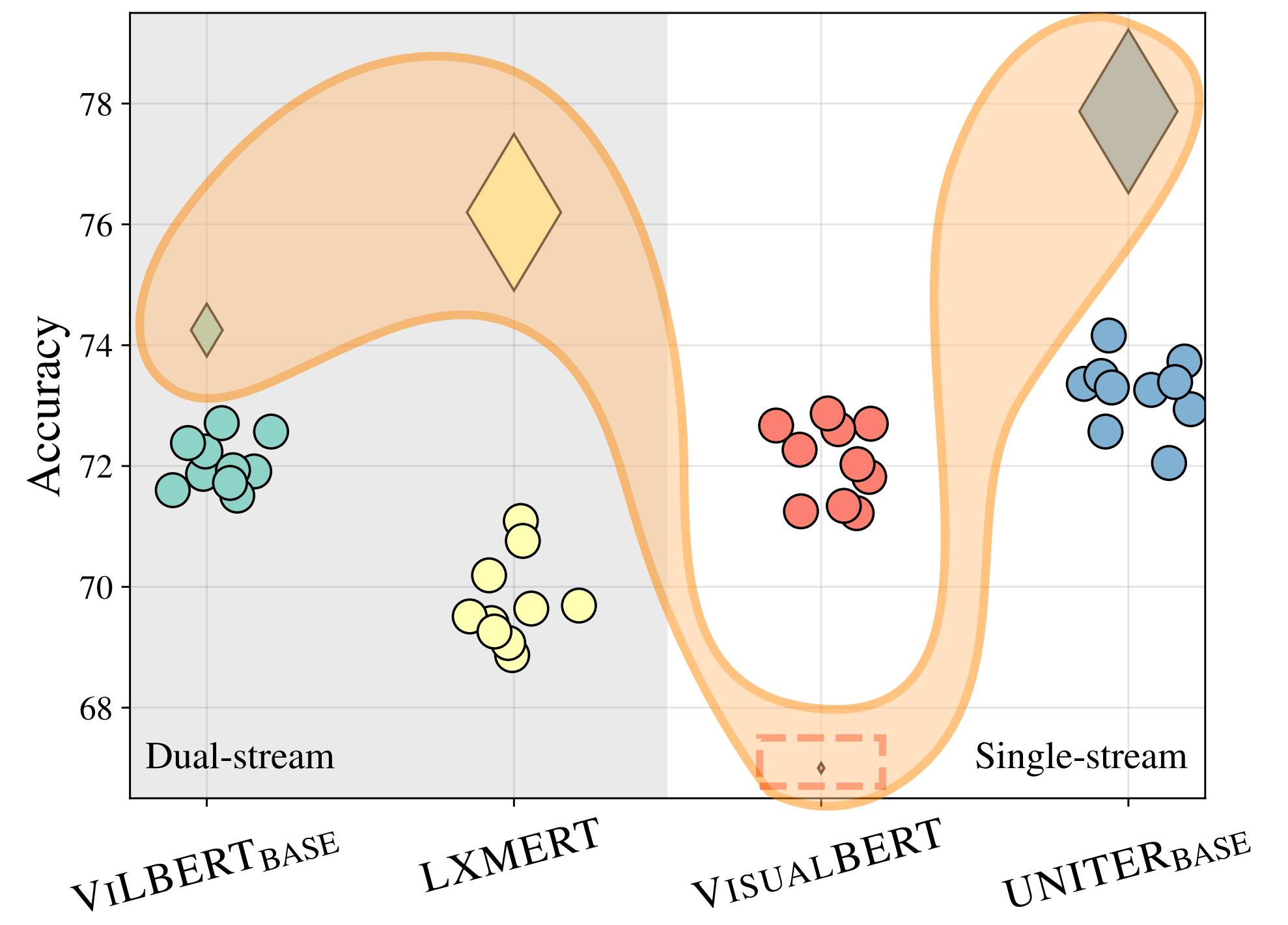
- Unified theoretical and software framework for current V&L BERTs
- Controlled experiments



Code, models & data online: github.com/e-bug/volta

TL;DW

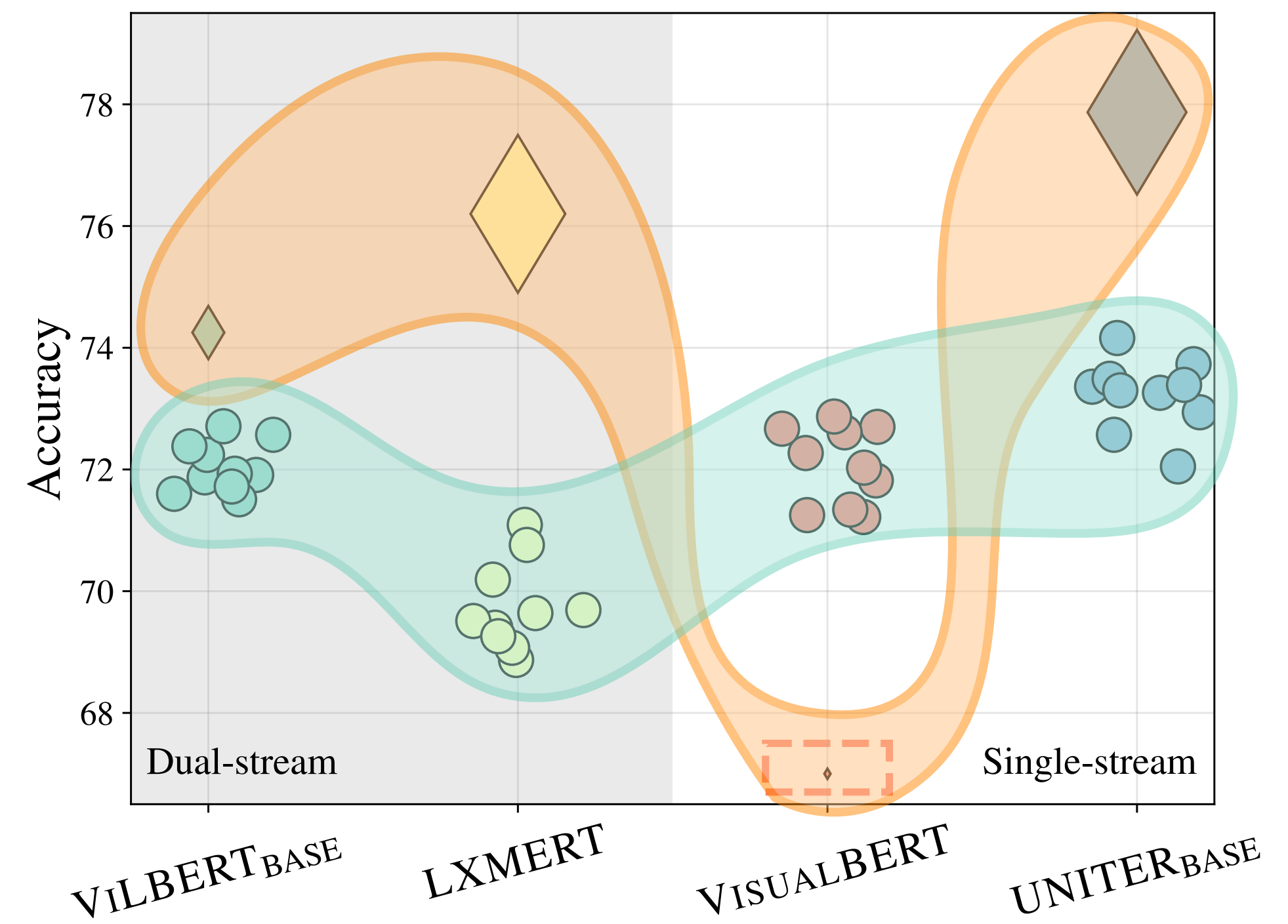
- Unified theoretical and software framework for current V&L BERTs
- Controlled experiments



Code, models & data online: github.com/e-bug/volta

TL;DW

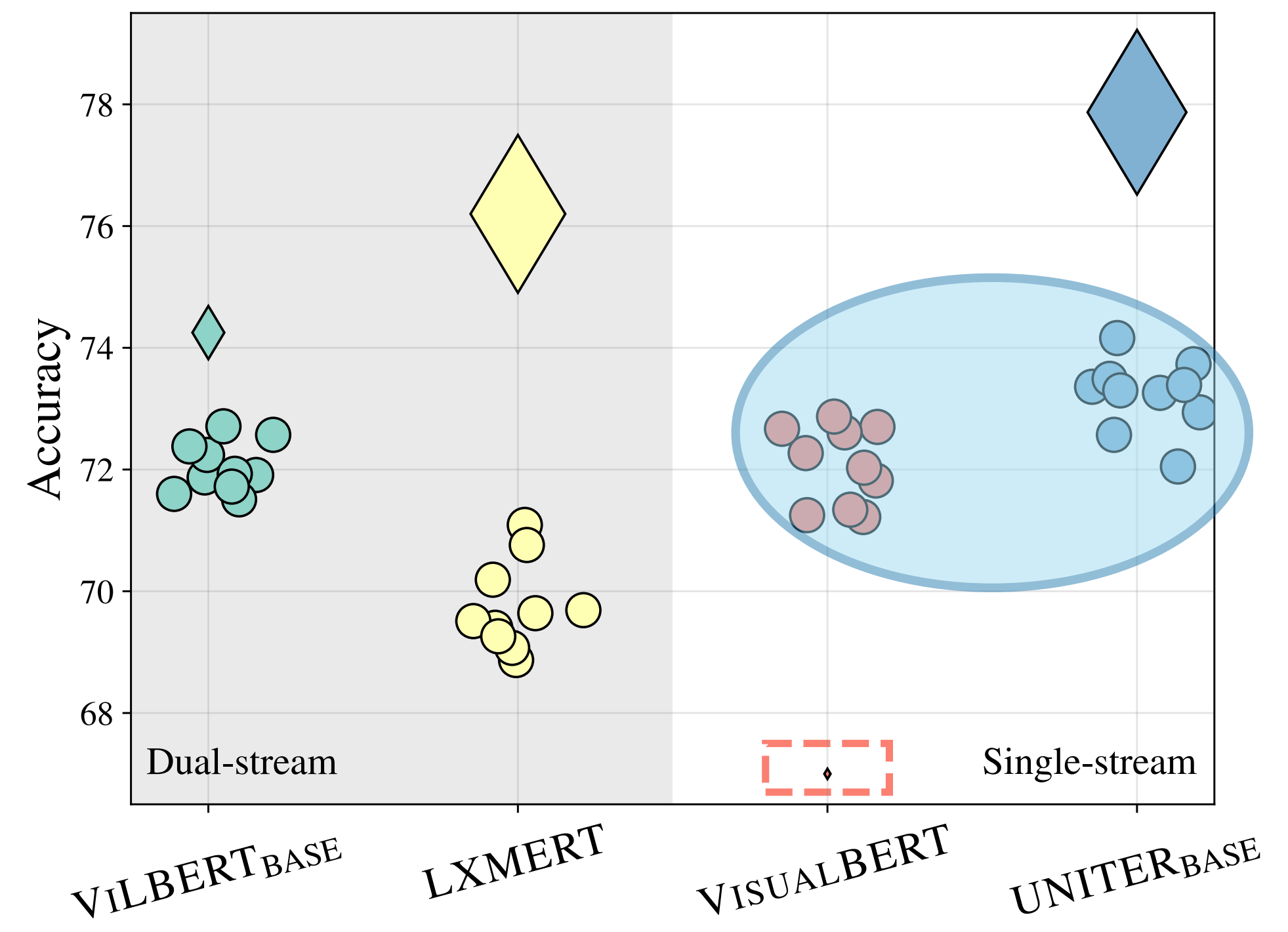
- Unified theoretical and software framework for current V&L BERTs
- Controlled experiments
 - Similar performance with same training data and hyperparameters



Code, models & data online: github.com/e-bug/volta

TL;DW

- Unified theoretical and software framework for current V&L BERTs
- Controlled experiments
 - Similar performance with same training data and hyperparameters
 - Embedding layer plays a crucial role



Code, models & data online: github.com/e-bug/volta

V&L Transformer Architectures

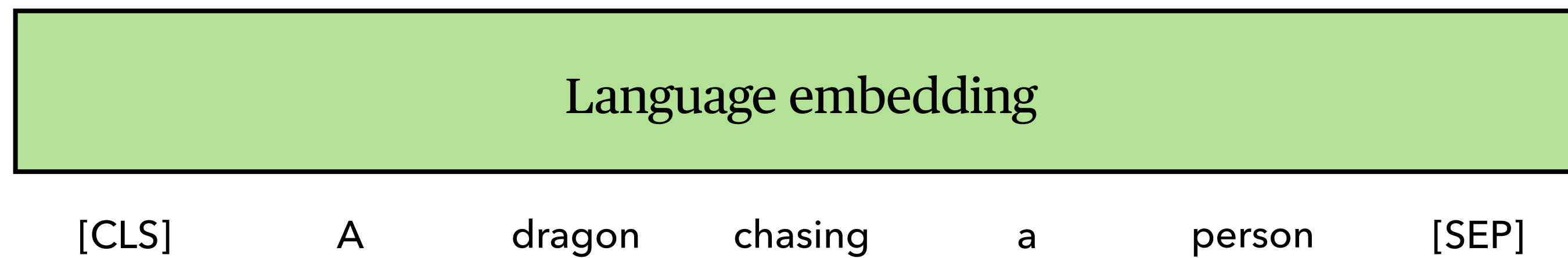
Overview

A dragon chasing a person

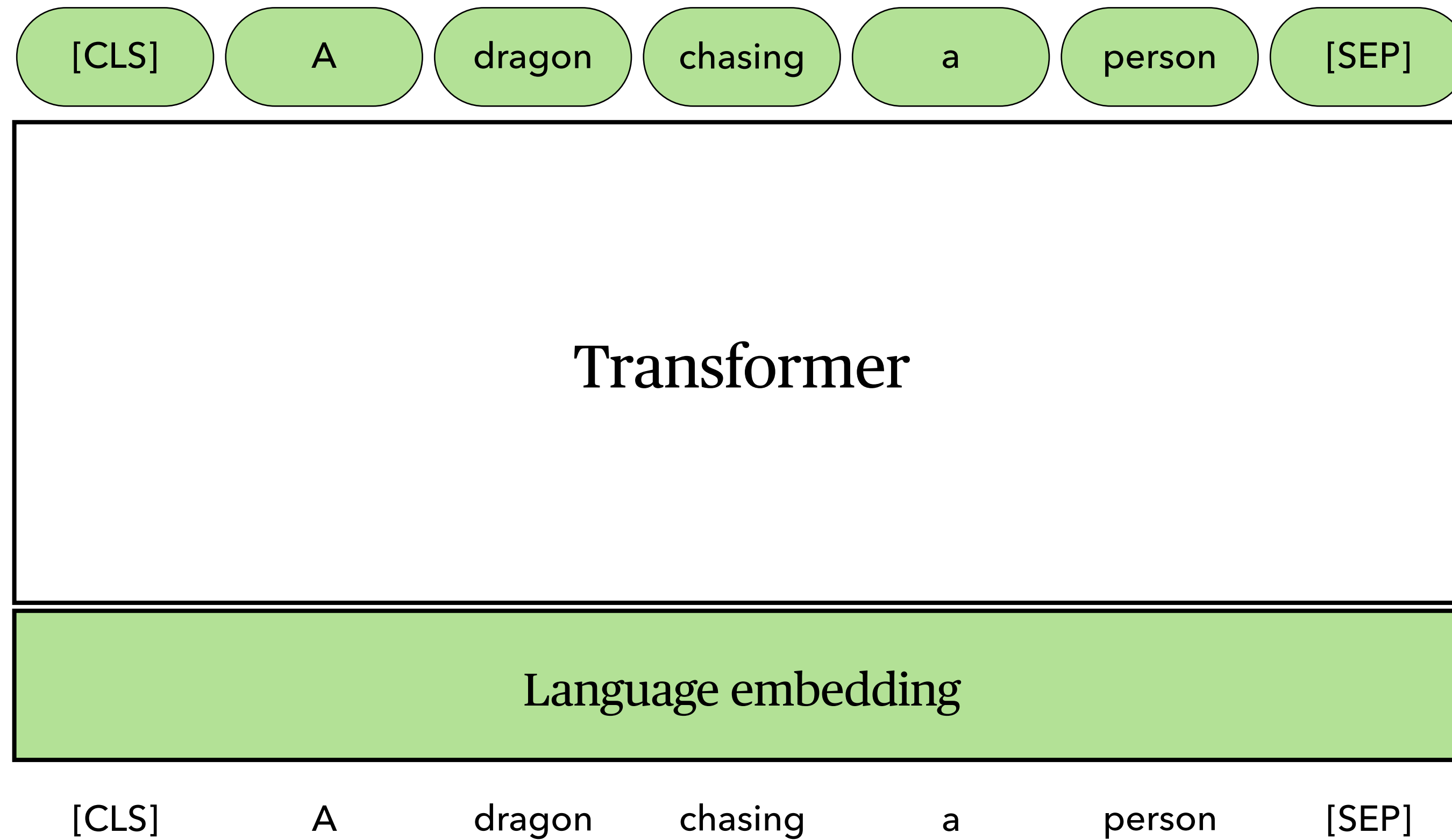
Overview

[CLS] A dragon chasing a person [SEP]

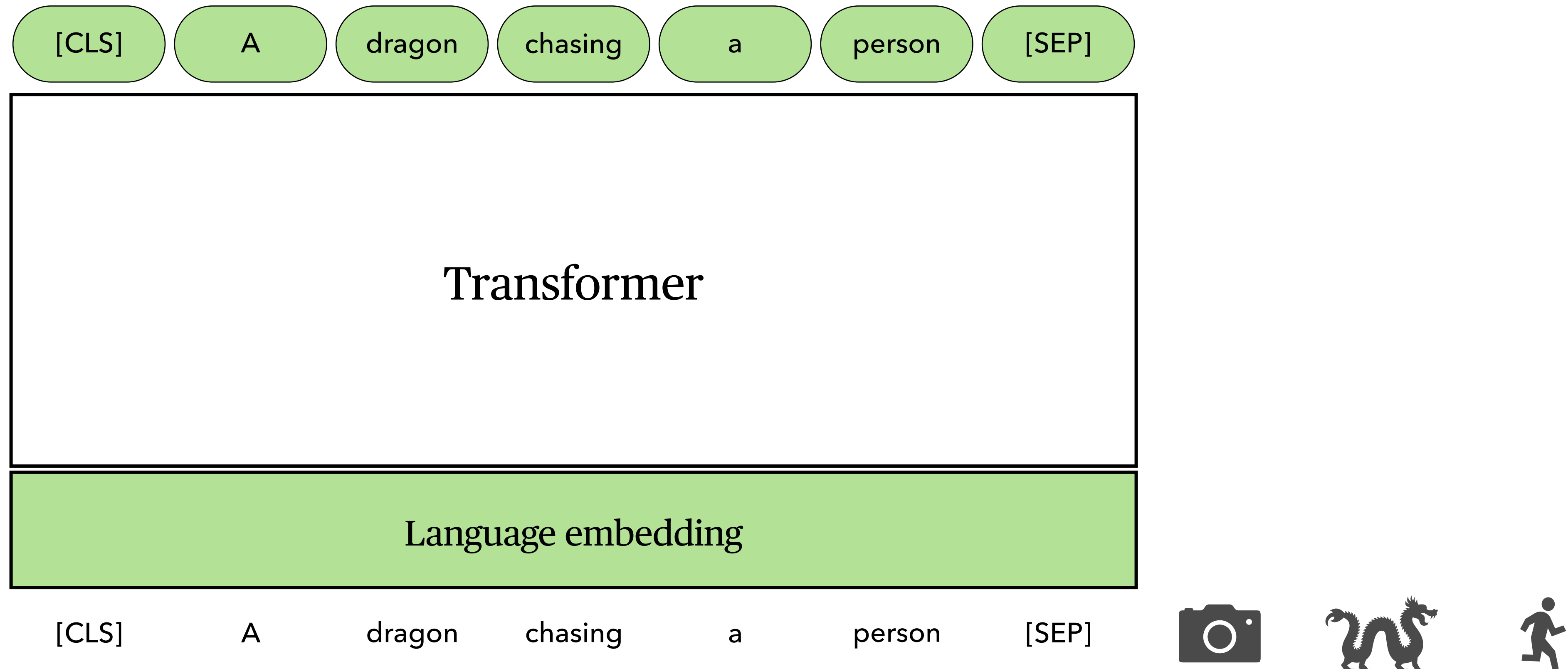
Overview



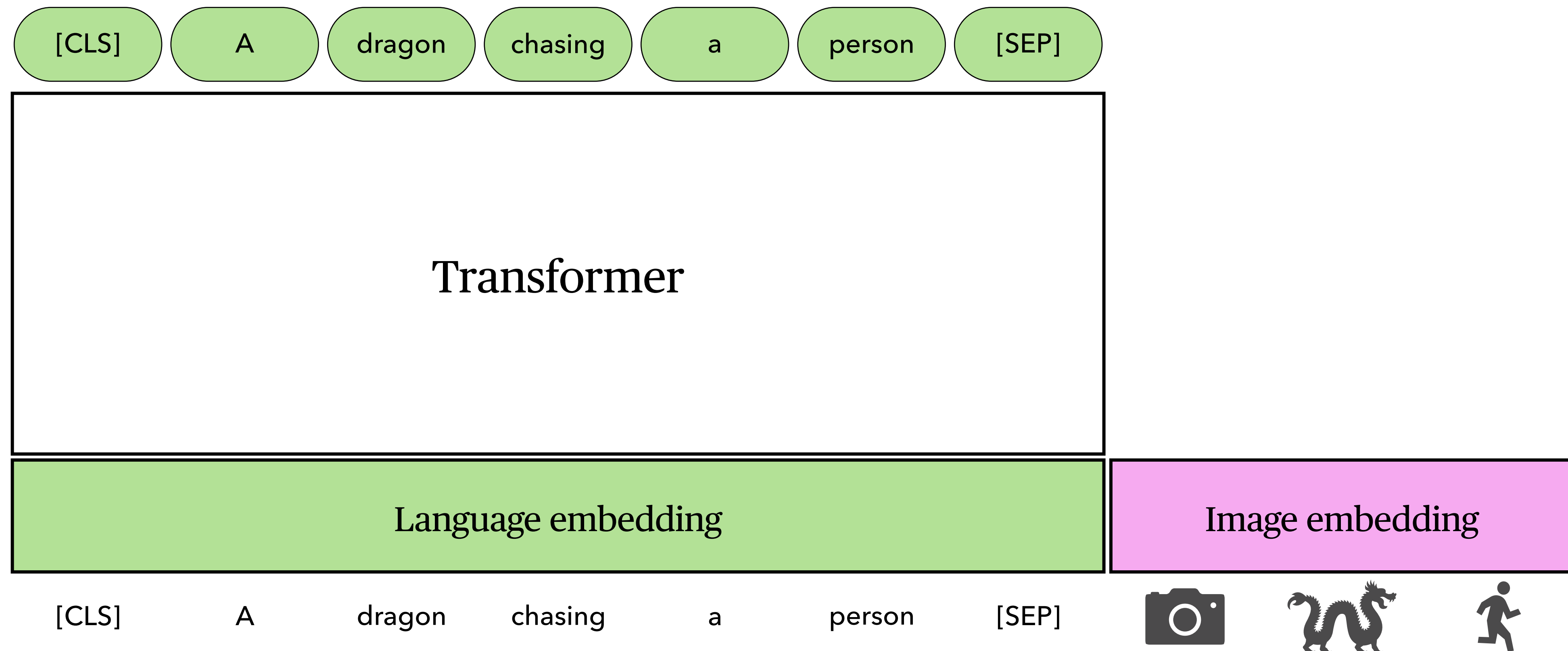
Overview



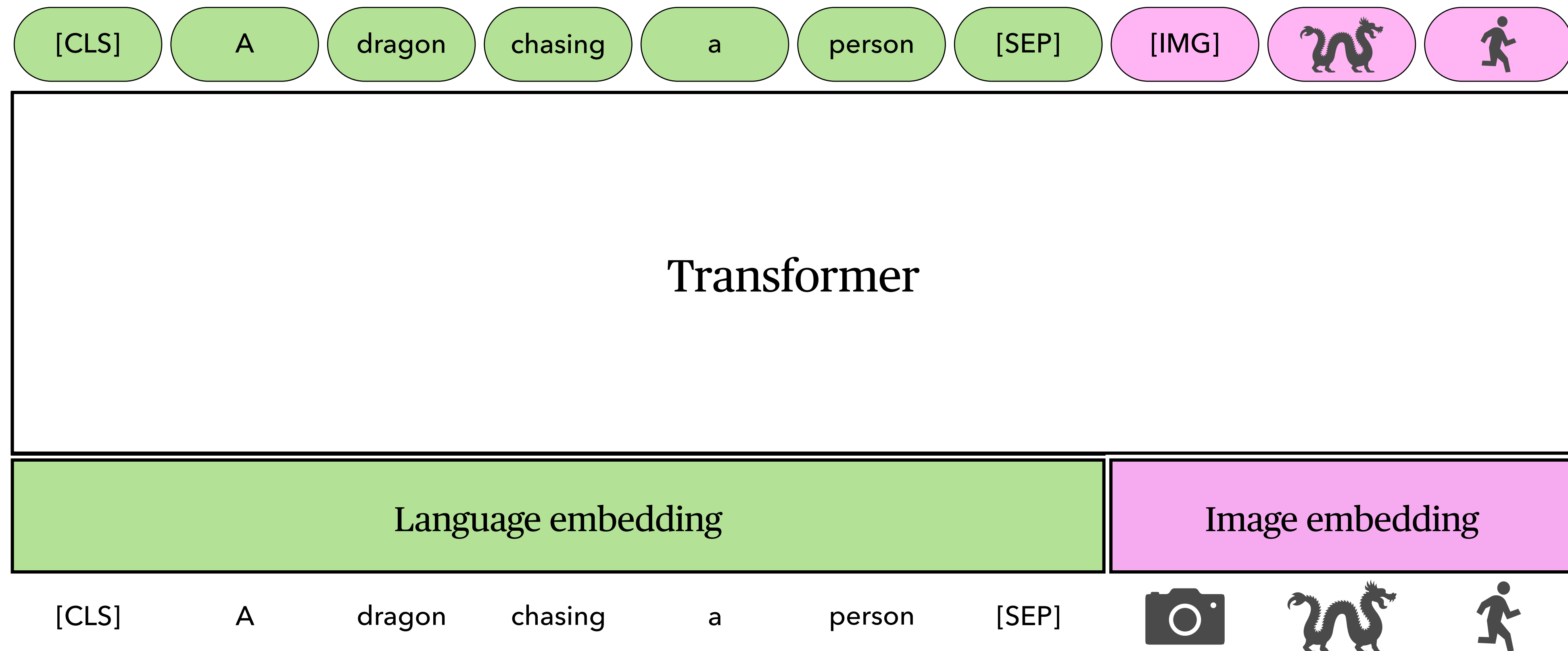
Overview



Overview



Overview



Single- & Dual-Stream Architectures

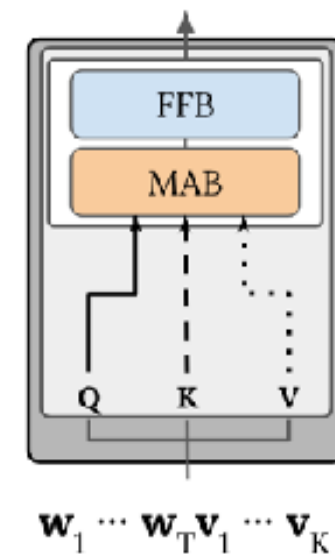
Single- & Dual-Stream Architectures

Single-Stream

Single- & Dual-Stream Architectures

Single-Stream

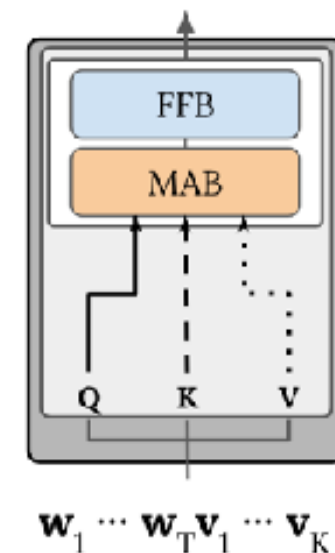
- Concat image–text inputs



Single- & Dual-Stream Architectures

Single-Stream

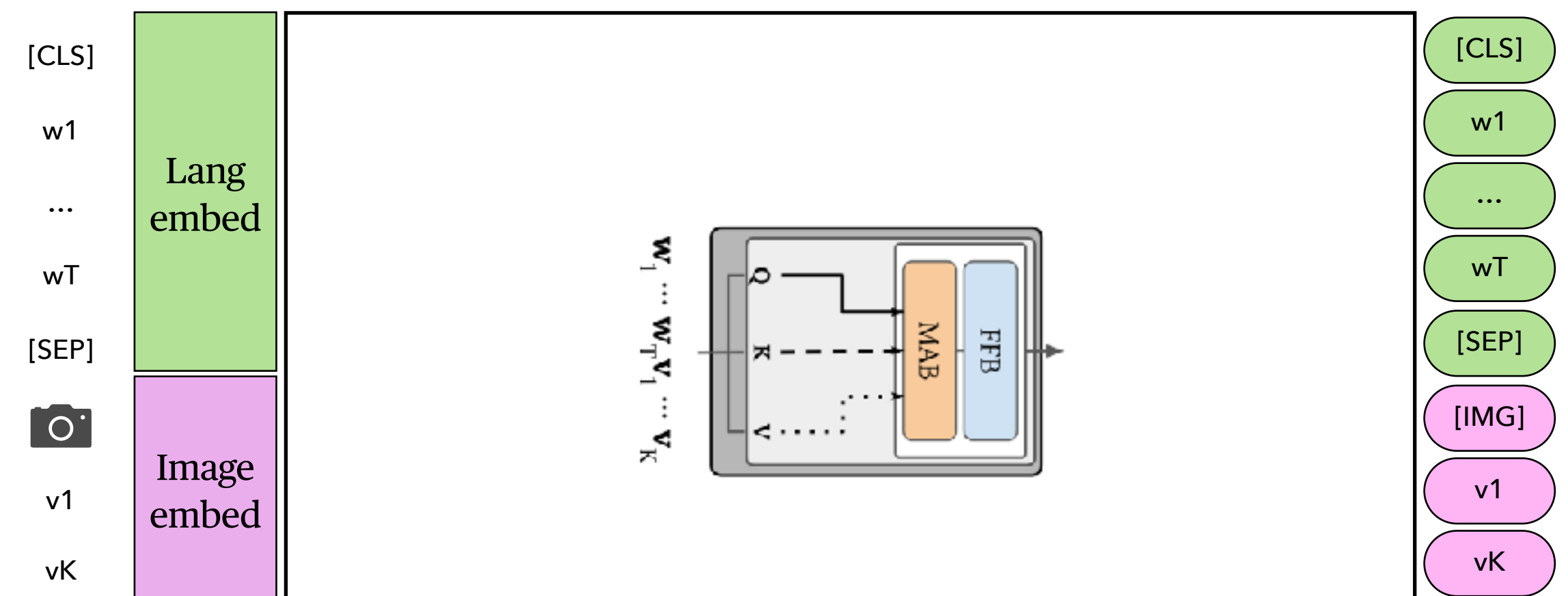
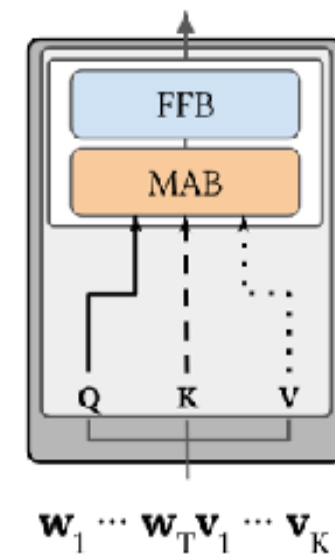
- Concat image–text inputs



Single- & Dual-Stream Architectures

Single-Stream

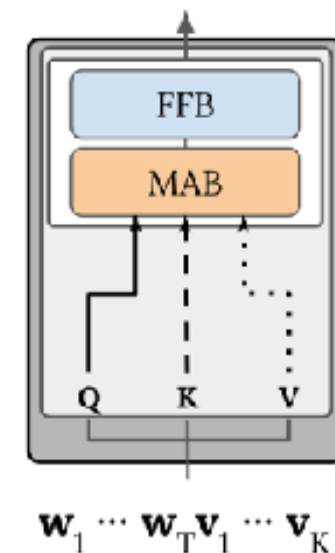
- Concat image–text inputs



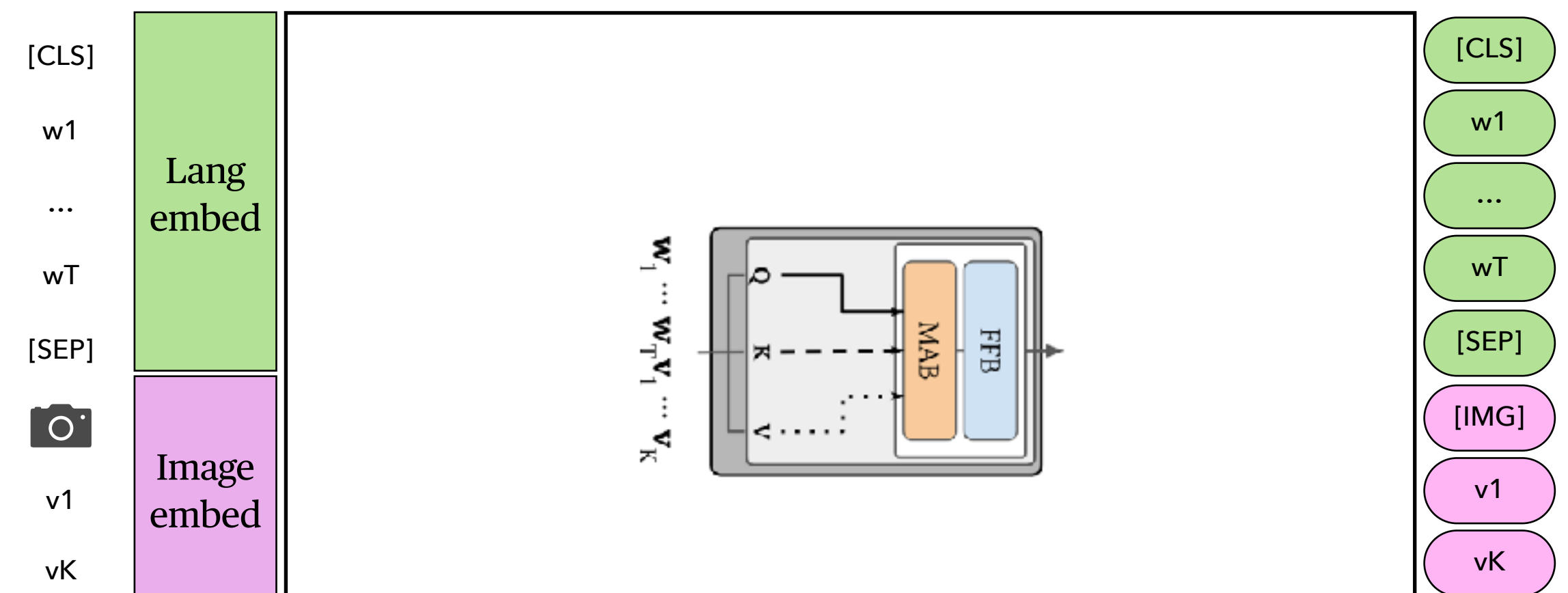
Single- & Dual-Stream Architectures

Single-Stream

- Concat image–text inputs



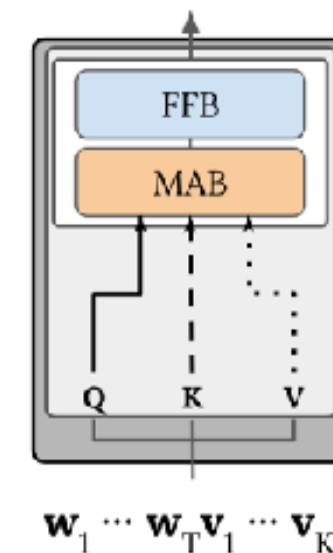
Dual-Stream



Single- & Dual-Stream Architectures

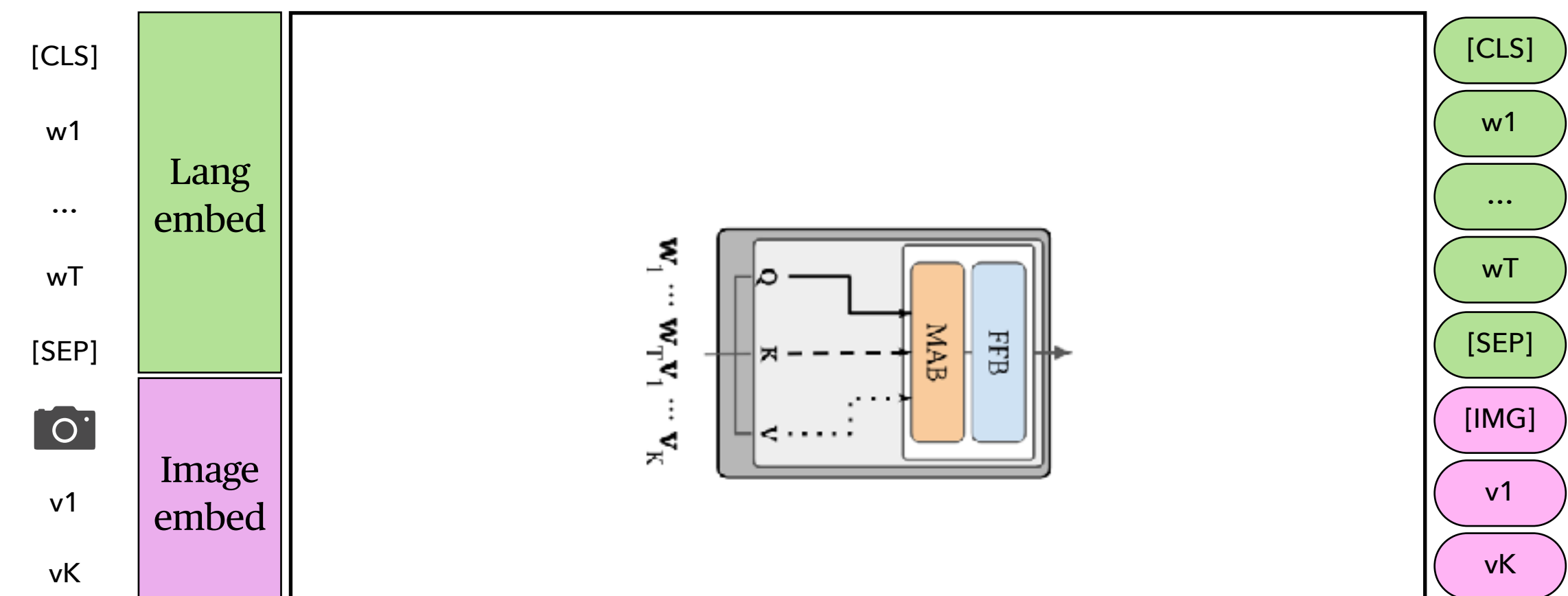
Single-Stream

- Concat image–text inputs



Dual-Stream

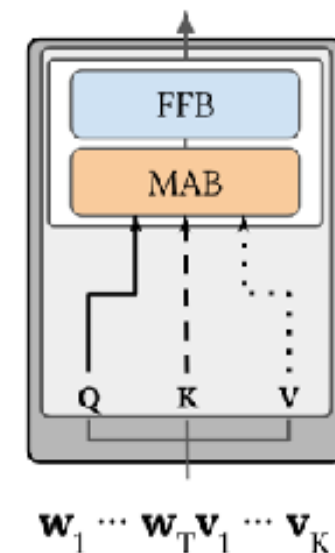
1. Image and text independently



Single- & Dual-Stream Architectures

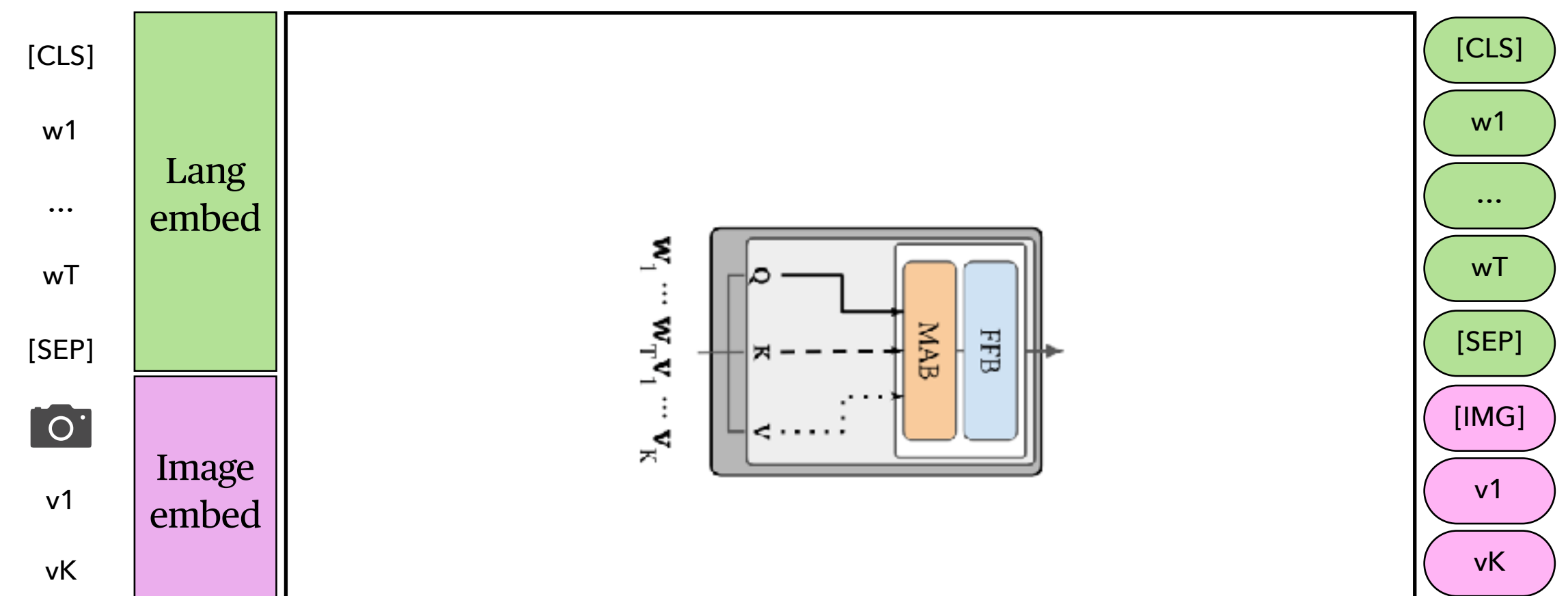
Single-Stream

- Concat image–text inputs



Dual-Stream

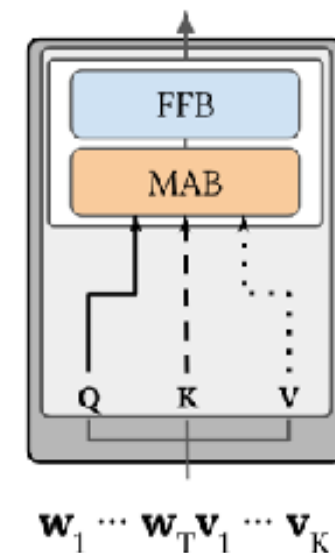
1. Image and text independently
2. Cross-modal layers



Single- & Dual-Stream Architectures

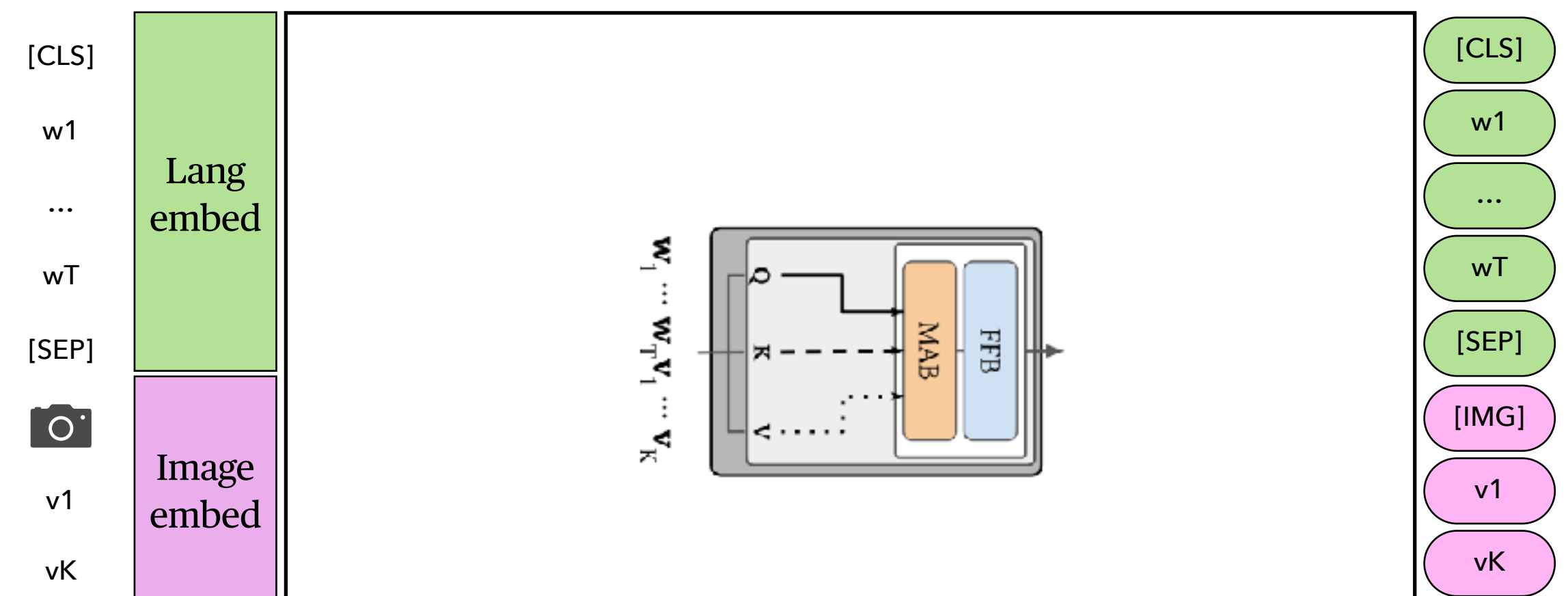
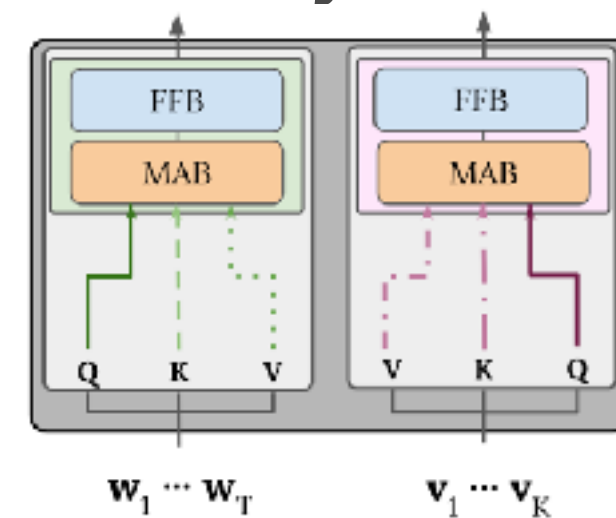
Single-Stream

- Concat image–text inputs



Dual-Stream

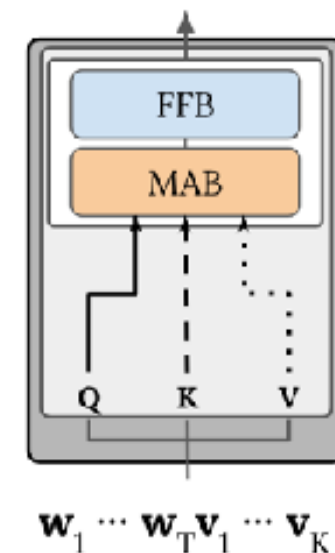
- Image and text independently
- Cross-modal layers
 - Intra-modal



Single- & Dual-Stream Architectures

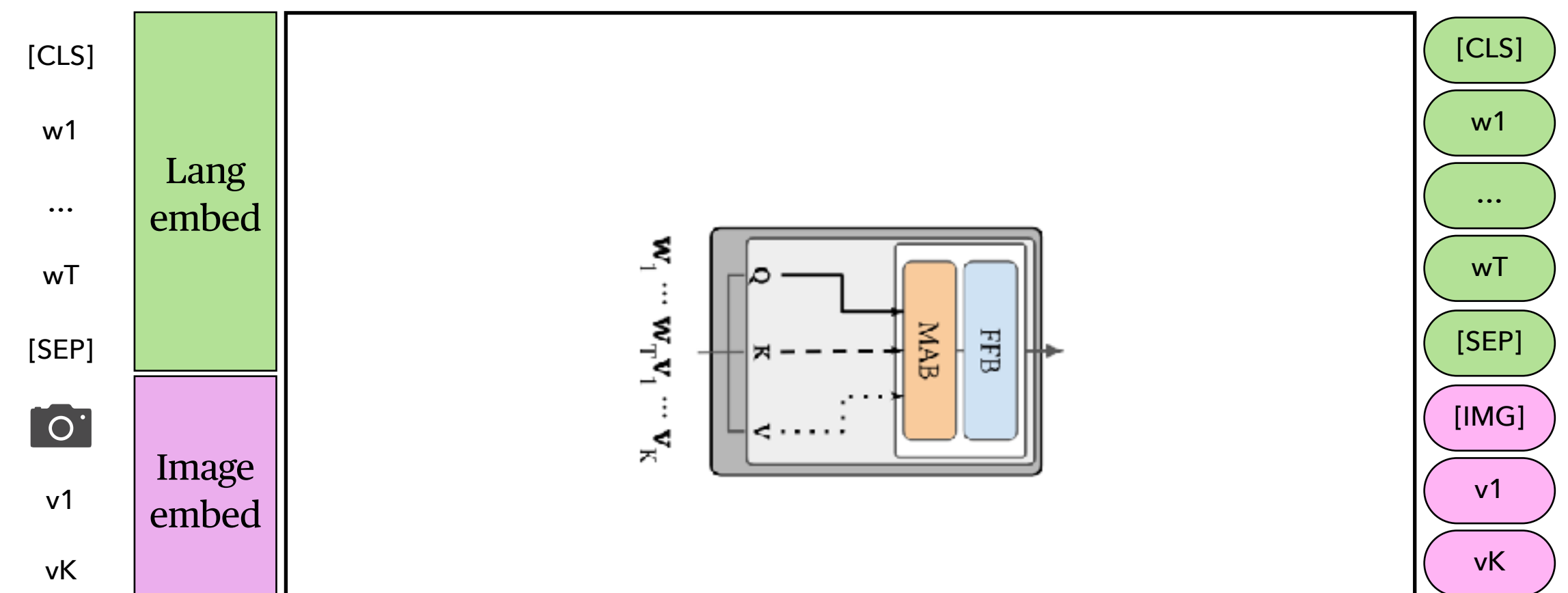
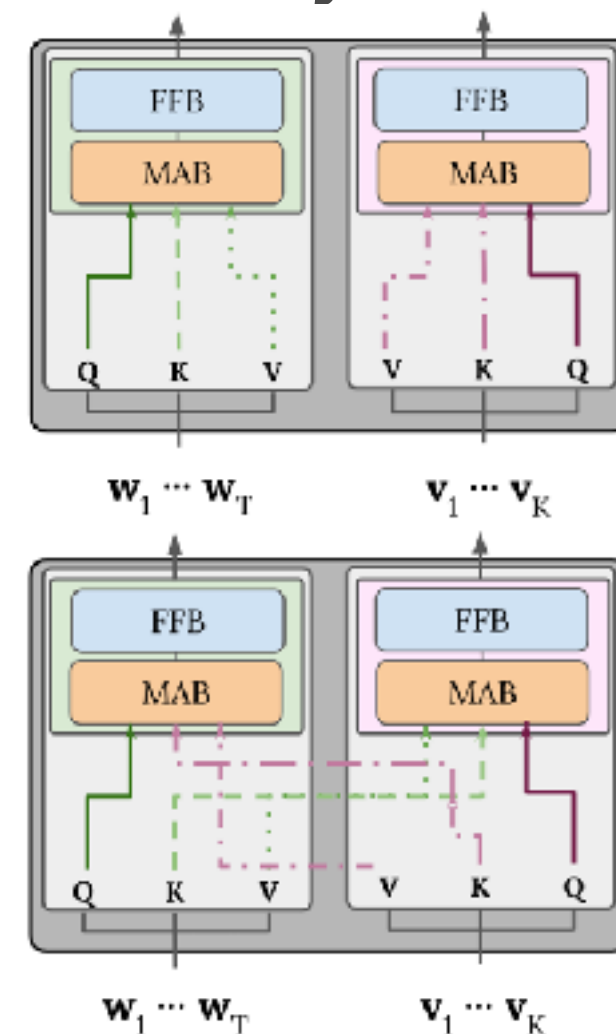
Single-Stream

- Concat image–text inputs



Dual-Stream

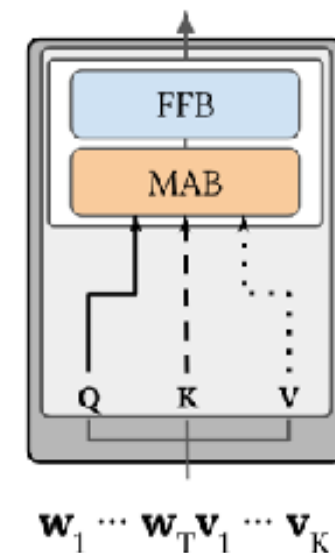
- Image and text independently
- Cross-modal layers
 - Intra-modal
 - Inter-modal



Single- & Dual-Stream Architectures

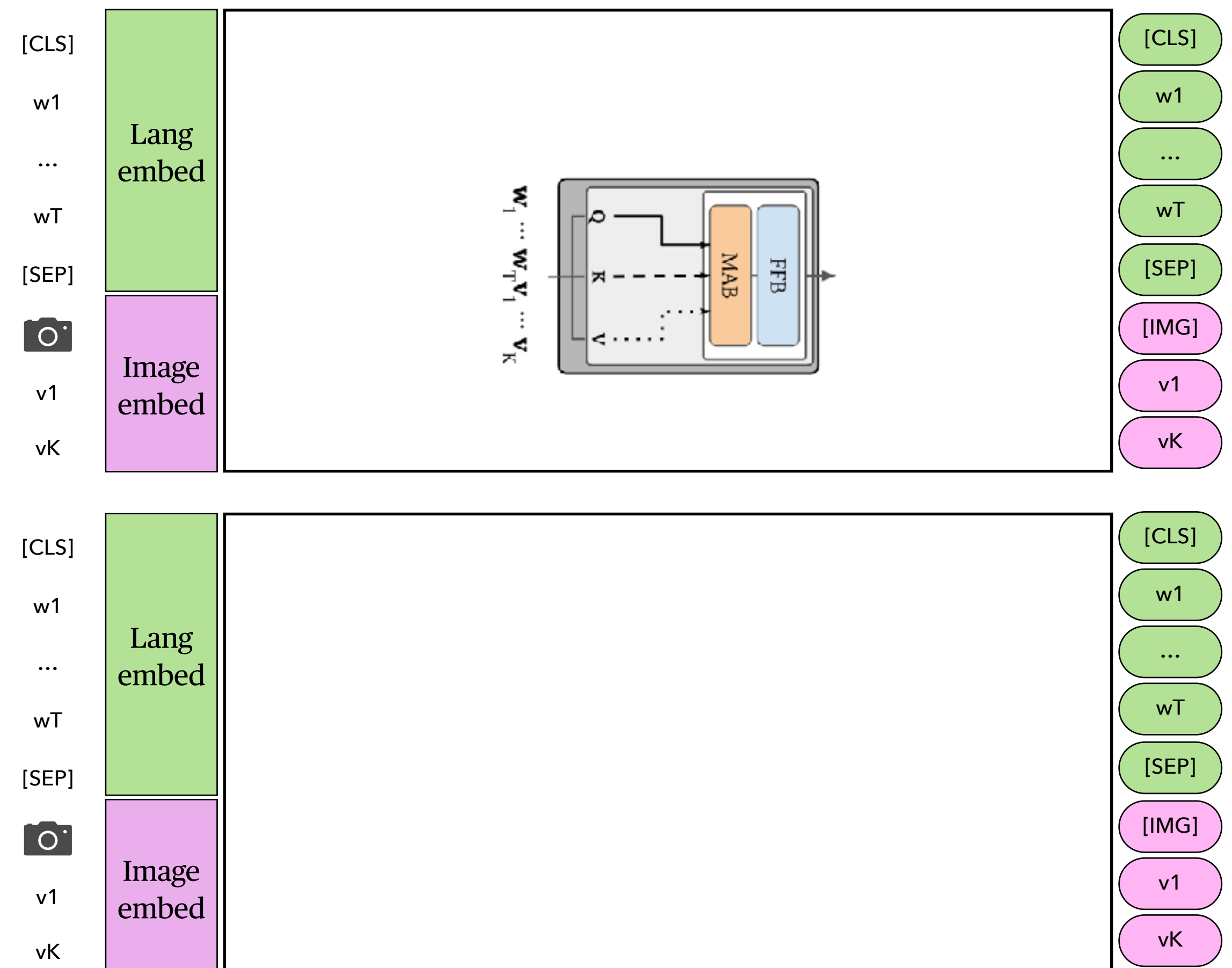
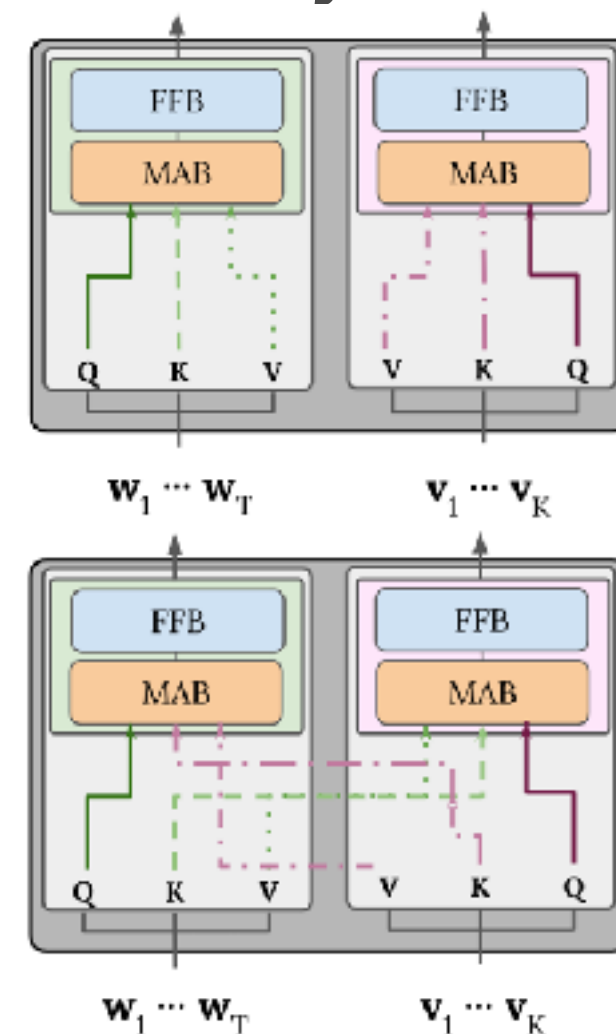
Single-Stream

- Concat image–text inputs



Dual-Stream

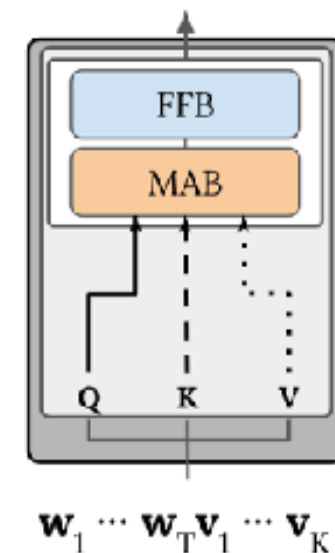
- Image and text independently
- Cross-modal layers
 - Intra-modal
 - Inter-modal



Single- & Dual-Stream Architectures

Single-Stream

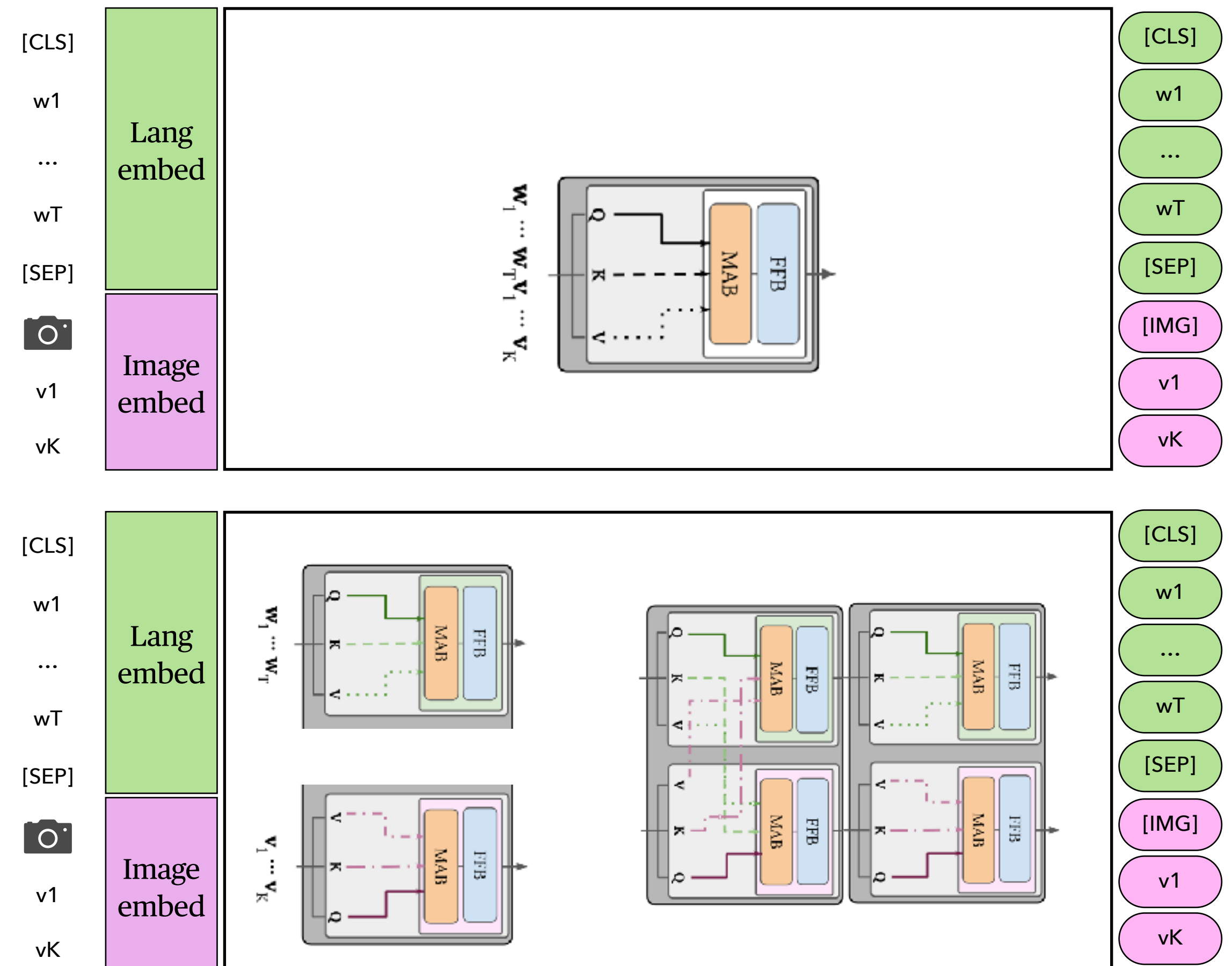
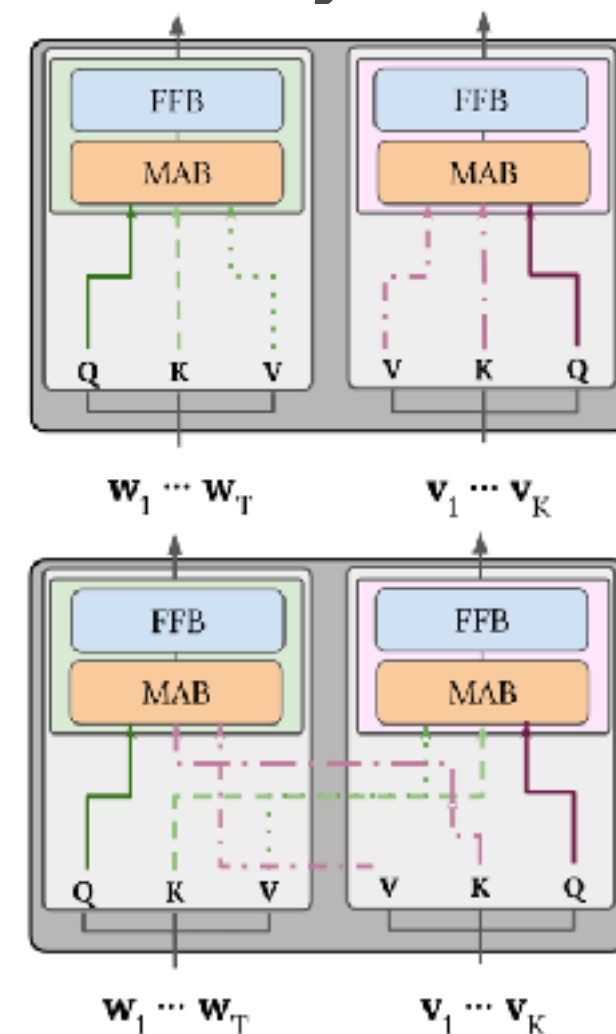
- Concat image–text inputs



Dual-Stream

- Image and text independently
- Cross-modal layers

- Intra-modal
- Inter-modal

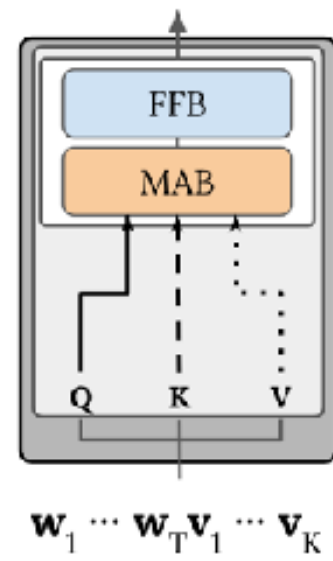


A Unified Framework

A Unified Framework

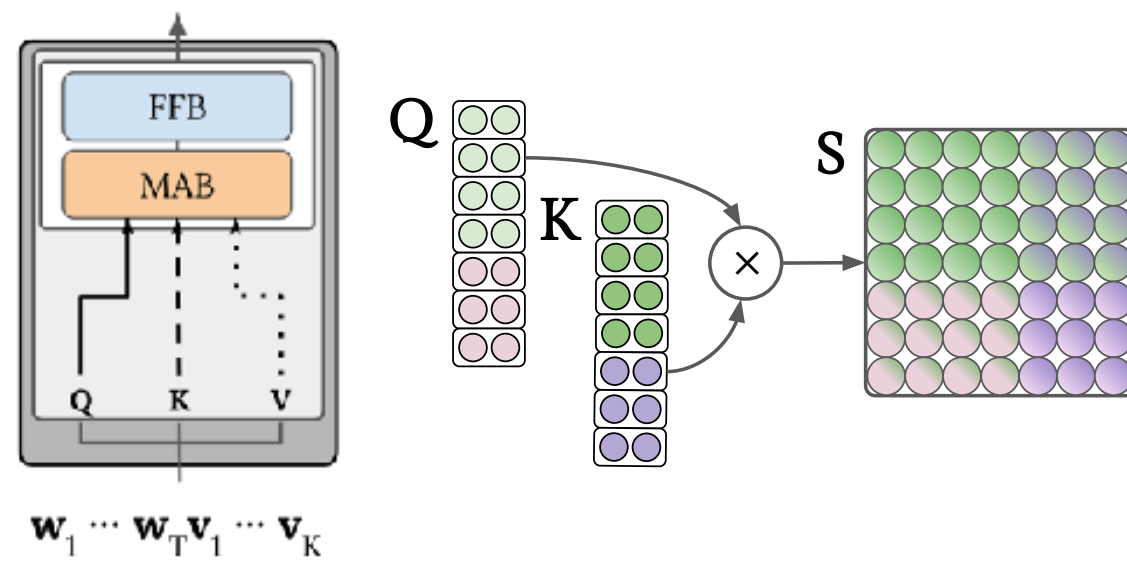
A Unified Framework

Single-Stream Attention



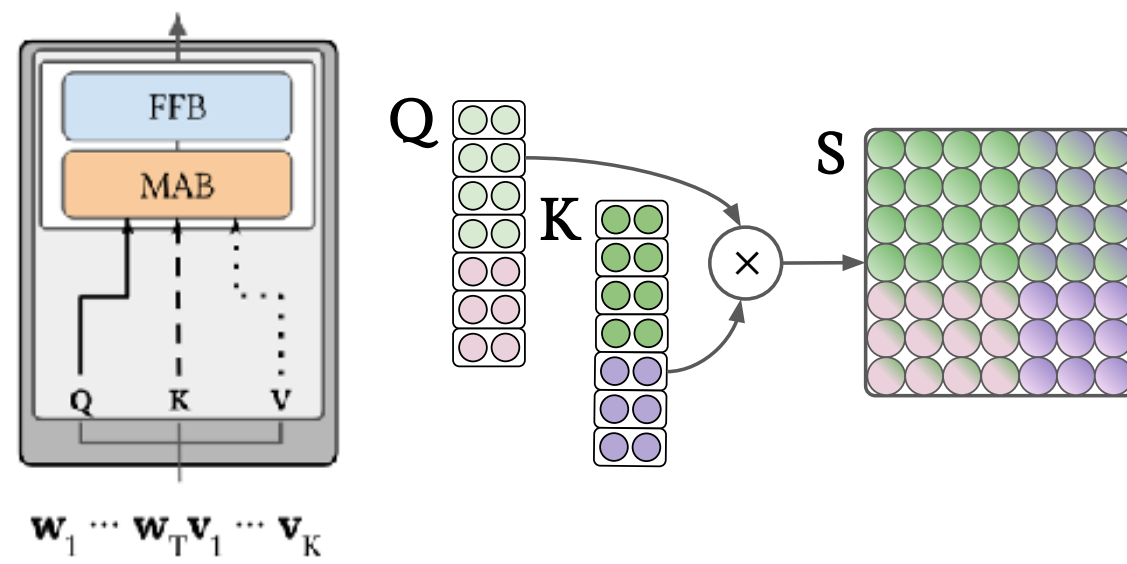
A Unified Framework

Single-Stream Attention

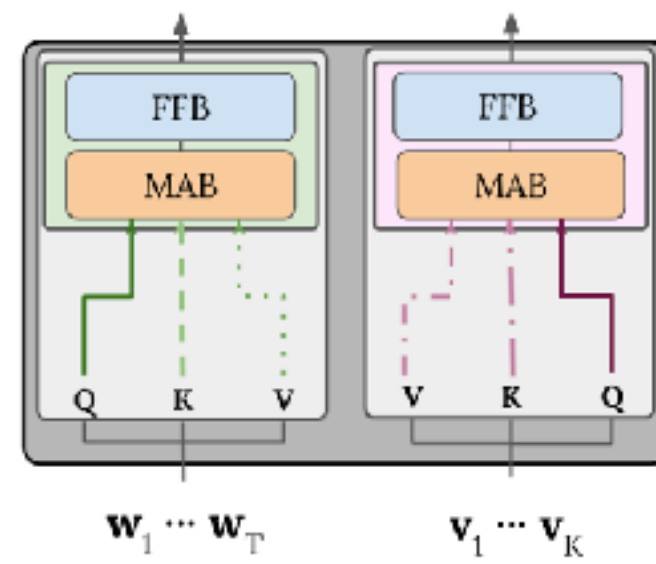


A Unified Framework

Single-Stream Attention

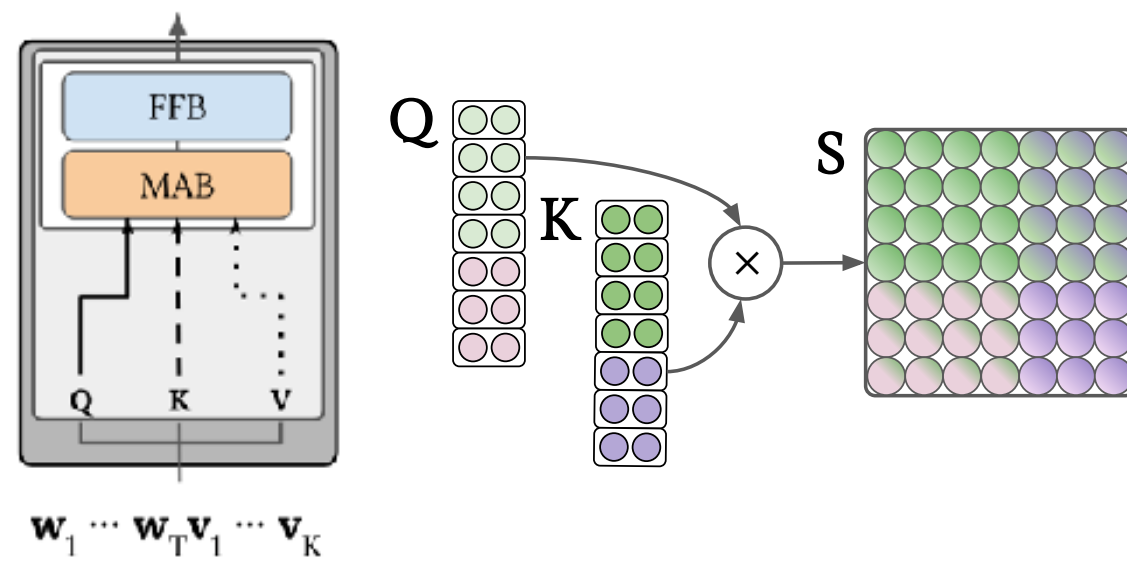


Dual-Stream Intra-modal Attention

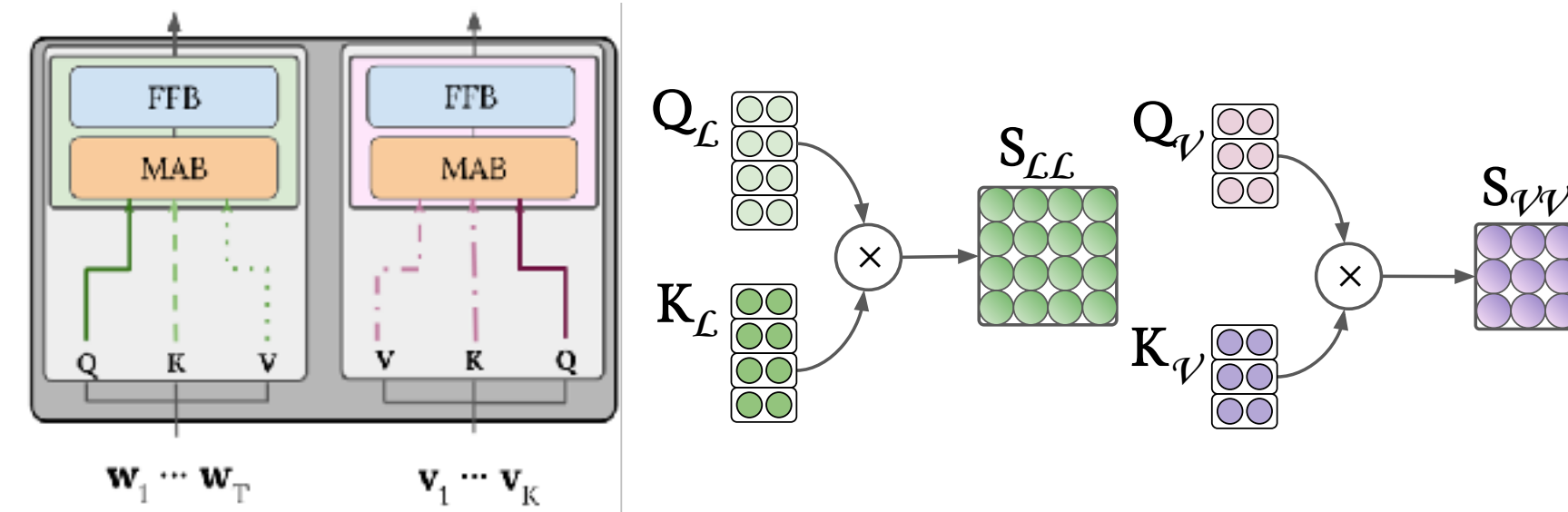


A Unified Framework

Single-Stream Attention

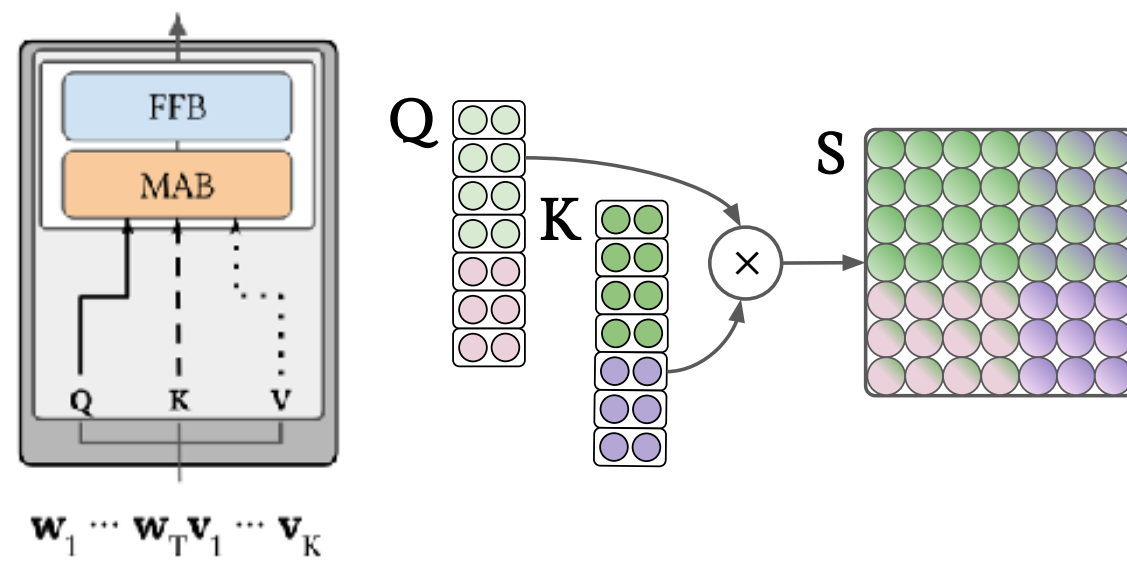


Dual-Stream Intra-modal Attention

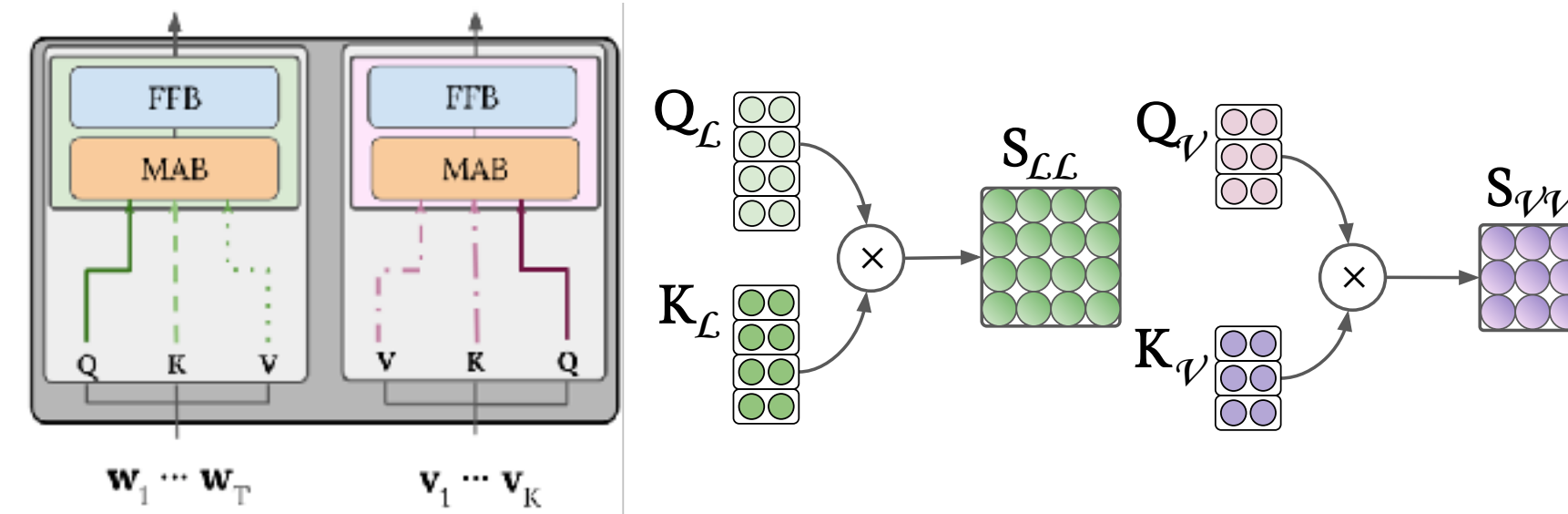


A Unified Framework

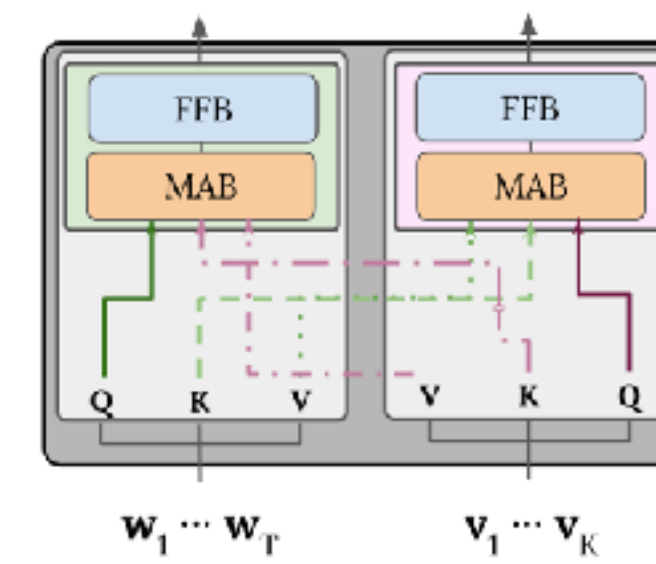
Single-Stream Attention



Dual-Stream Intra-modal Attention

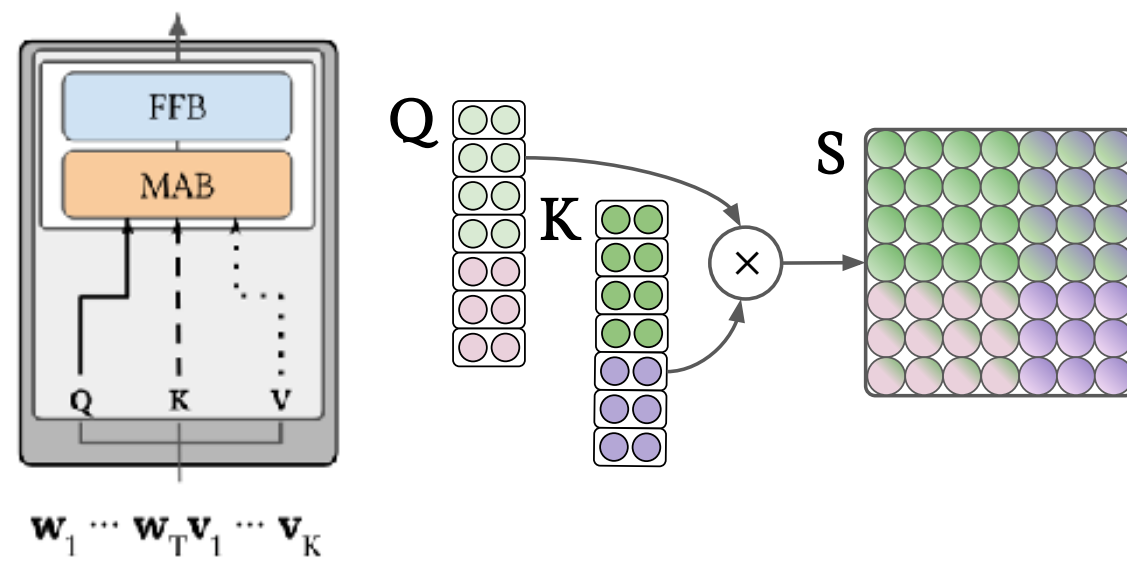


Dual-Stream Inter-modal Attention

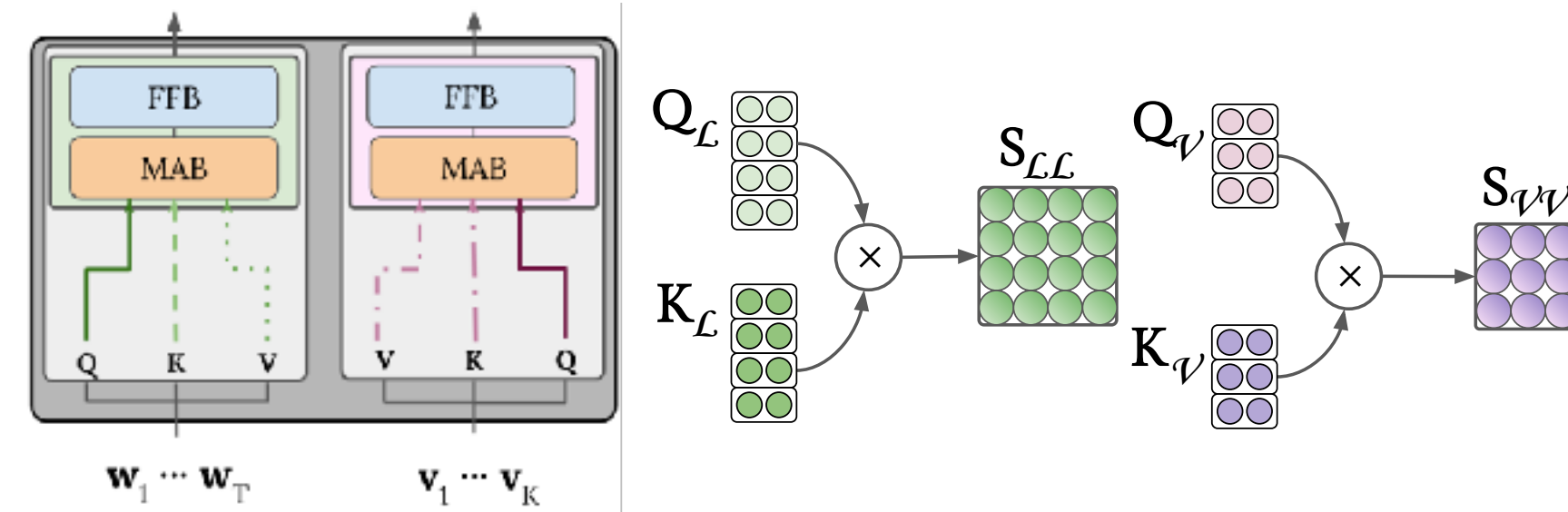


A Unified Framework

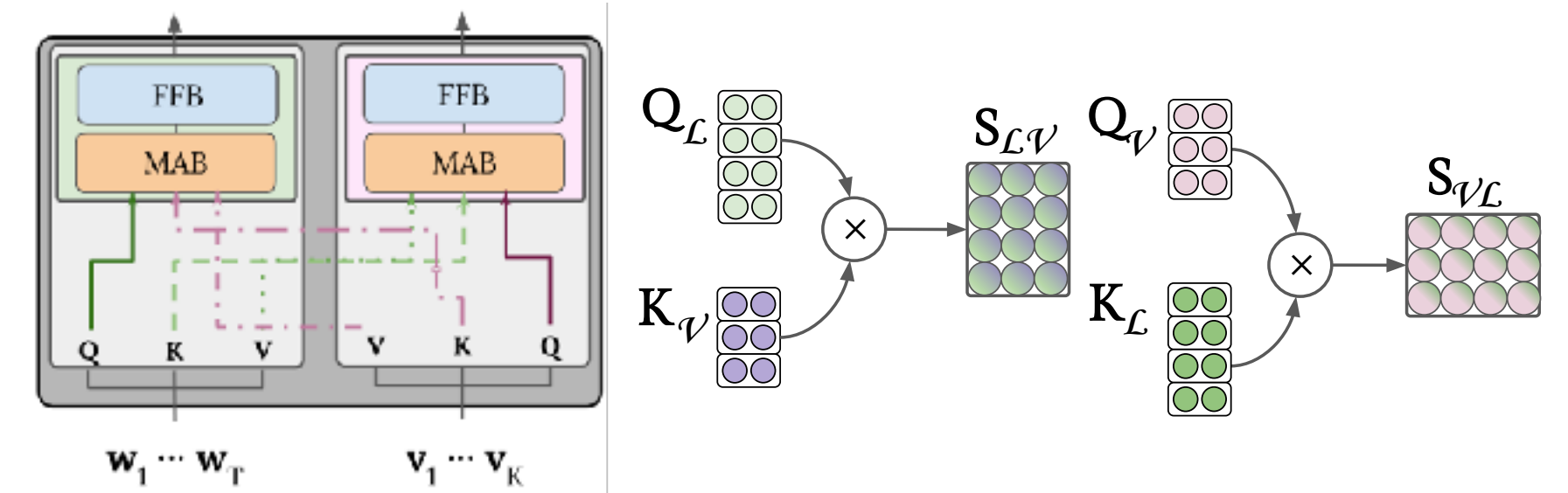
Single-Stream Attention



Dual-Stream Intra-modal Attention

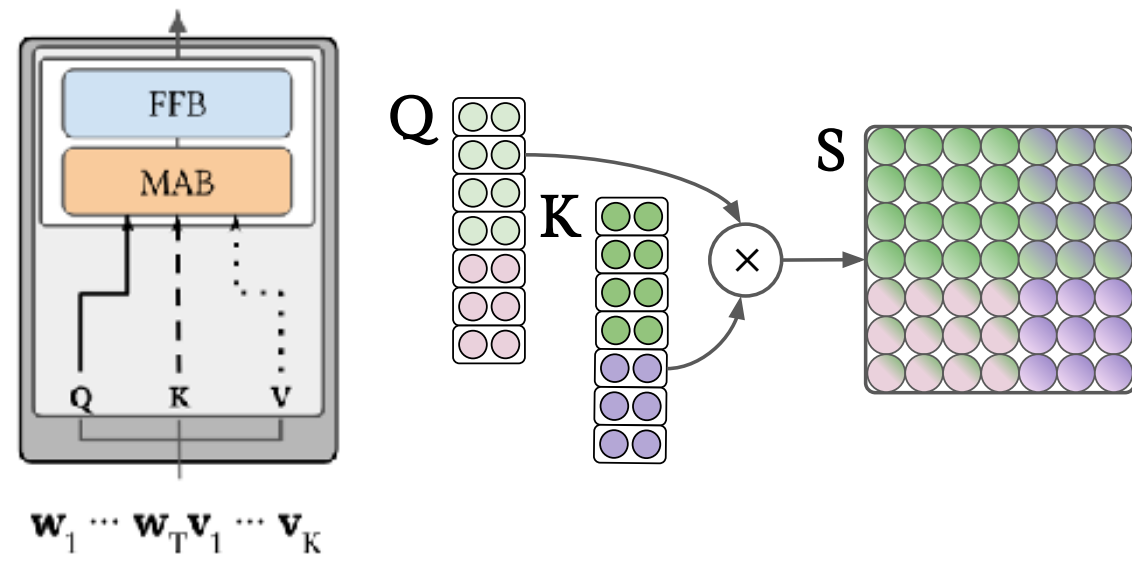


Dual-Stream Inter-modal Attention

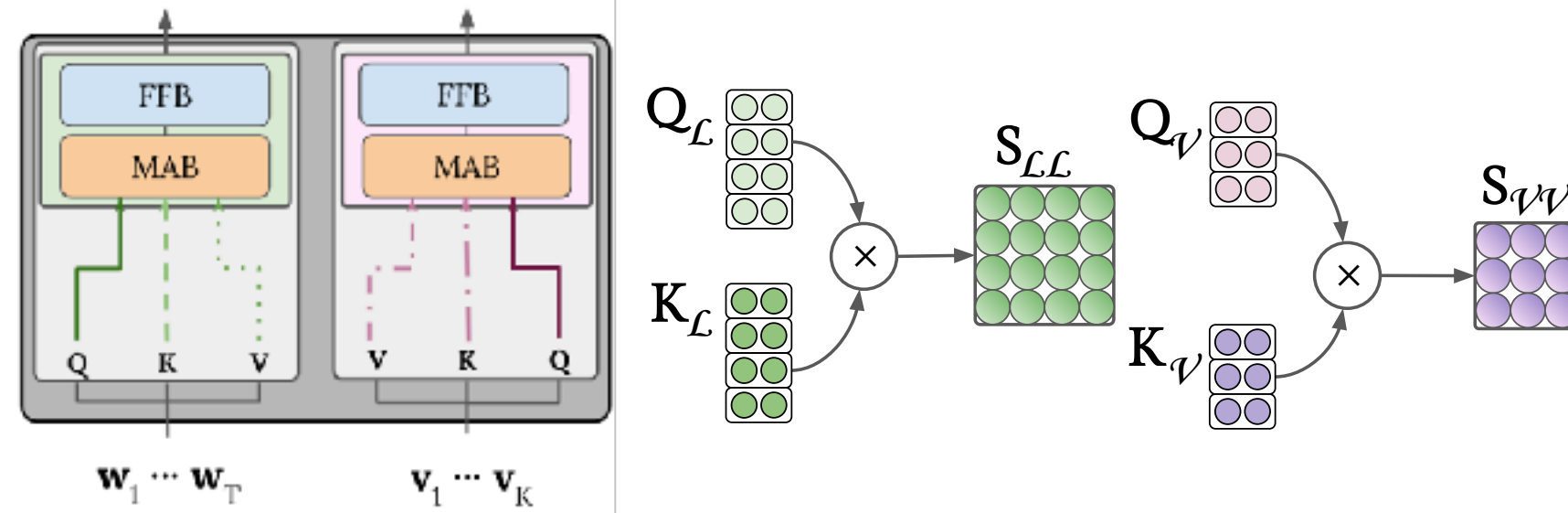


A Unified Framework

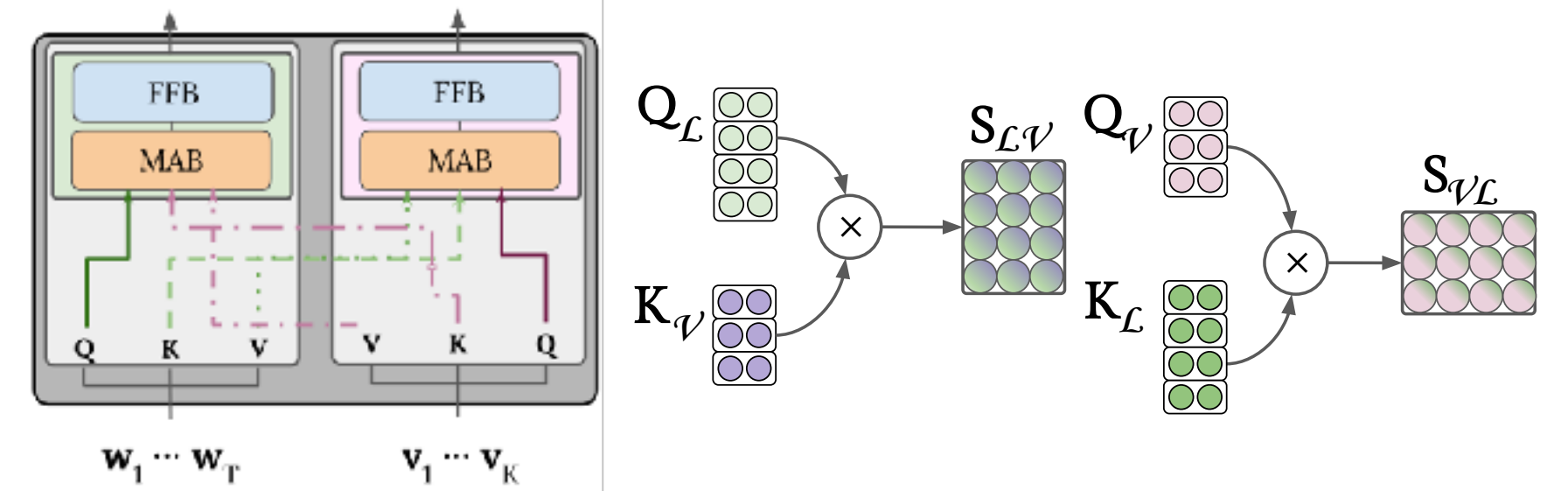
Single-Stream Attention



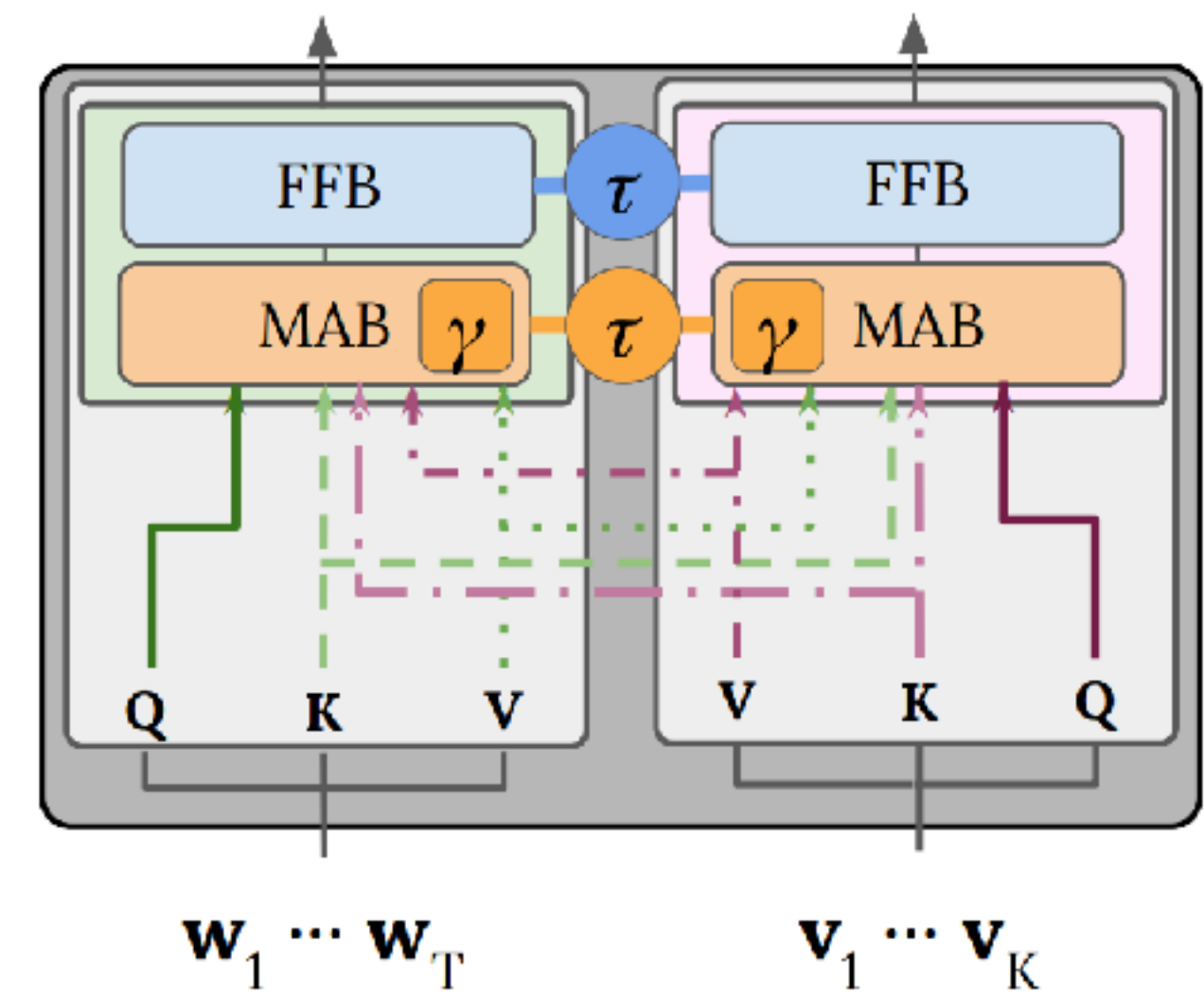
Dual-Stream Intra-modal Attention



Dual-Stream Inter-modal Attention

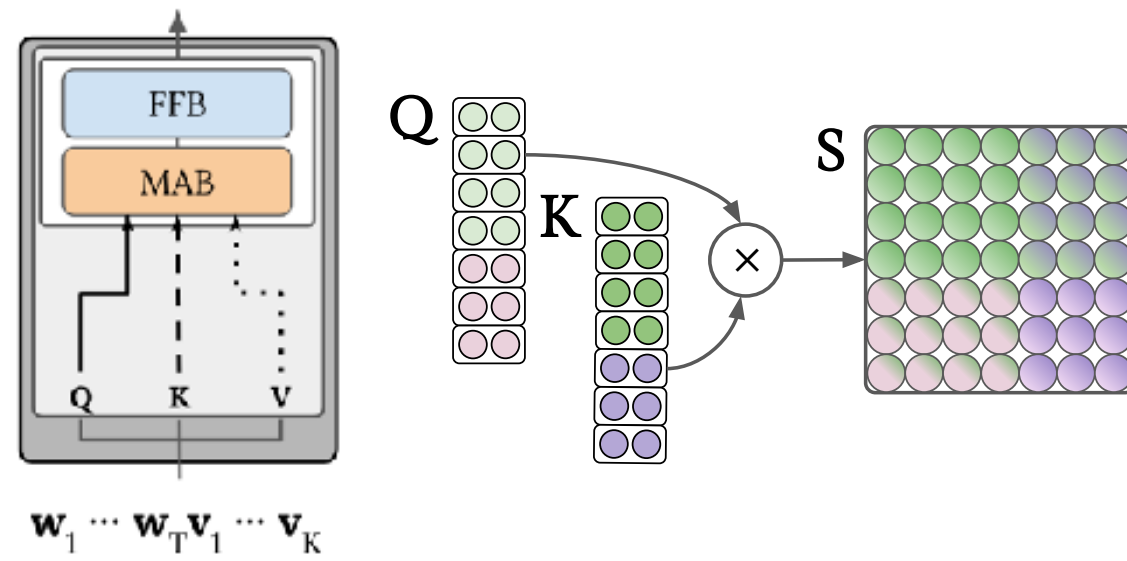


Gated Bimodal Transformer Layer

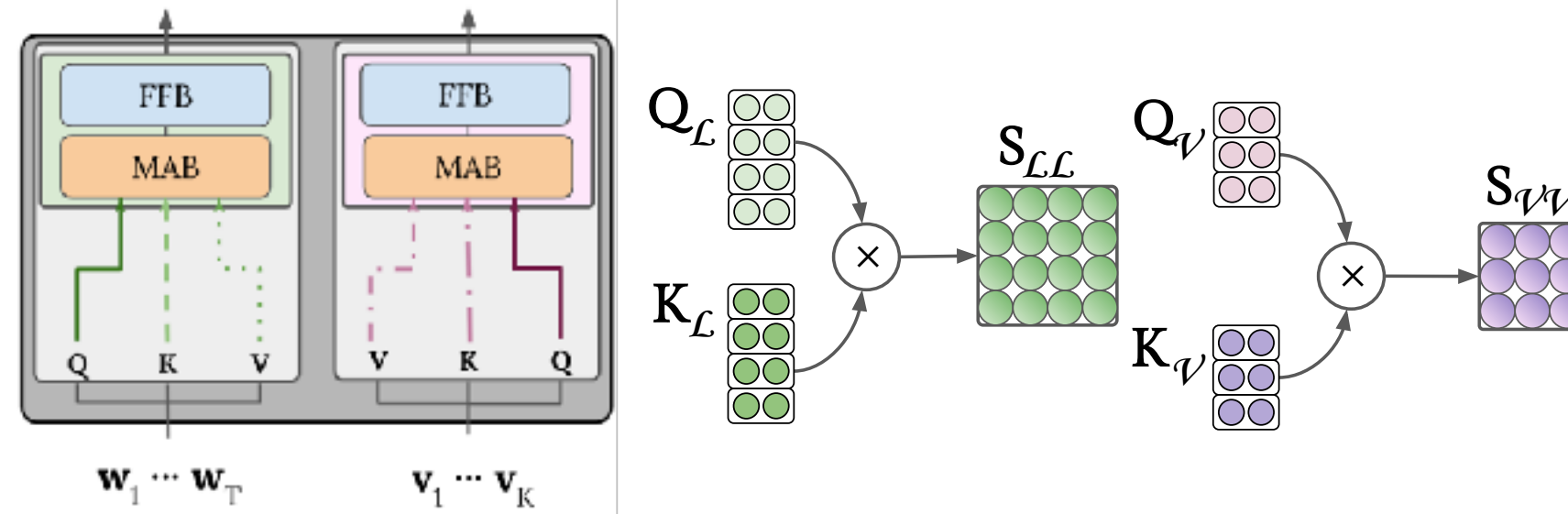


A Unified Framework

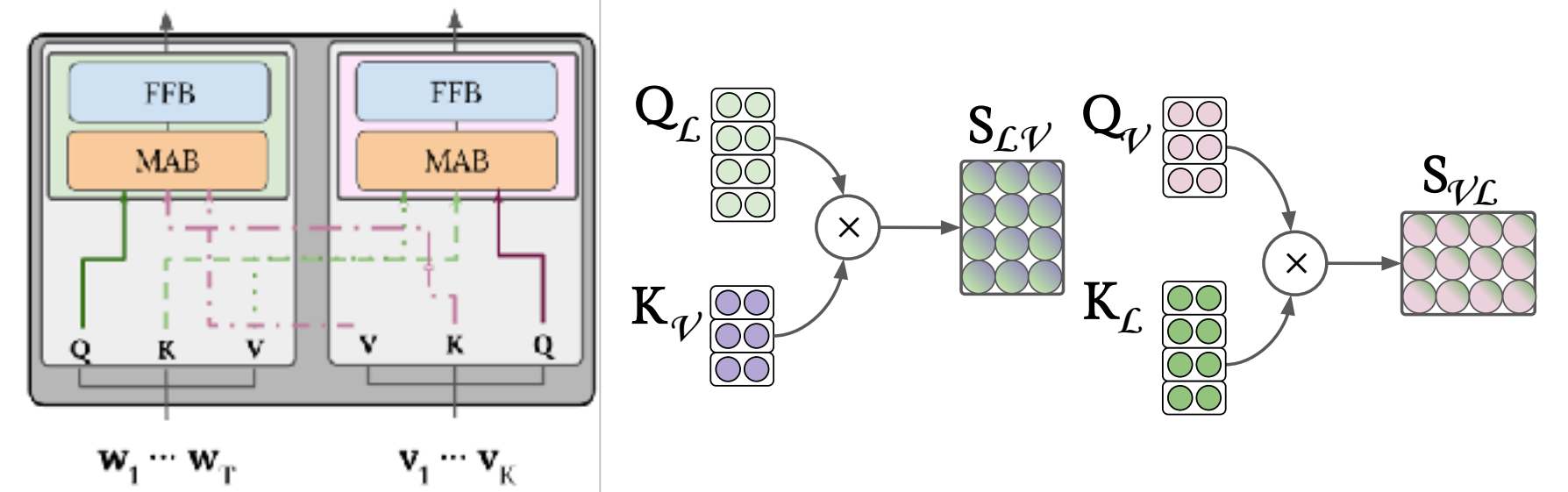
Single-Stream Attention



Dual-Stream Intra-modal Attention

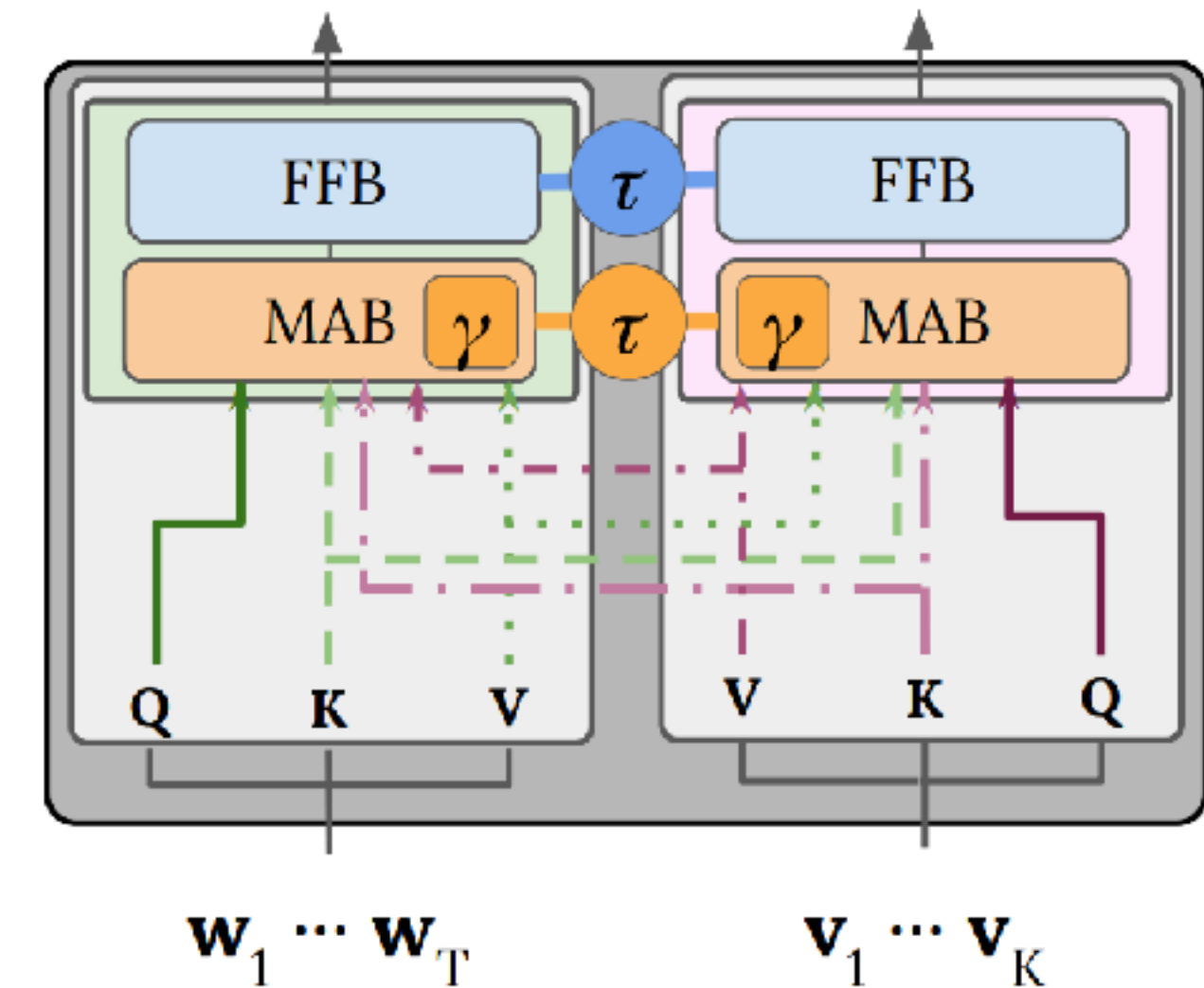


Dual-Stream Inter-modal Attention



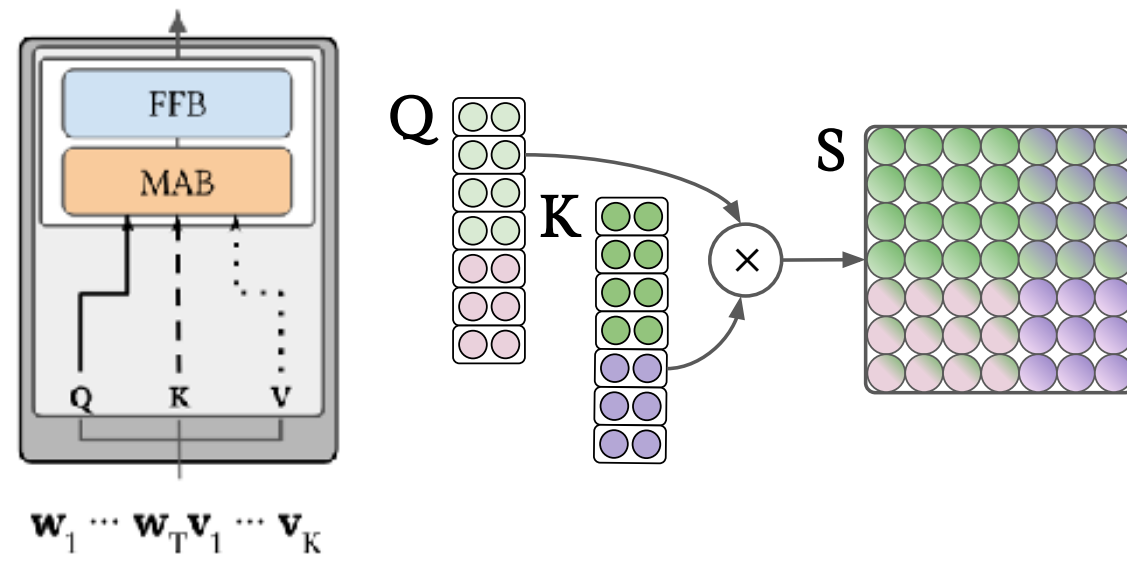
Gated Bimodal Transformer Layer

- Single- and dual-stream layers are special cases

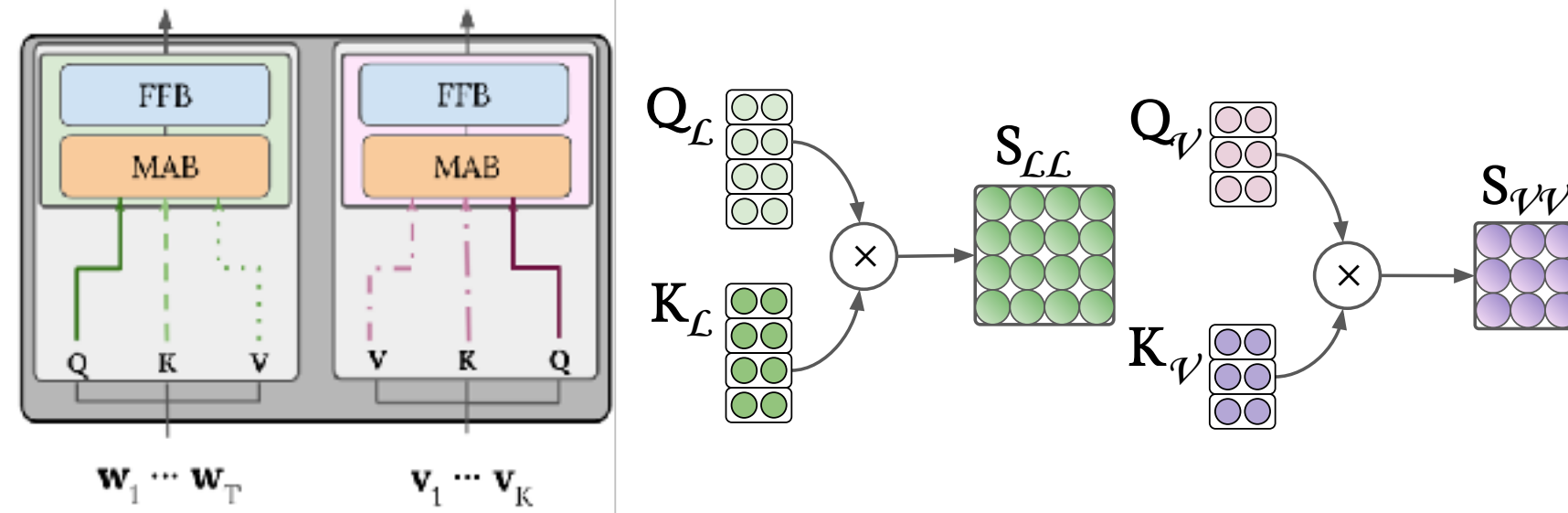


A Unified Framework

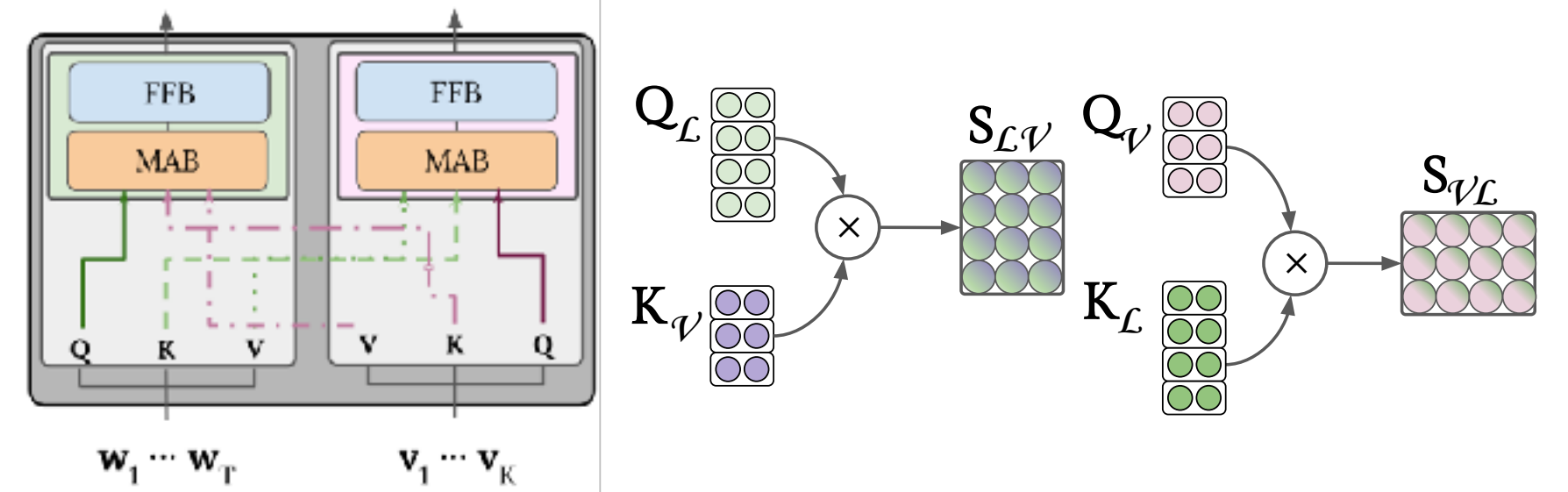
Single-Stream Attention



Dual-Stream Intra-modal Attention

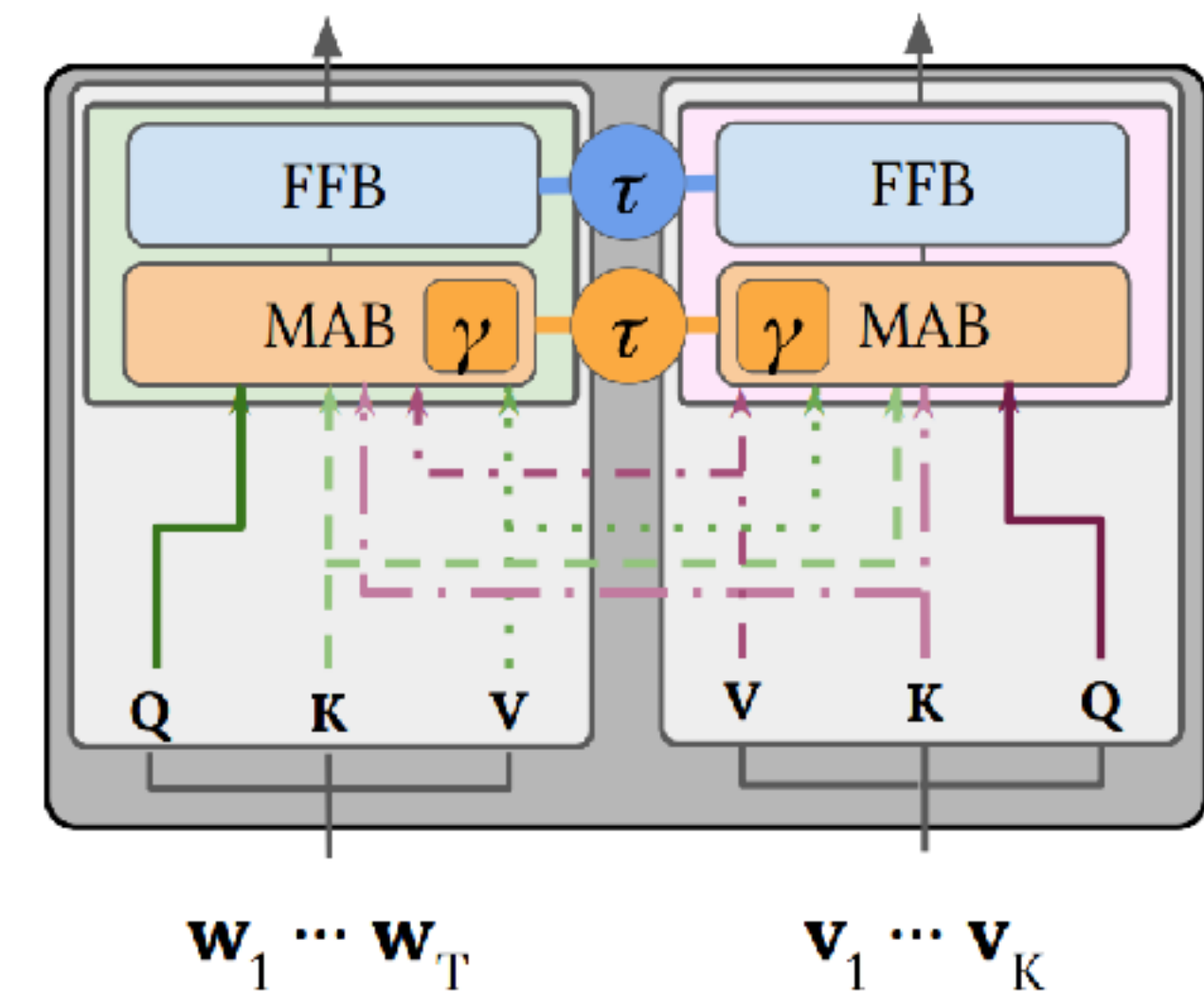


Dual-Stream Inter-modal Attention



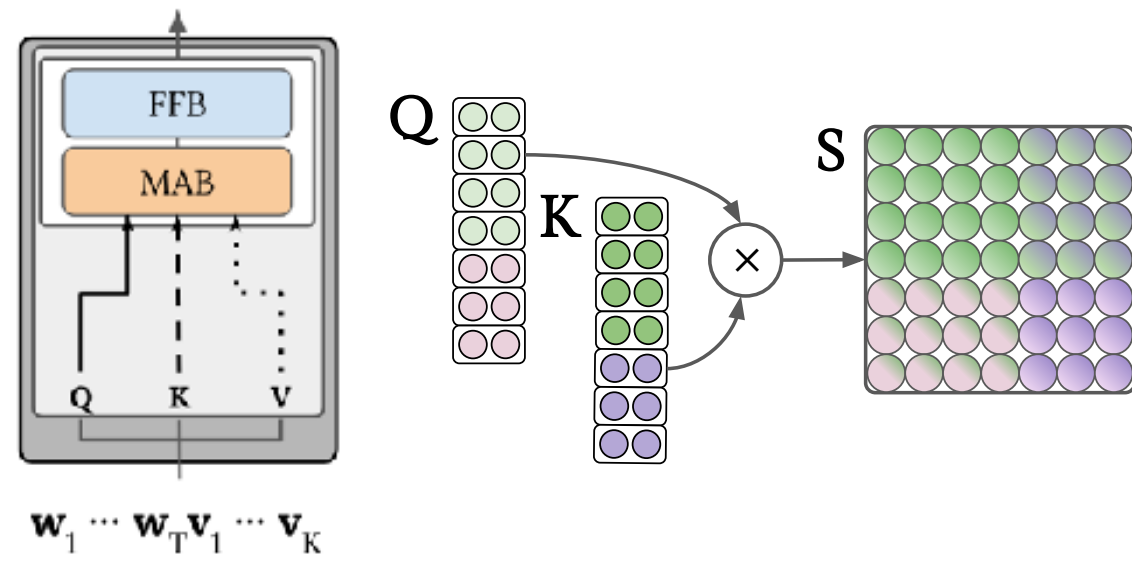
Gated Bimodal Transformer Layer

- Single- and dual-stream layers are special cases
- Takes a set of fixed binary variables $\{\gamma, \tau\}$

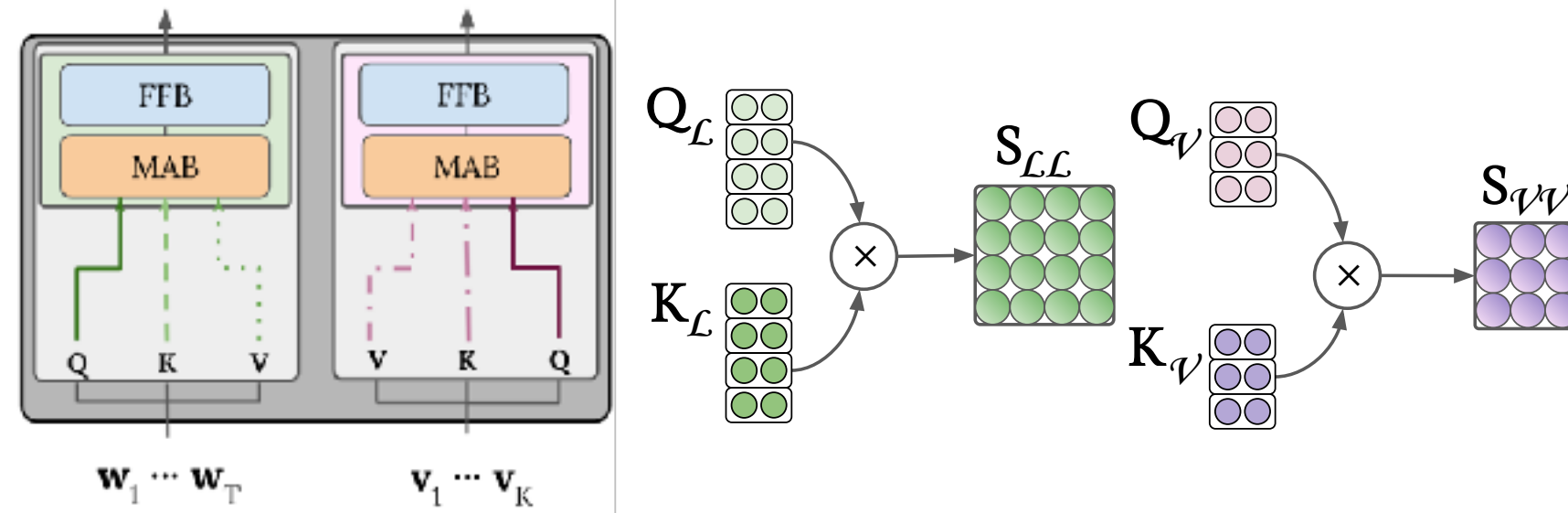


A Unified Framework

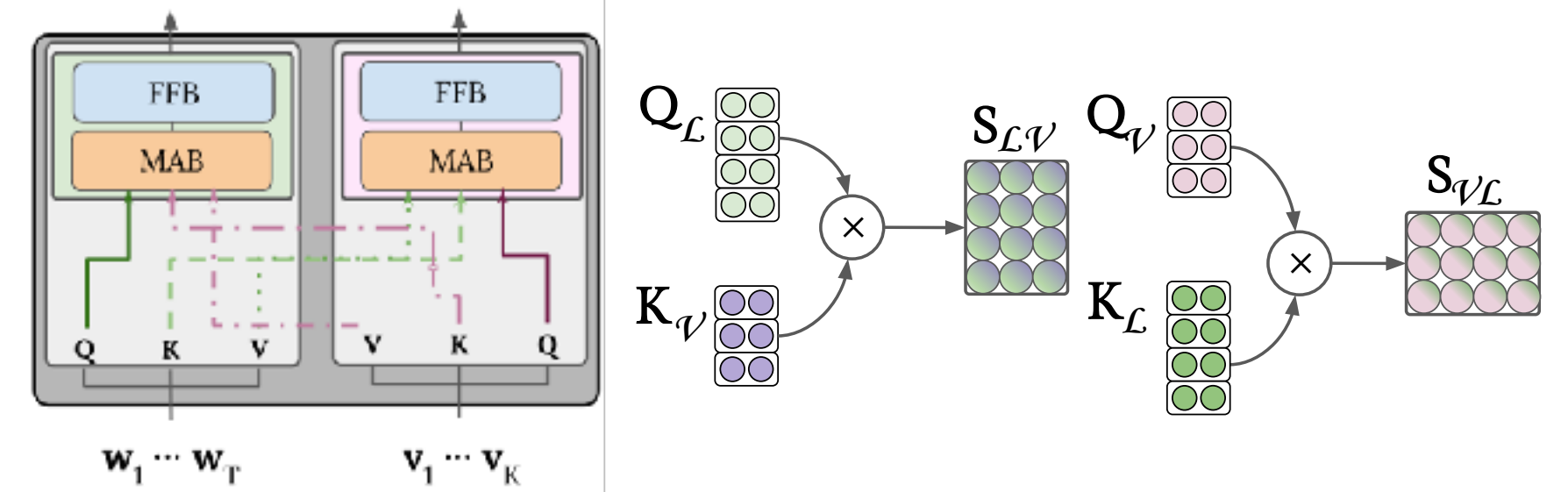
Single-Stream Attention



Dual-Stream Intra-modal Attention

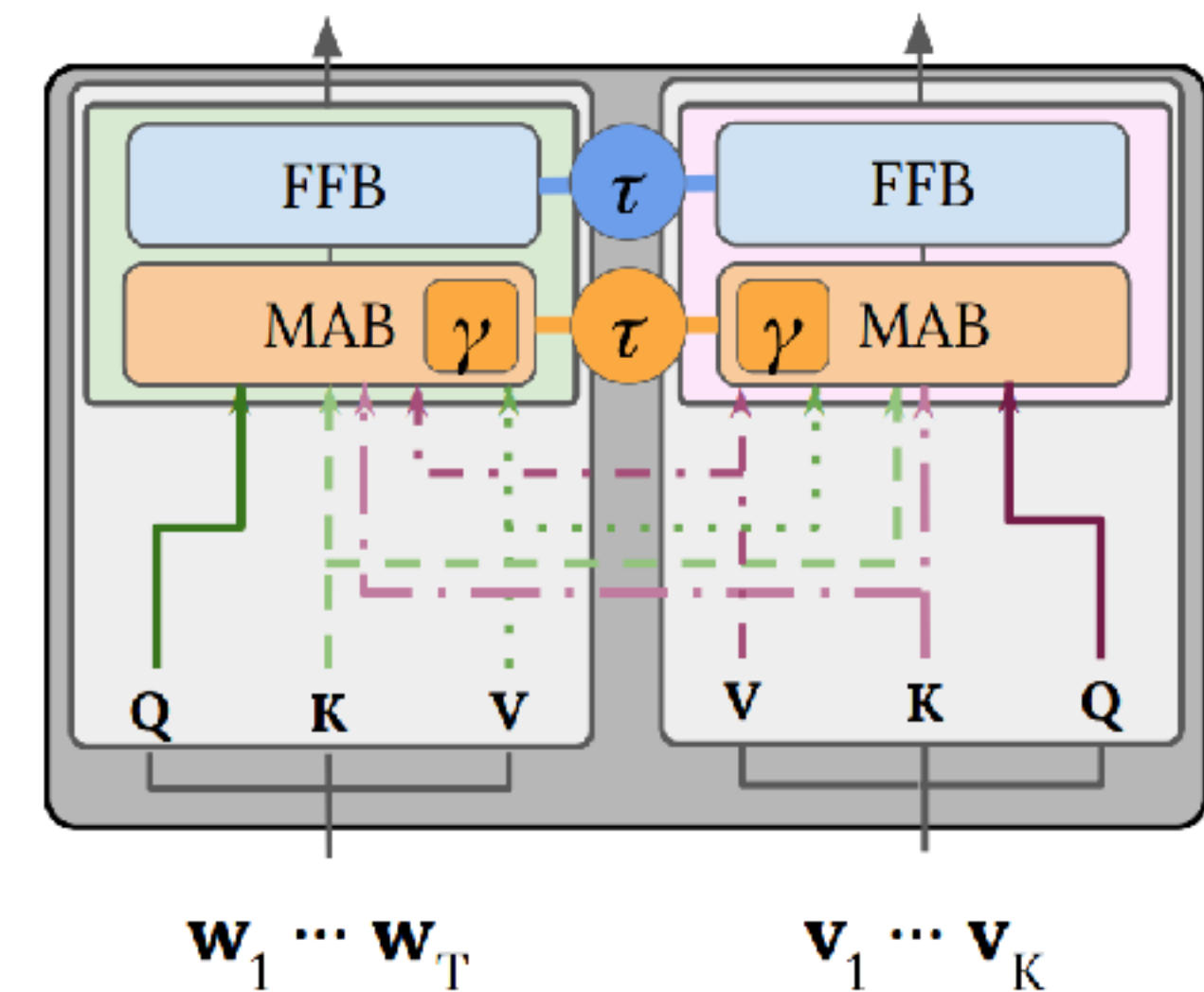


Dual-Stream Inter-modal Attention



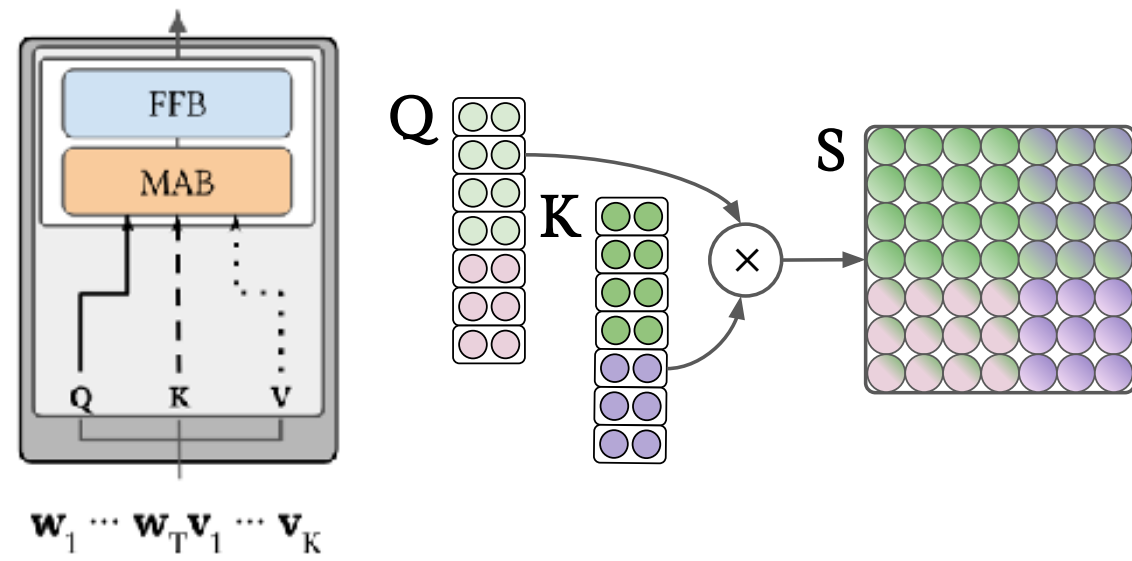
Gated Bimodal Transformer Layer

- Single- and dual-stream layers are special cases
- Takes a set of fixed binary variables $\{\gamma, \tau\}$
 - γ : regulate the cross-modal interactions

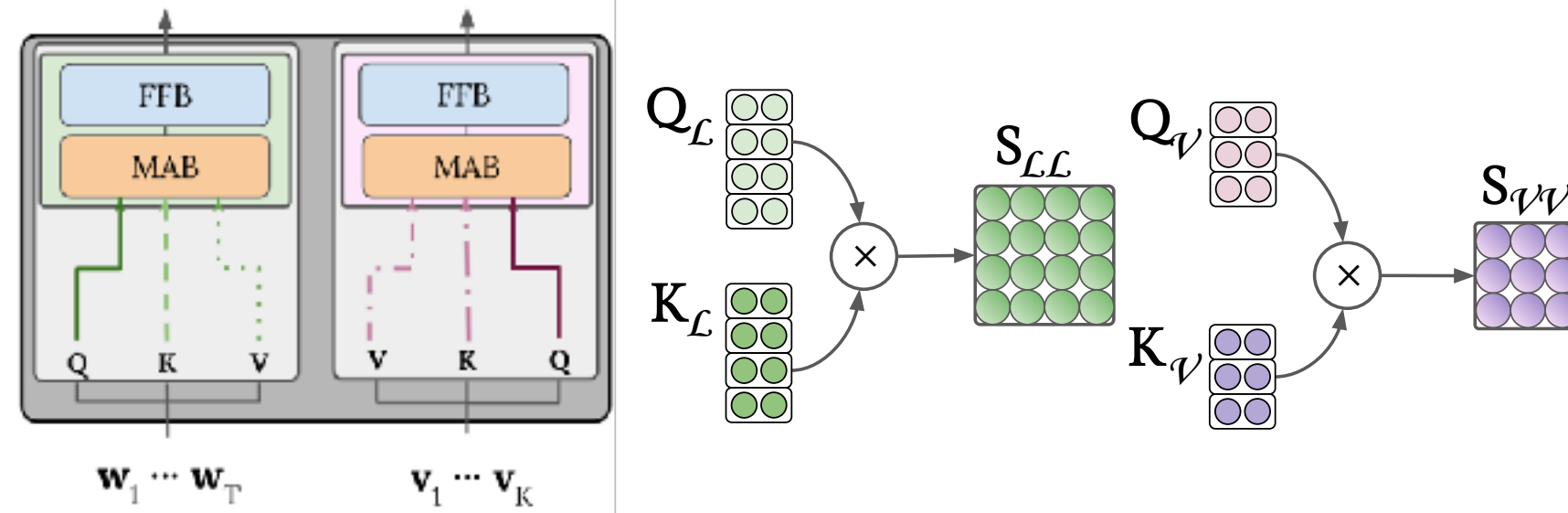


A Unified Framework

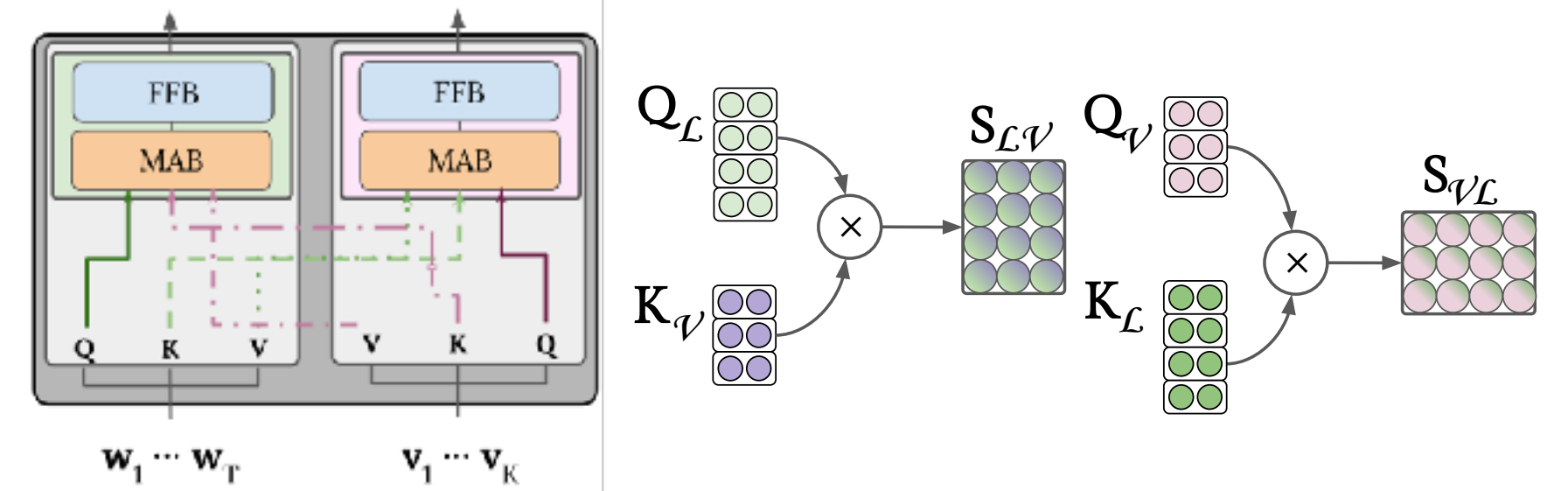
Single-Stream Attention



Dual-Stream Intra-modal Attention

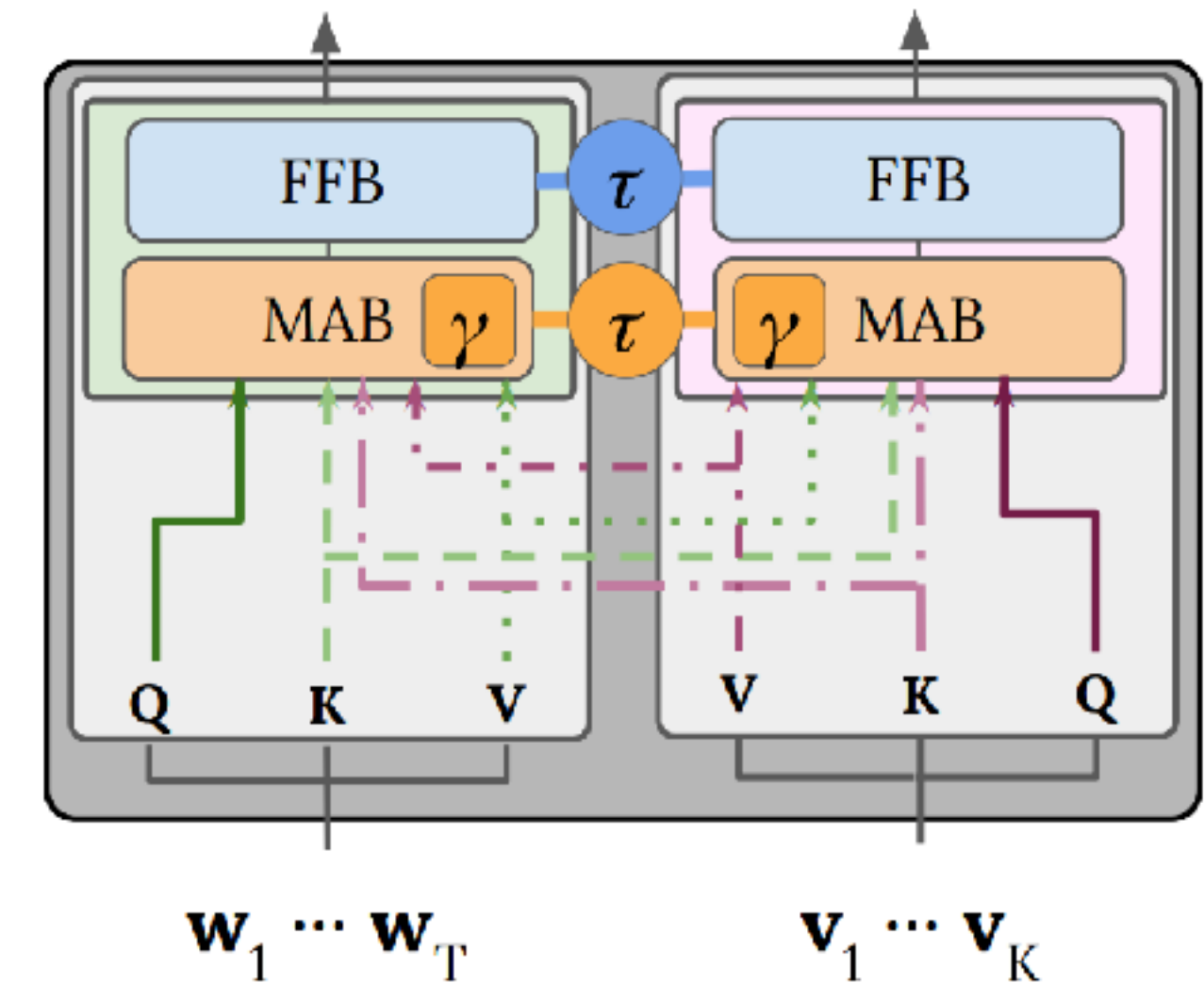


Dual-Stream Inter-modal Attention



Gated Bimodal Transformer Layer

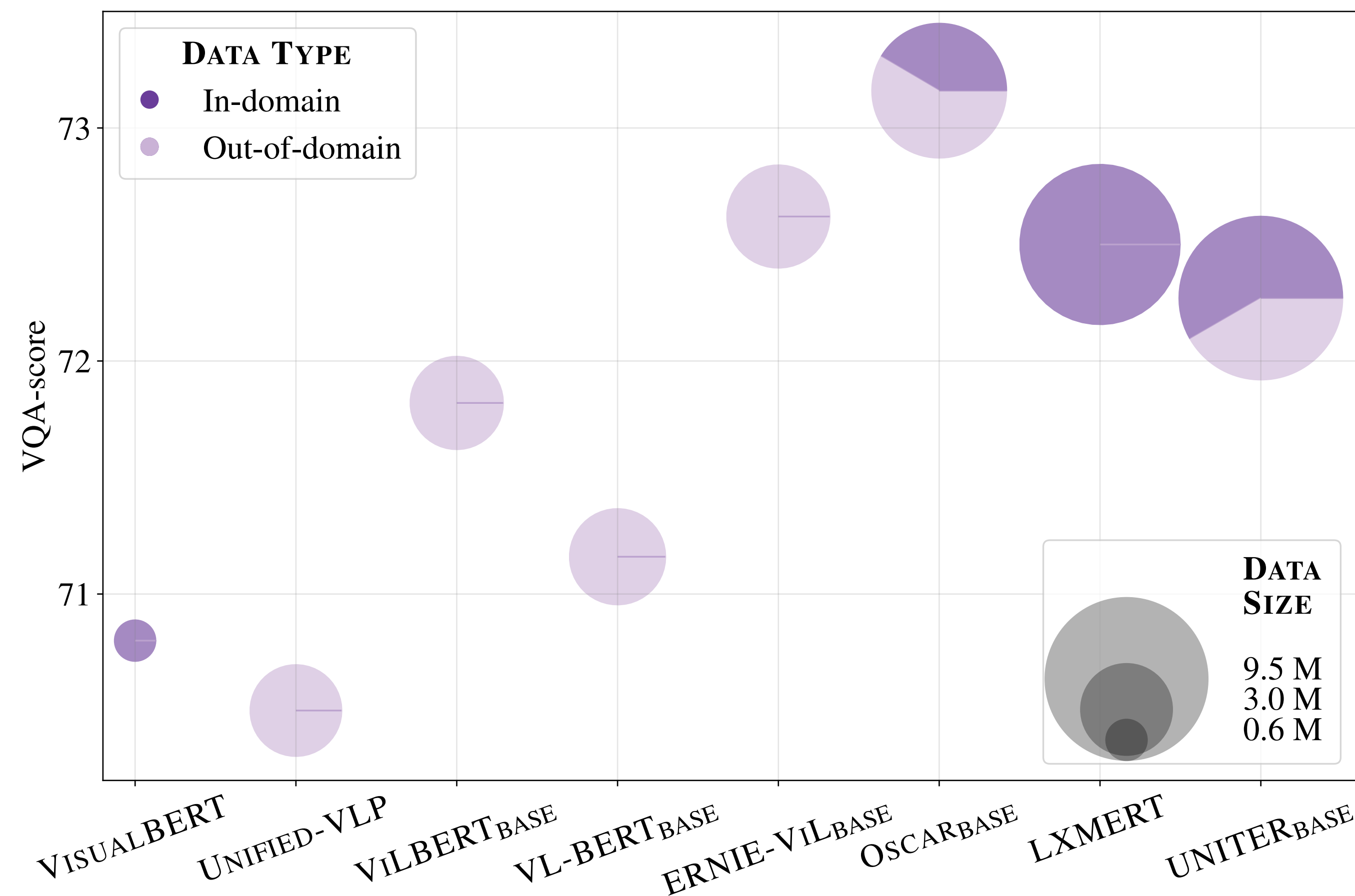
- Single- and dual-stream layers are special cases
- Takes a set of fixed binary variables $\{\gamma, \tau\}$
 - γ : regulate the cross-modal interactions
 - τ : control if parameters are tied between modalities



Downstream Performance

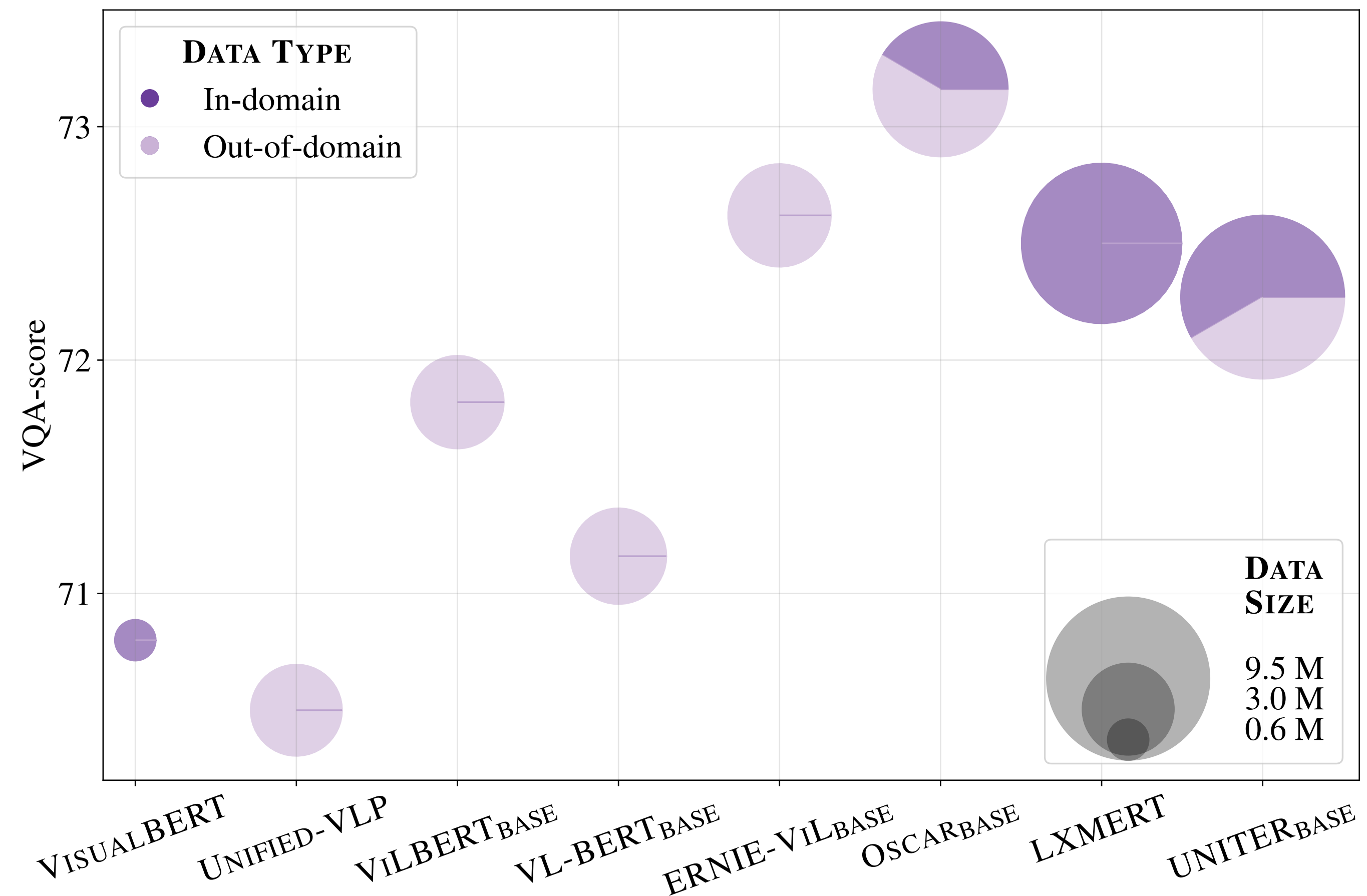
Pretraining Data Confound?

Pretraining Data Confound?



Pretraining Data Confound?

V&L BERTs are pretrained on different amounts and types of data



Controlled Experiments Setup

Controlled Experiments Setup

Models

- *Single-Stream*: VL-BERT (Su+, 2020) VisualBERT (Li+, 2019) and UNITER (Chen+, 2020)
- *Dual-Stream*: ViLBERT (Lu+, 2019) LXMERT (Tan&Bansal, 2019)
- Re-implemented using our unified framework in VOLTA (github.com/e-bug/volta)

Controlled Experiments Setup

Models

- *Single-Stream*: VL-BERT (Su+, 2020) VisualBERT (Li+, 2019) and UNITER (Chen+, 2020)
- *Dual-Stream*: ViLBERT (Lu+, 2019) LXMERT (Tan&Bansal, 2019)
- Re-implemented using our unified framework in VOLTA (github.com/e-bug/volta)

Pretraining data

- Conceptual Captions (2.77M)

Controlled Experiments Setup

Models

- *Single-Stream*: VL-BERT (Su+, 2020) VisualBERT (Li+, 2019) and UNITER (Chen+, 2020)
- *Dual-Stream*: ViLBERT (Lu+, 2019) LXMERT (Tan&Bansal, 2019)
- Re-implemented using our unified framework in VOLTA (github.com/e-bug/volta)

Pretraining data

- Conceptual Captions (2.77M)

Downstream tasks

- VQAv2, RefCOCO+, NLVR2 and Flickr30K

Controlled Experiments Setup

Models

- *Single-Stream*: VL-BERT (Su+, 2020) VisualBERT (Li+, 2019) and UNITER (Chen+, 2020)
- *Dual-Stream*: ViLBERT (Lu+, 2019) LXMERT (Tan&Bansal, 2019)
- Re-implemented using our unified framework in VOLTA (github.com/e-bug/volta)

Pretraining data

- Conceptual Captions (2.77M)

Downstream tasks

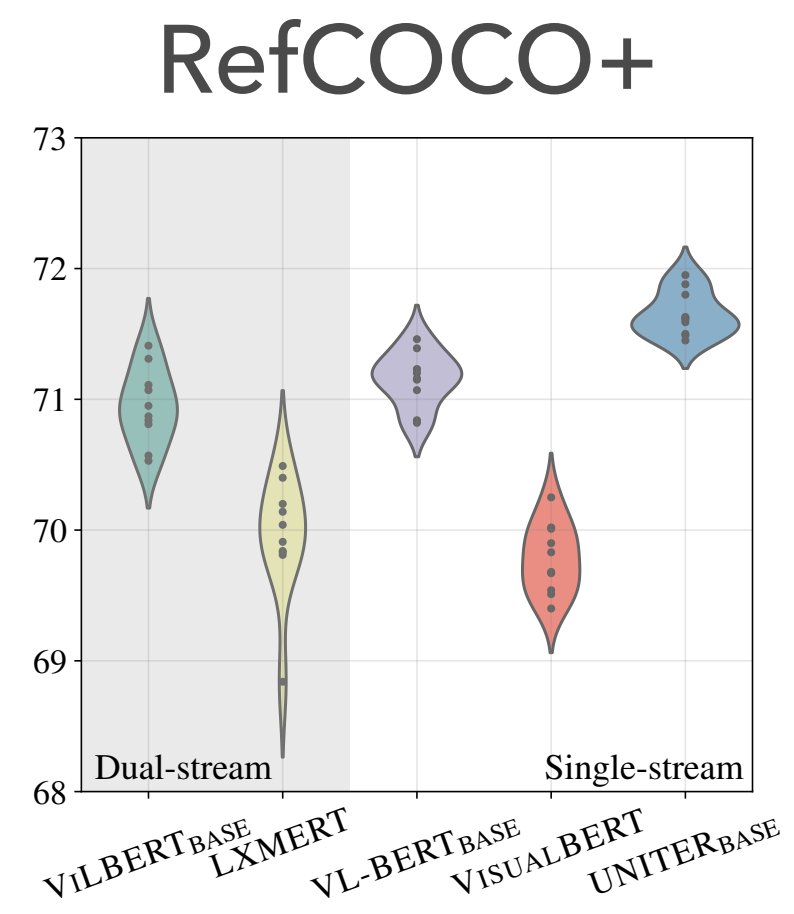
- VQAv2, RefCOCO+, NLVR2 and Flickr30K

Controlled setup

- Same data & fixed set of hyperparameters
- Multiple pretraining and fine-tuning seeds

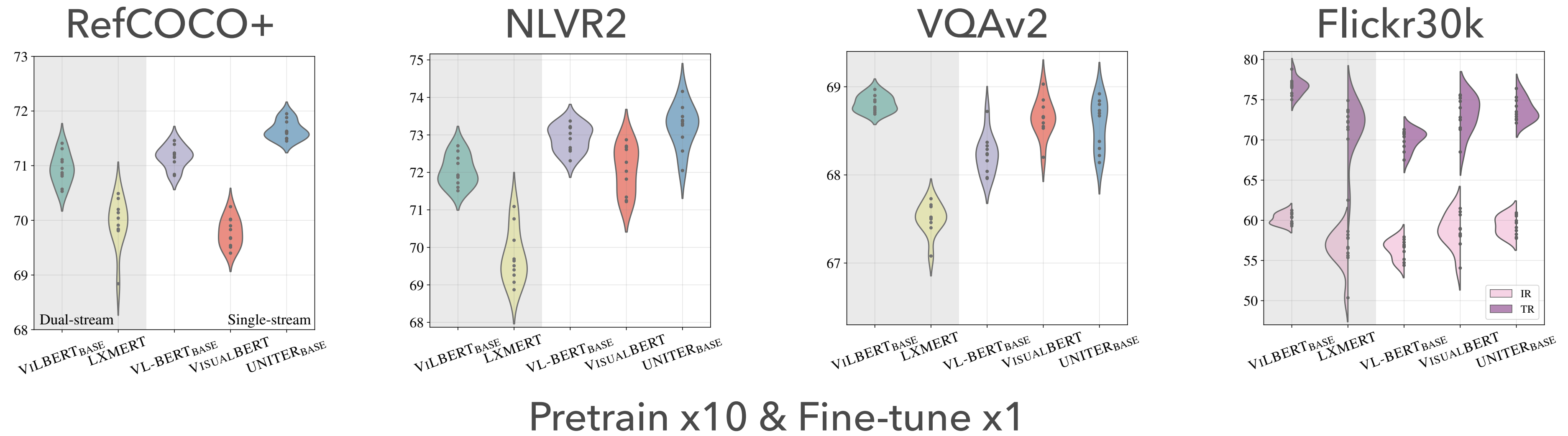
Controlled Experiments Results

Controlled Experiments Results

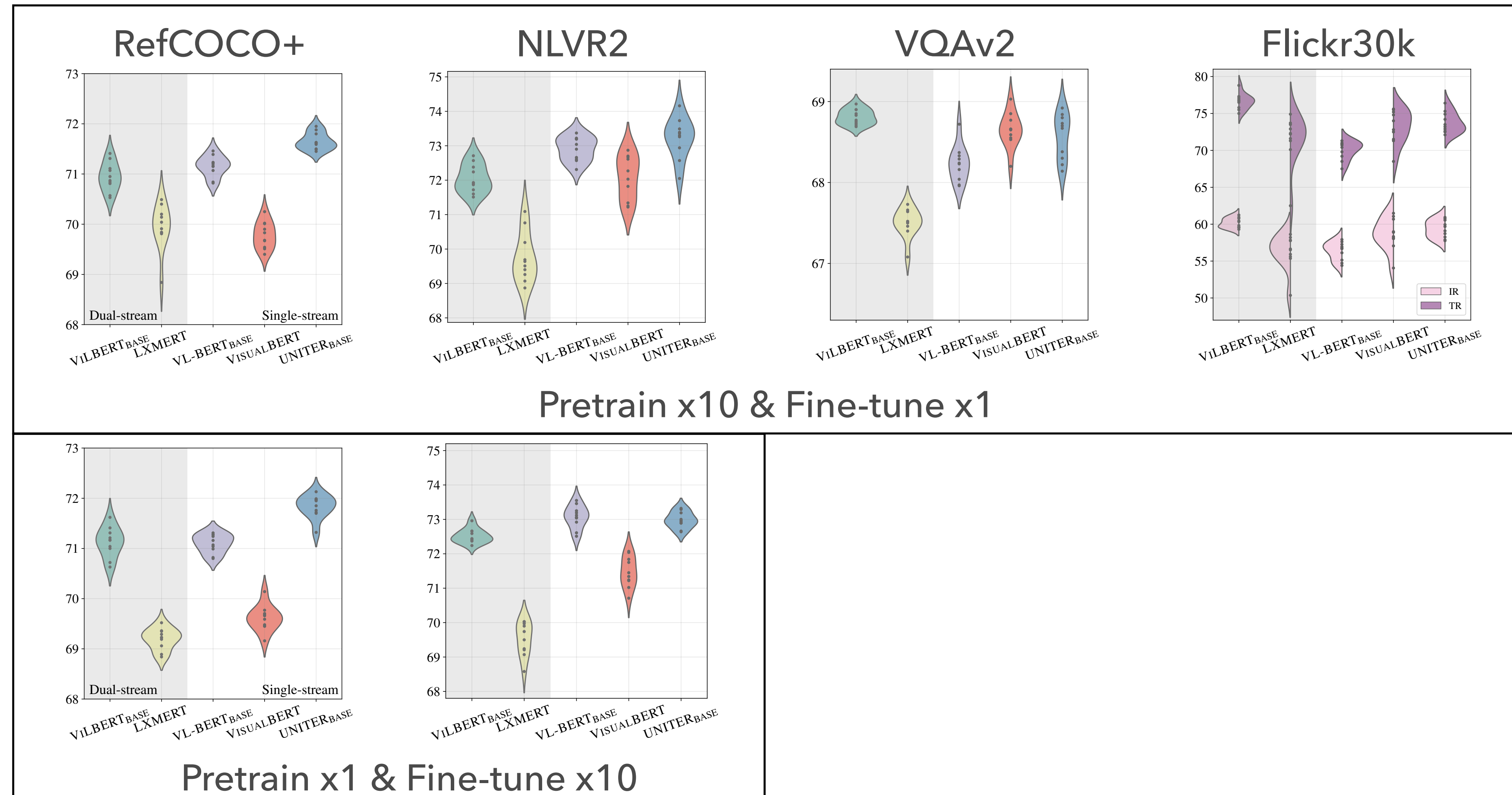


Pretrain x10 & Fine-tune x1

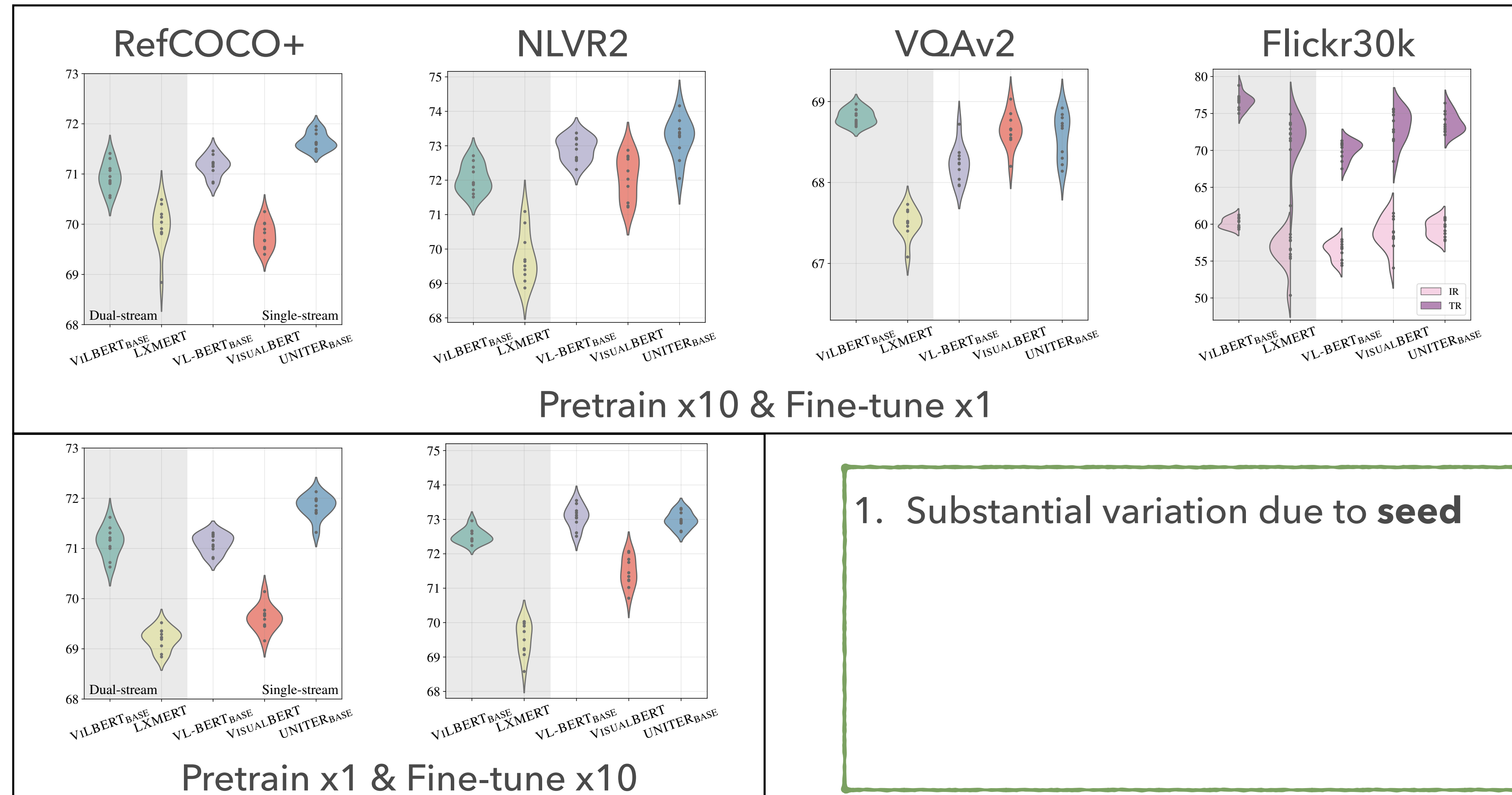
Controlled Experiments Results



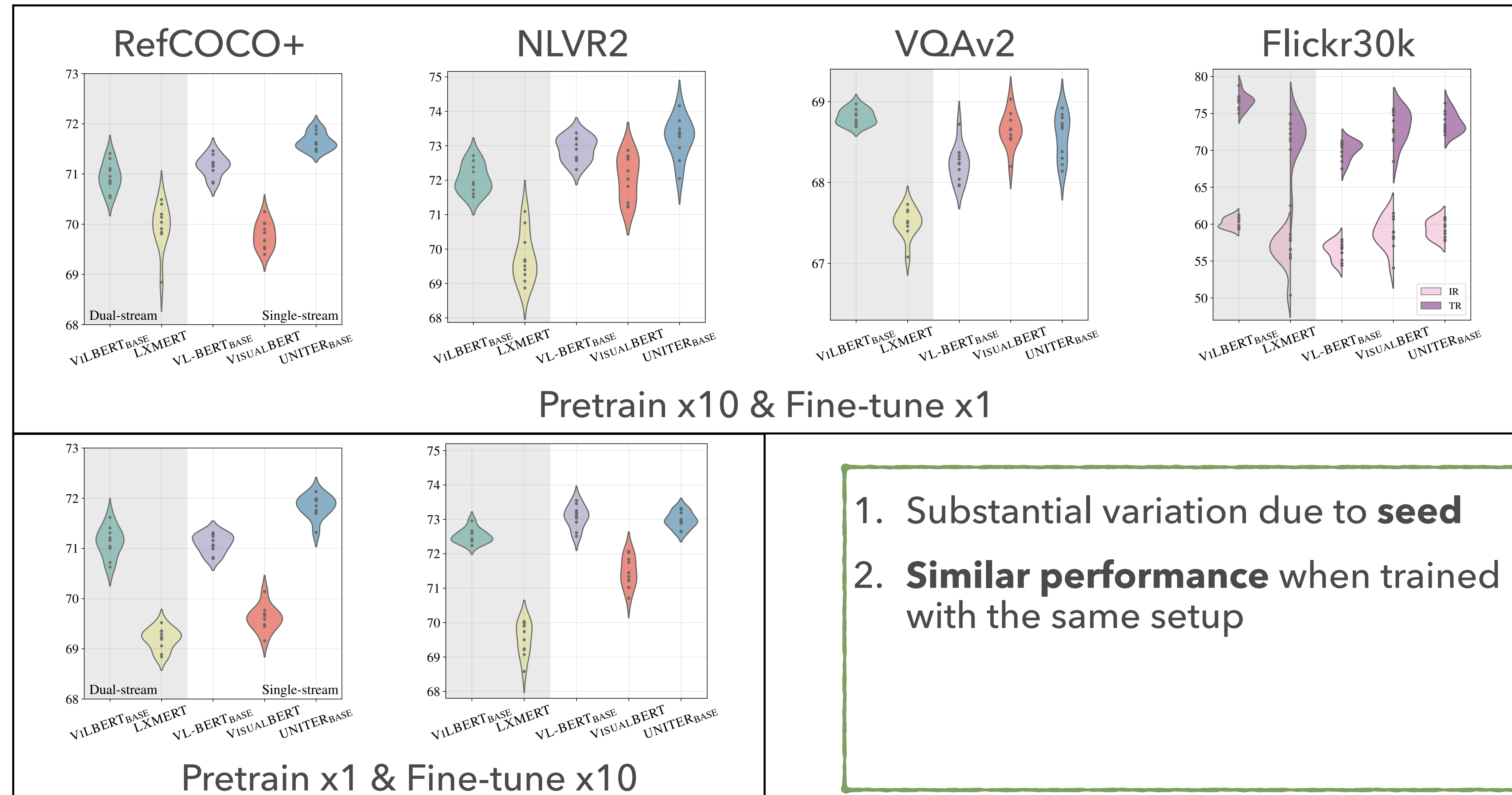
Controlled Experiments Results



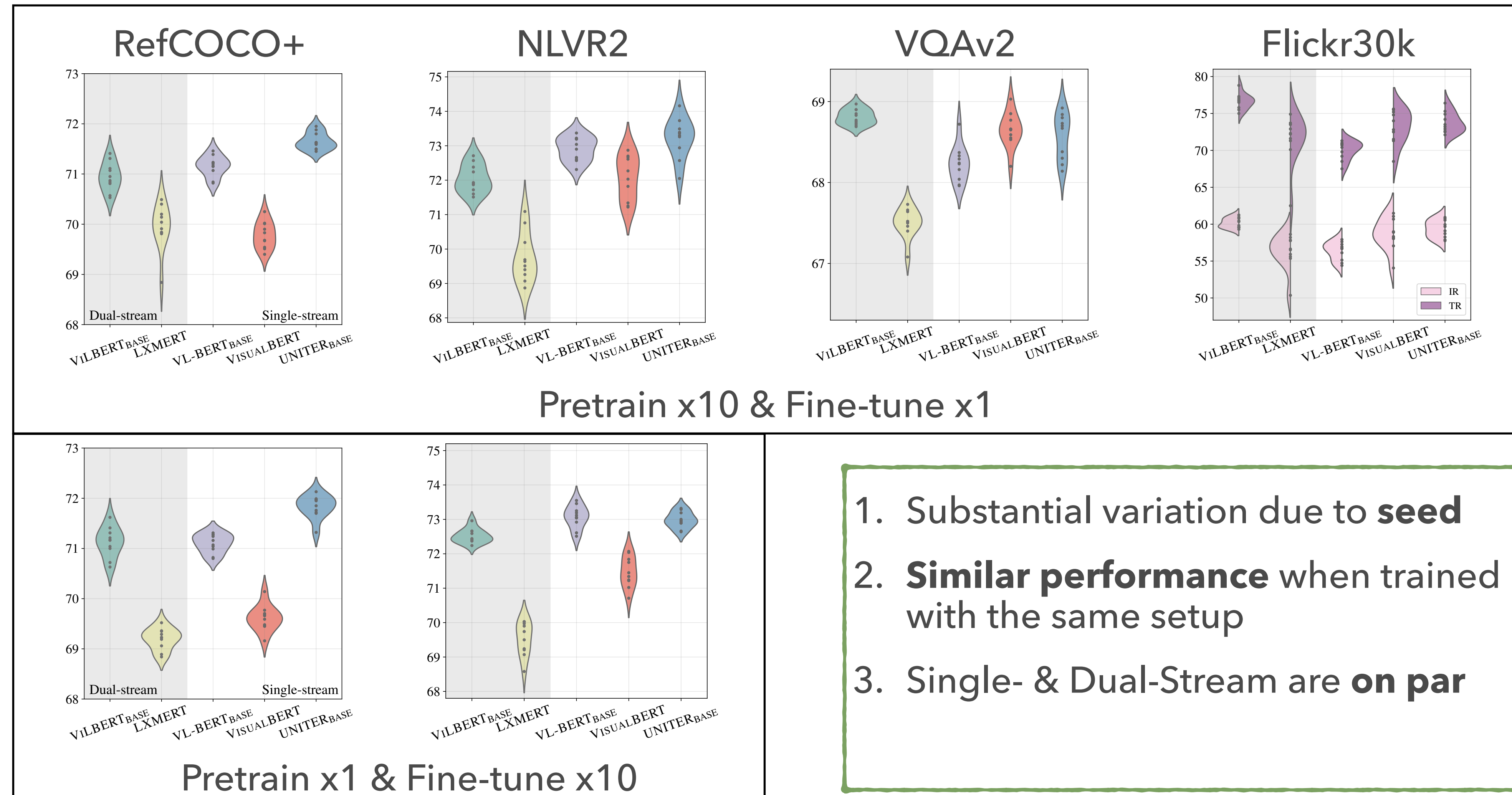
Controlled Experiments Results



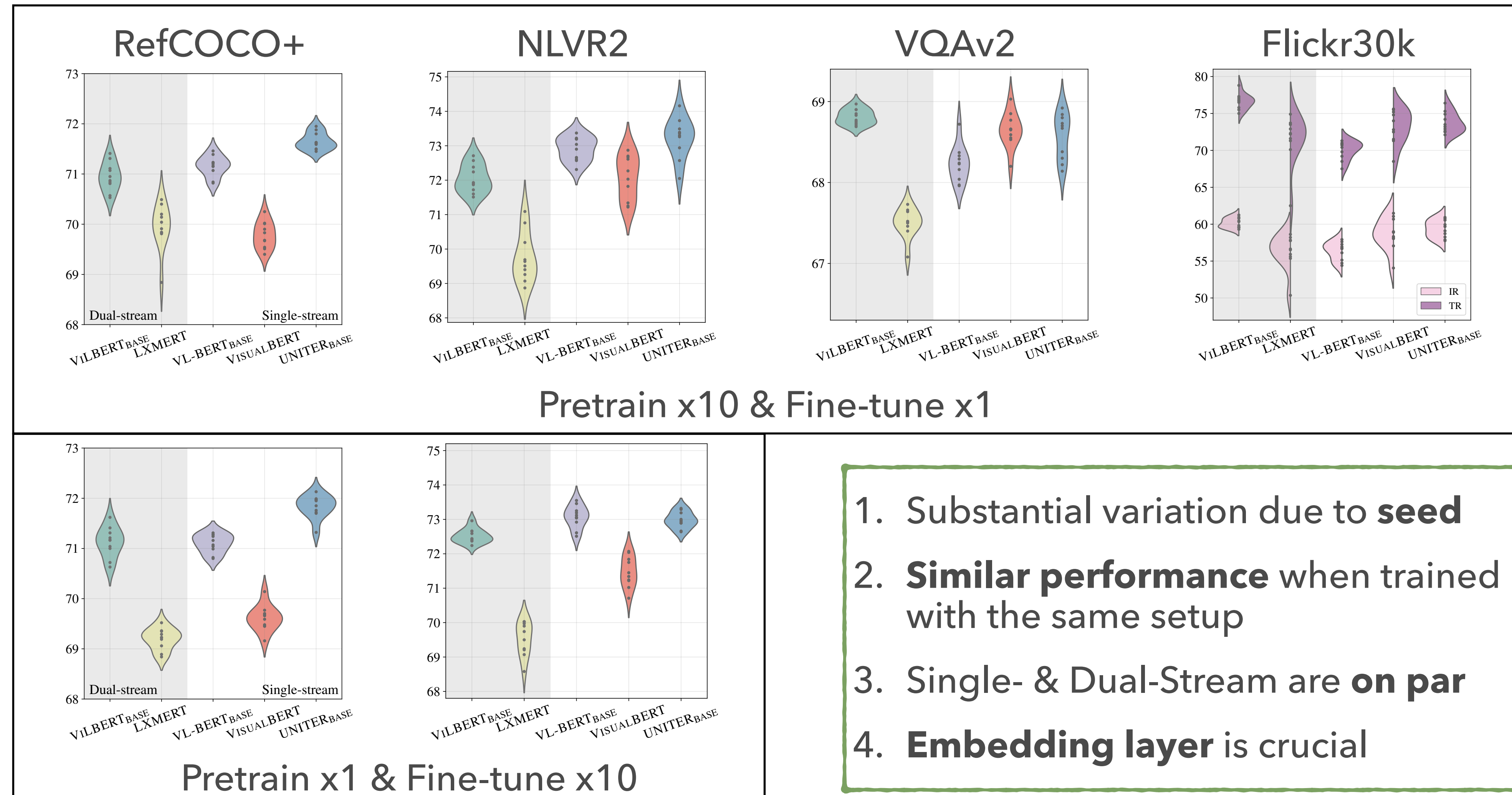
Controlled Experiments Results



Controlled Experiments Results



Controlled Experiments Results



Conclusions

Conclusions

Unified mathematical and software framework for V&L BERTs

Conclusions

Unified mathematical and software framework for V&L BERTs

Meta-Analysis of 5 V&L BERTs

Conclusions

Unified mathematical and software framework for V&L BERTs

Meta-Analysis of 5 V&L BERTs

- Most differences due to training data & hyperparameters

Conclusions

Unified mathematical and software framework for V&L BERTs

Meta-Analysis of 5 V&L BERTs

- Most differences due to training data & hyperparameters
- Embedding layer plays a crucial role

Conclusions

Unified mathematical and software framework for V&L BERTs

Meta-Analysis of 5 V&L BERTs

- Most differences due to training data & hyperparameters
- Embedding layer plays a crucial role

Future work *should*

Conclusions

Unified mathematical and software framework for V&L BERTs

Meta-Analysis of 5 V&L BERTs

- Most differences due to training data & hyperparameters
- Embedding layer plays a crucial role

Future work *should*

- Train on similar datasets
- Report variance

Conclusions

Unified mathematical and software framework for V&L BERTs

Meta-Analysis of 5 V&L BERTs

- Most differences due to training data & hyperparameters
- Embedding layer plays a crucial role

Future work *should*

- Train on similar datasets
- Report variance

Code, models (50!) and data available online

- github.com/e-bug/volta
- github.com/e-bug/mpre-unmasked

Conclusions

Thank you

Unified mathematical and software framework for V&L BERTs

Meta-Analysis of 5 V&L BERTs

- Most differences due to training data & hyperparameters
- Embedding layer plays a crucial role

Future work *should*

- Train on similar datasets
- Report variance

Code, models (50!) and data available online

- github.com/e-bug/volta
- github.com/e-bug/mpre-unmasked