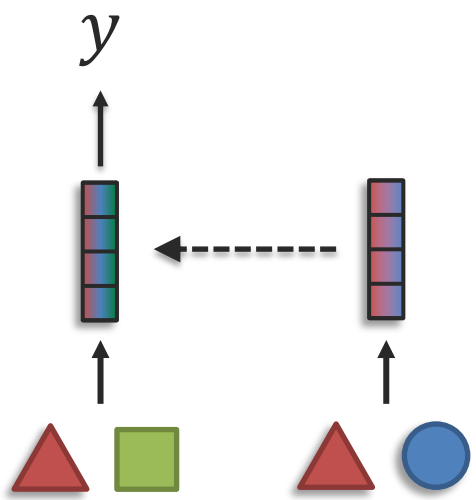
Challenge 5: Transference

Transference

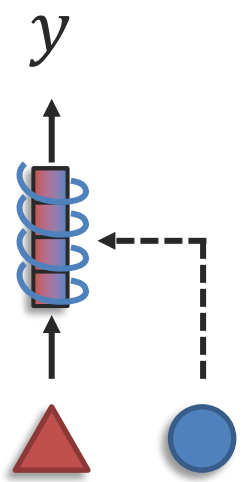
Definition: Transfer knowledge between modalities, usually to help the primary modality which may be noisy or with limited resources

Sub-challenges:

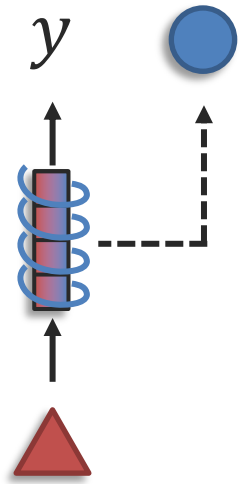
Transfer



Co-learning via representation

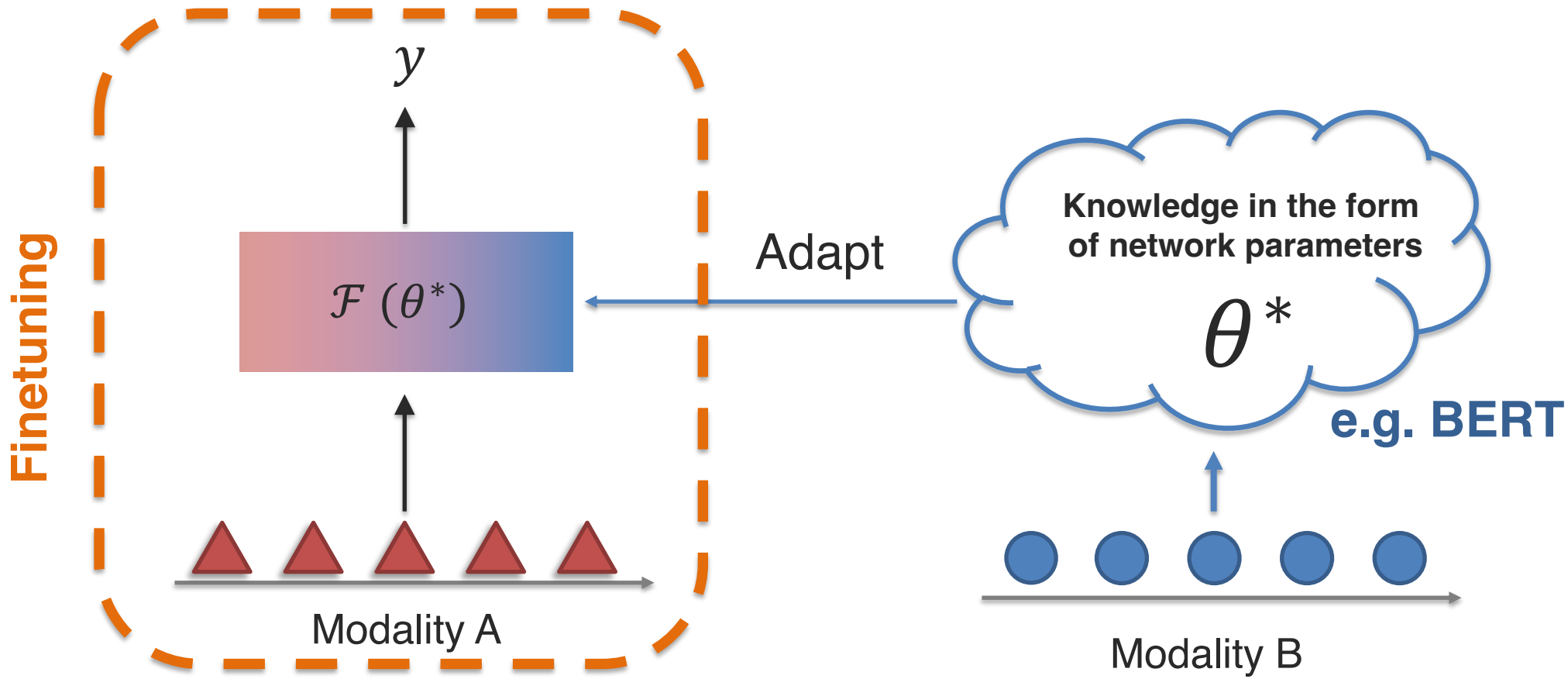


Co-learning via generation



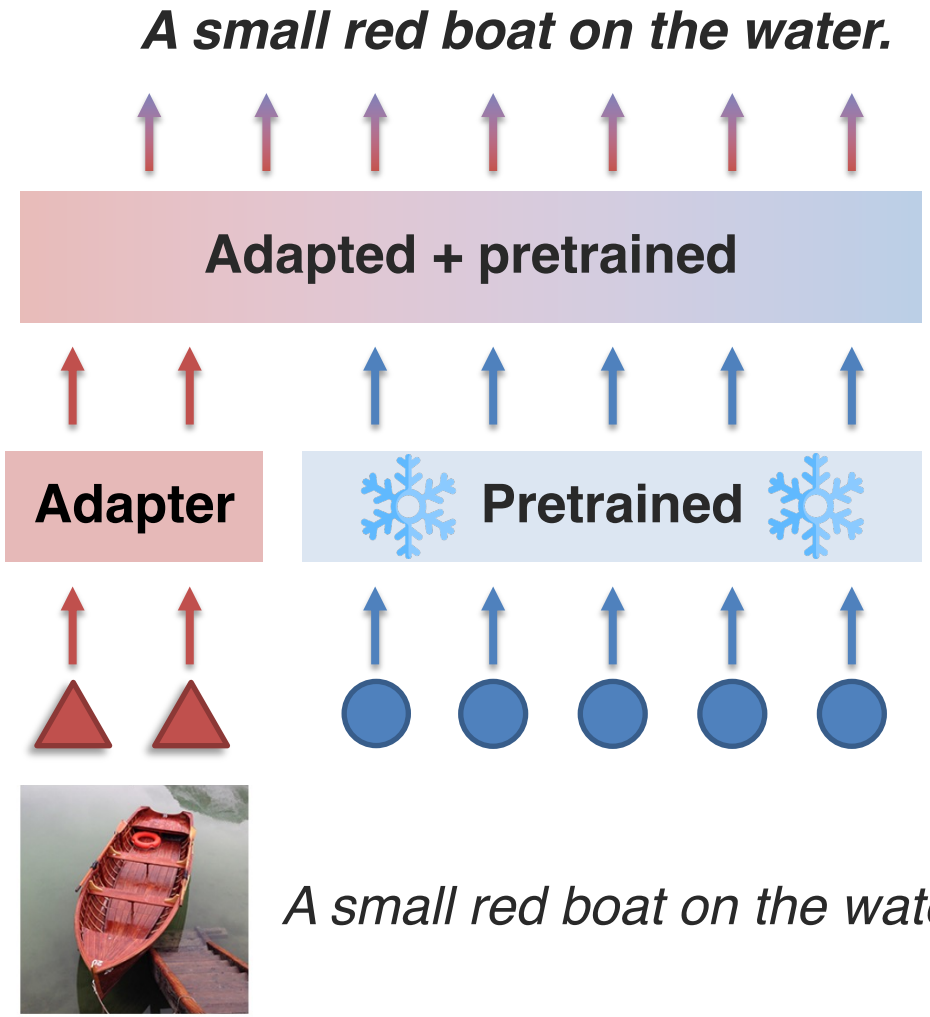
Sub-Challenge 5a: Transfer via Pretrained Models

Definition: Transferring knowledge from large-scale pretrained models to downstream tasks involving the primary modality.



Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via prefix tuning

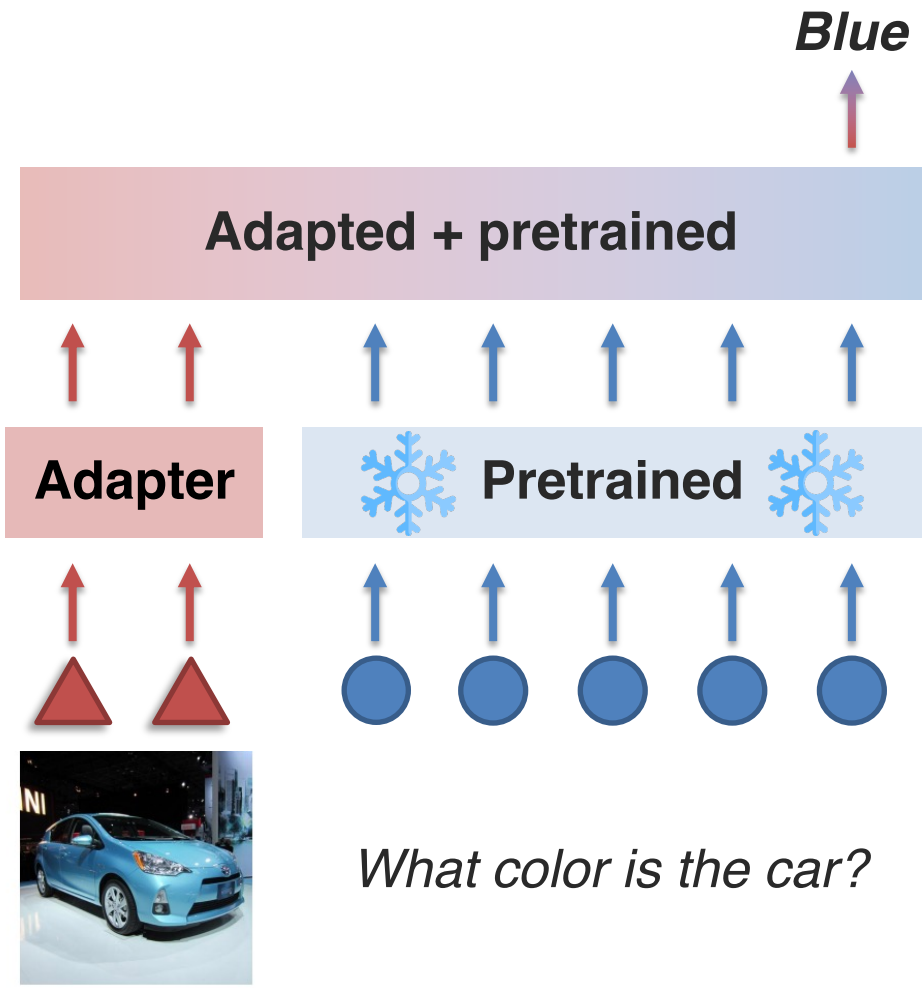


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via prefix tuning

0-shot VQA:



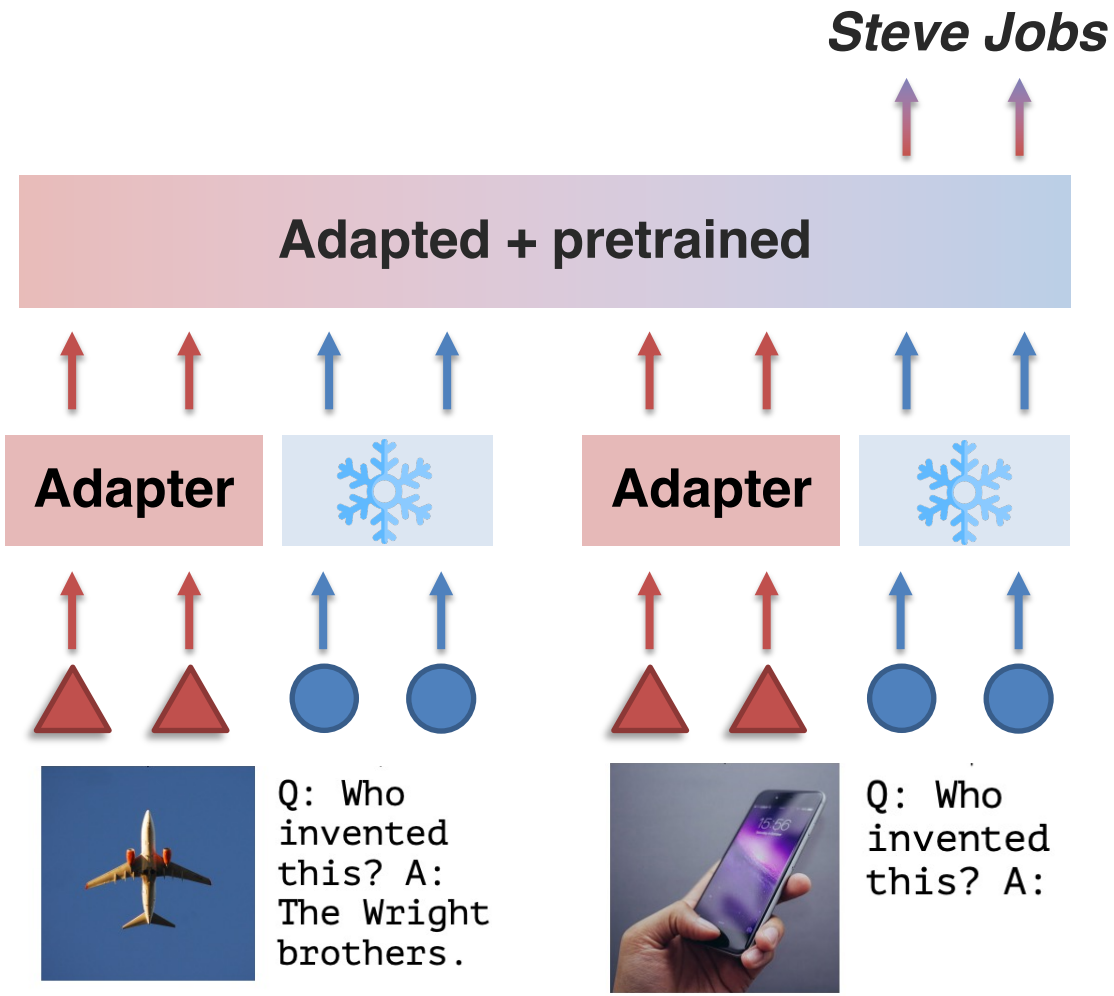
[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via prefix tuning

1-shot outside knowledge VQA:

Recall reasoning – leverage implicit knowledge in LMs

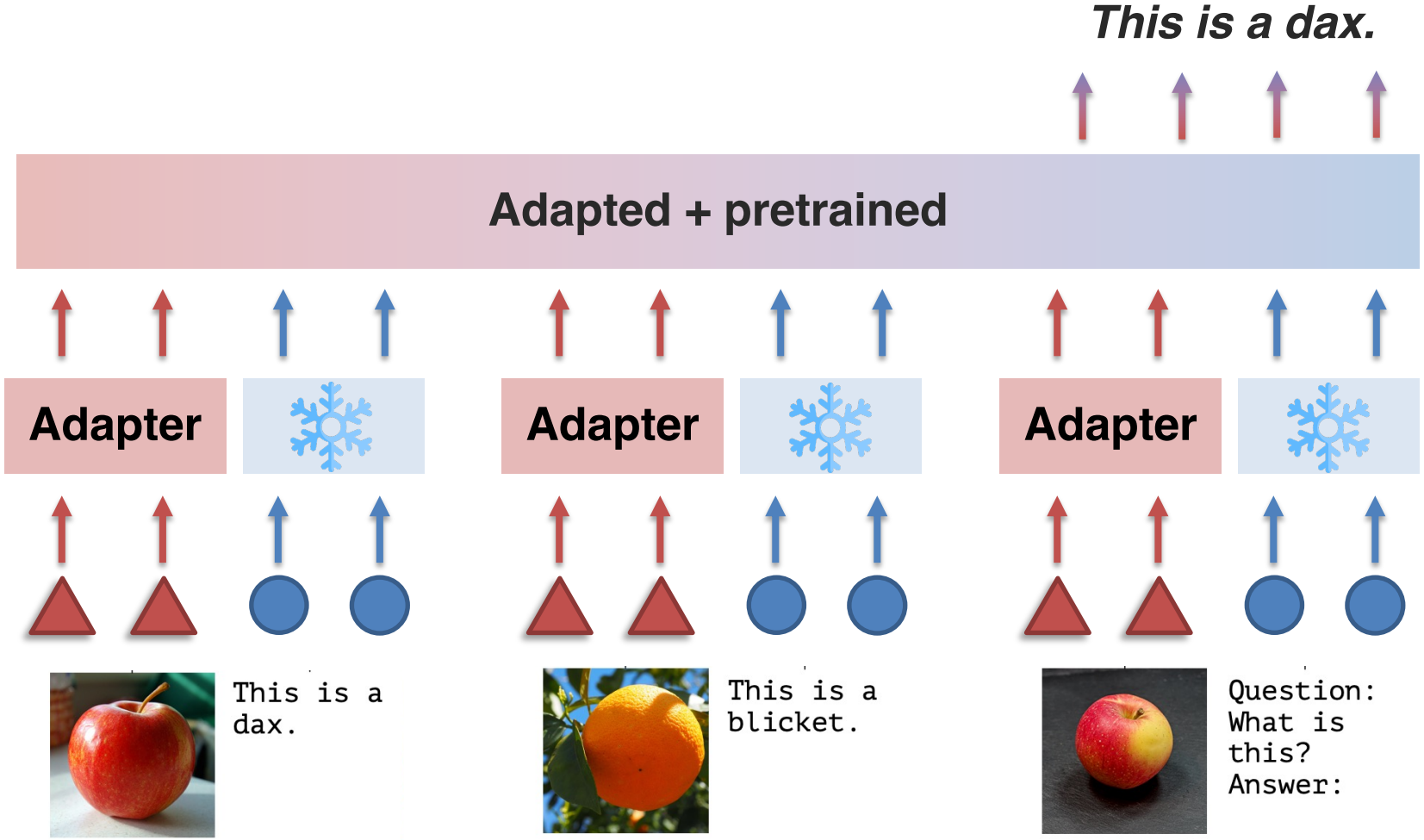


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via prefix tuning

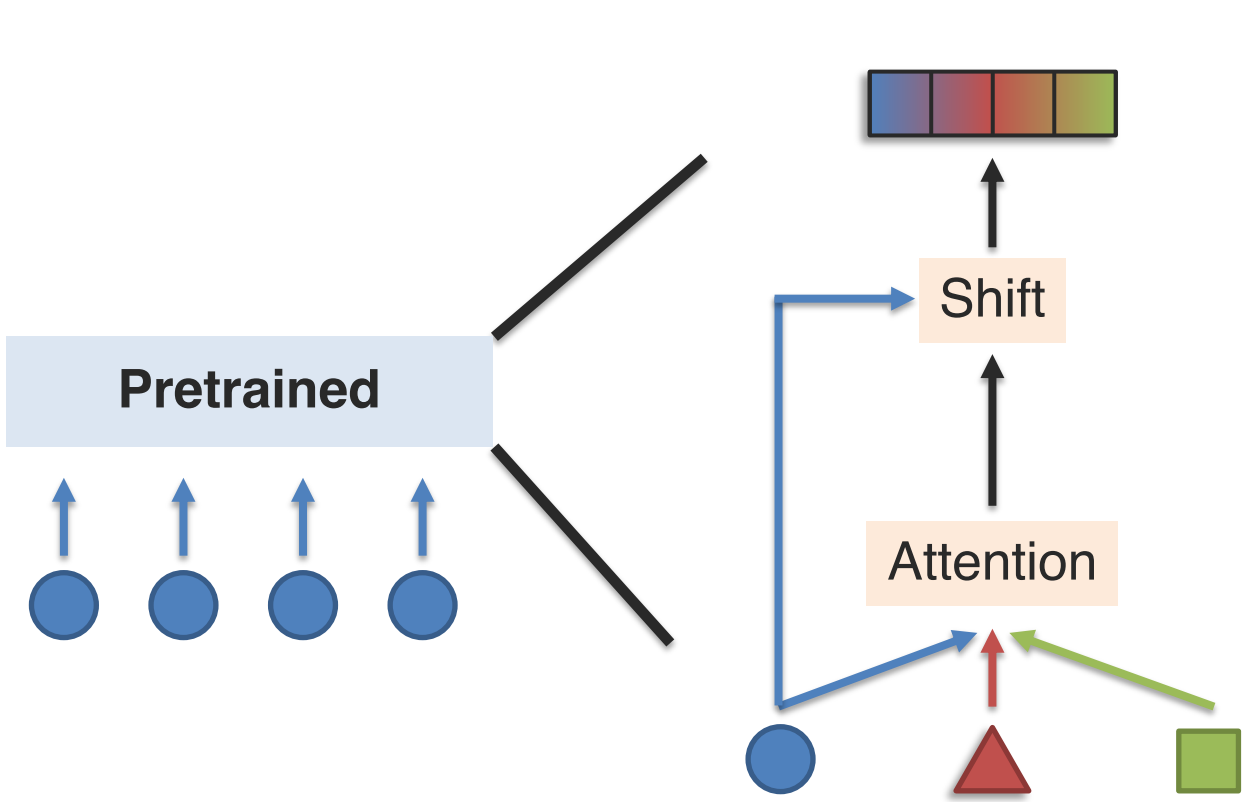
Few-shot image classification:

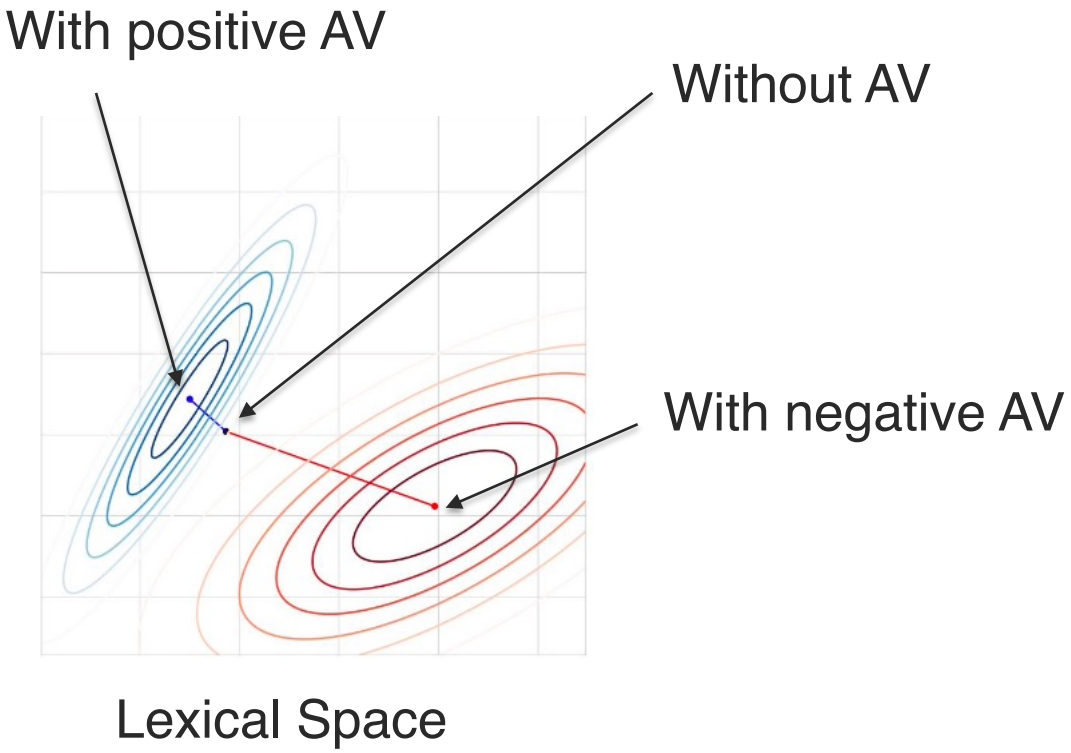


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via representation tuning





[Ziegler et al., Encoder-Agnostic Adaptation for Conditional Language Generation. arXiv 2019]

[Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers. ACL 2020]

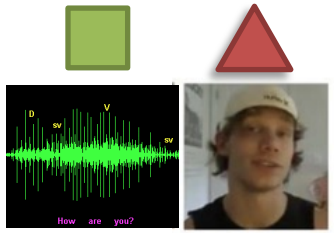
Sub-Challenge 5a: Transfer via Pretrained Models

How can we transfer knowledge across multiple tasks, each over a different subset of modalities?

Video classification

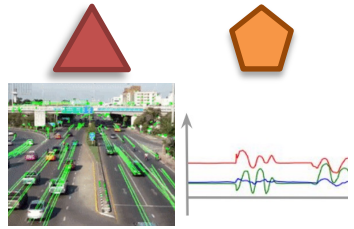
Language Video Audio

Sentiment, emotions



Audio Video

Robot dynamics



Video Time-series

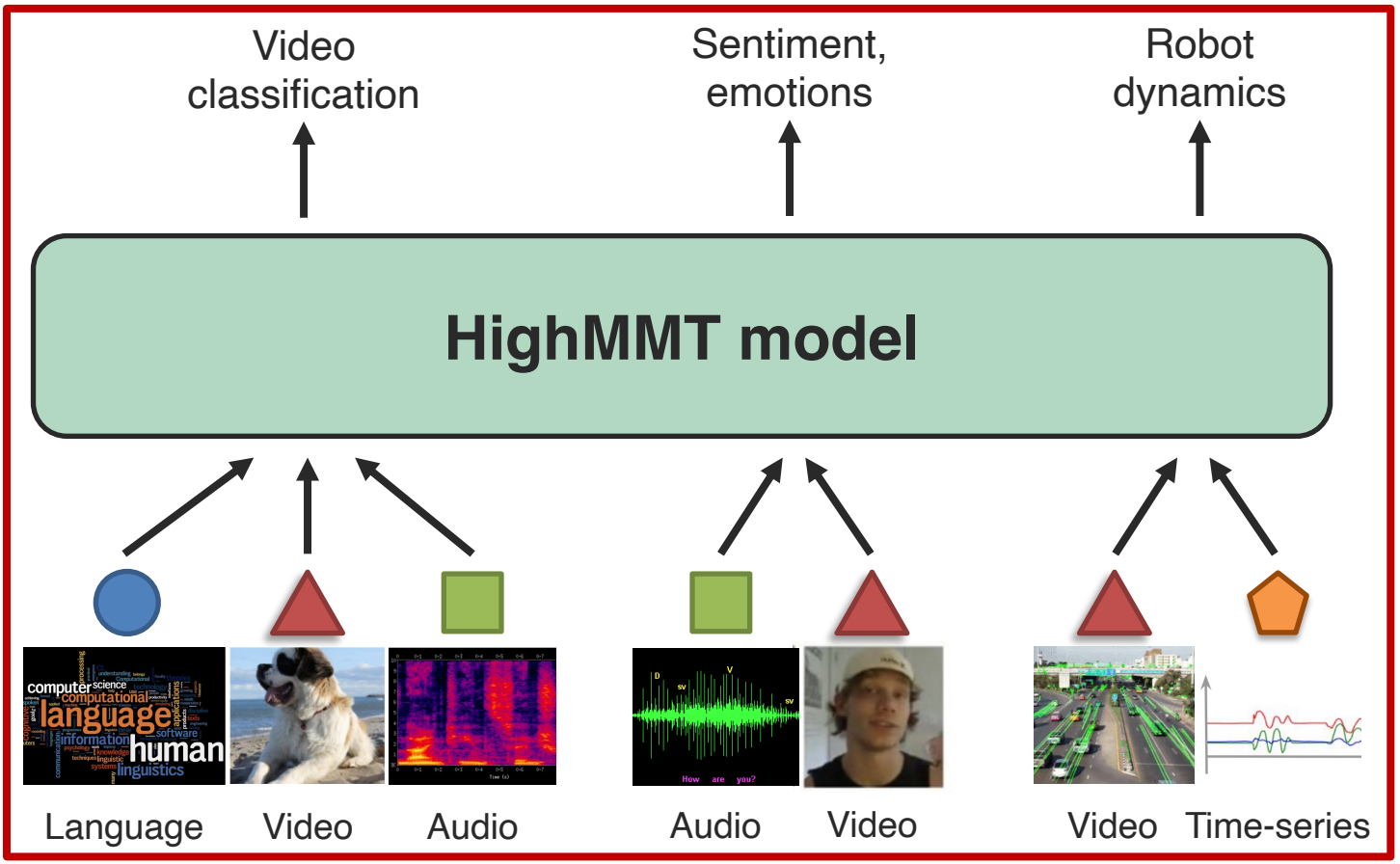
Generalization across modalities and tasks
Important if some tasks are low-resource

[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning



Non-parallel multitask learning

Task-specific classifiers

Shared multimodal model

Modality-specific embeddings

Standardized input sequence

Same model architecture!

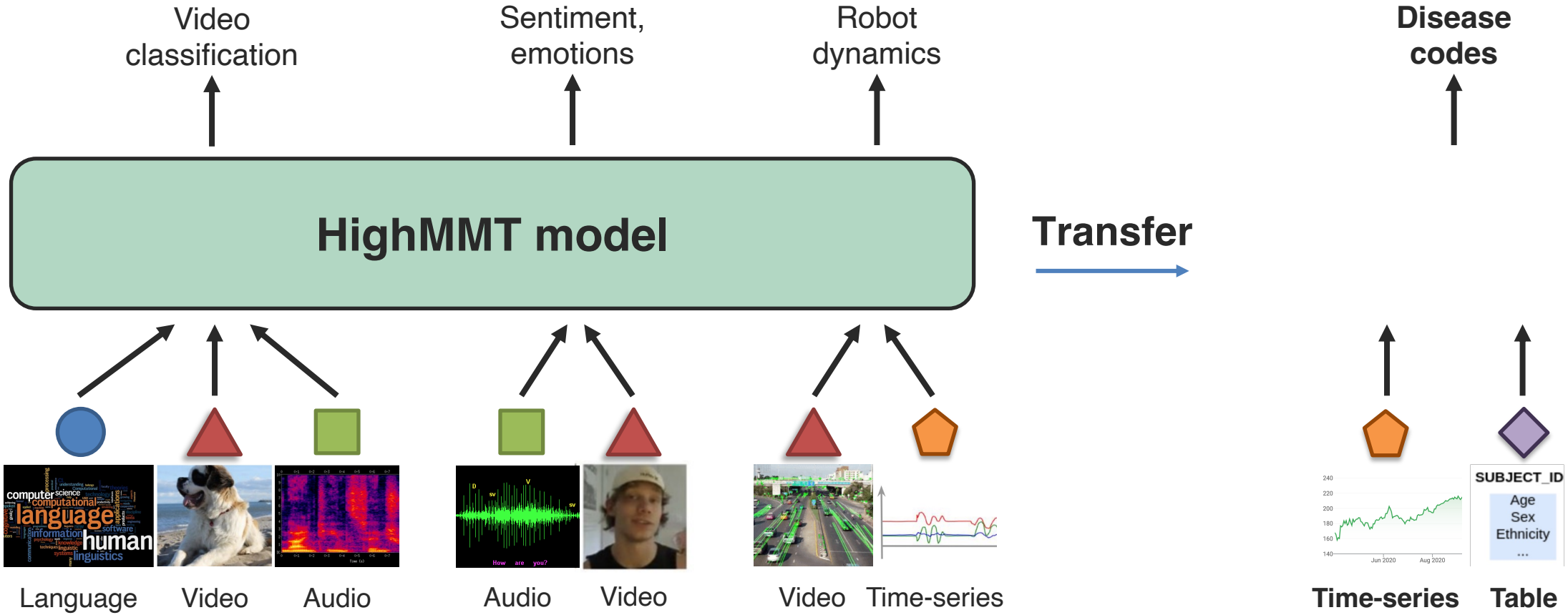
Same parameters!

[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning

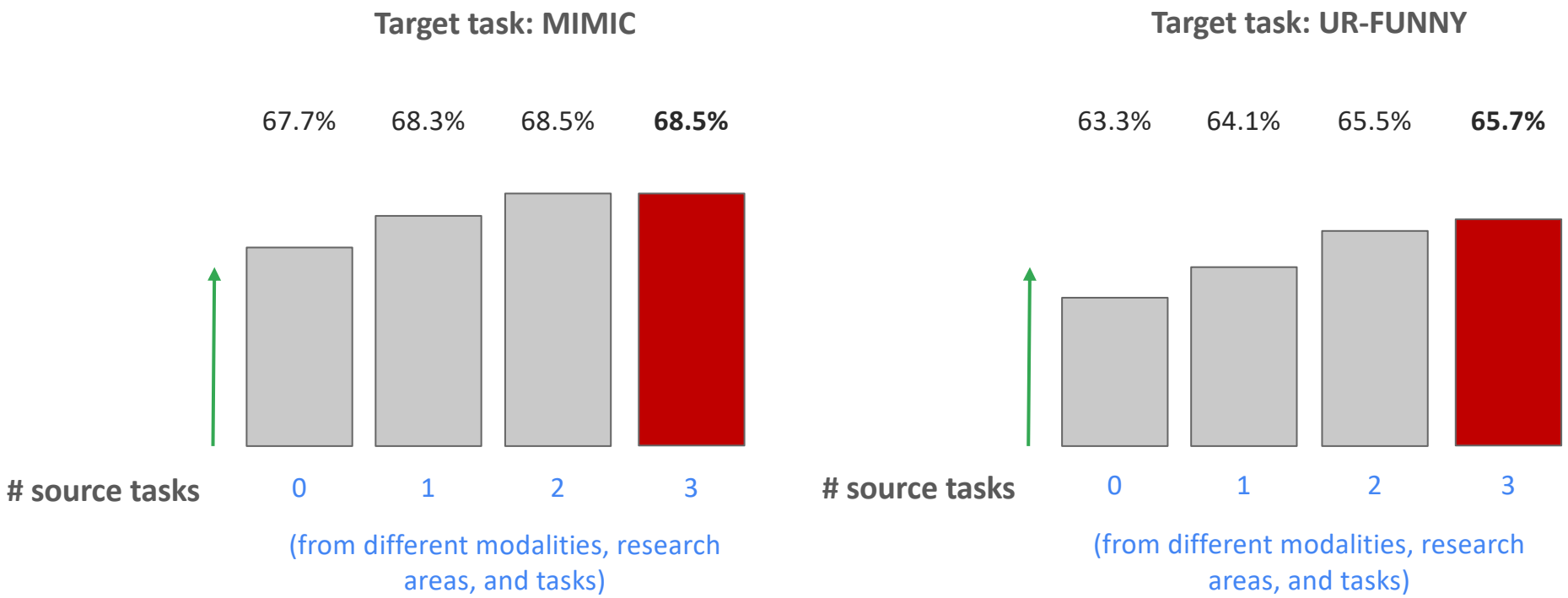


[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning



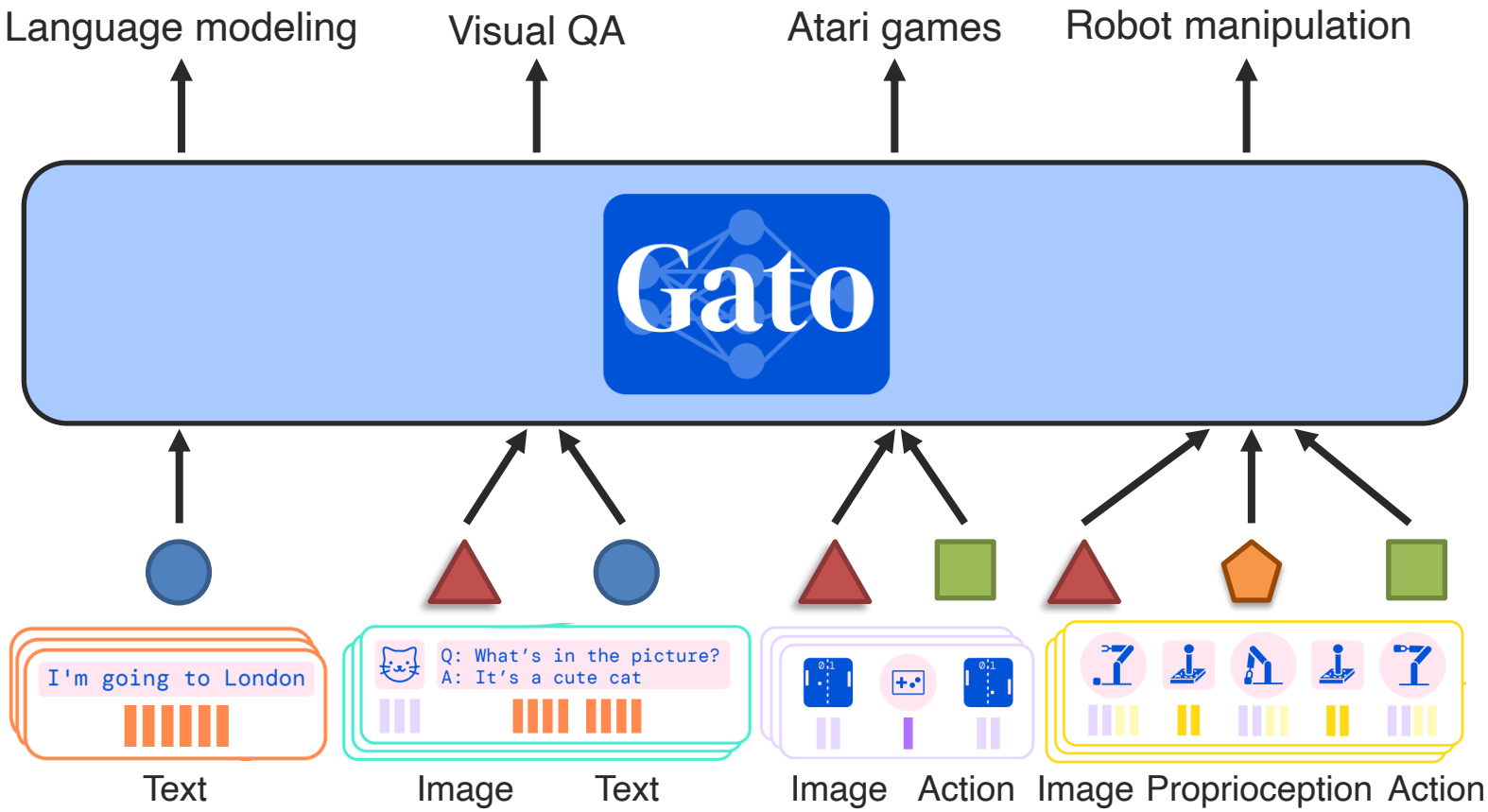
Achieves both multitask and transfer capabilities across modalities and tasks

[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer across partially observable modalities

Gato: unified model + parameter sharing + multitask learning



Non-parallel multitask learning

Task-specific classifiers

Same model architecture!

Shared multimodal model

Same parameters!

Modality-specific embeddings

Standardized input sequence

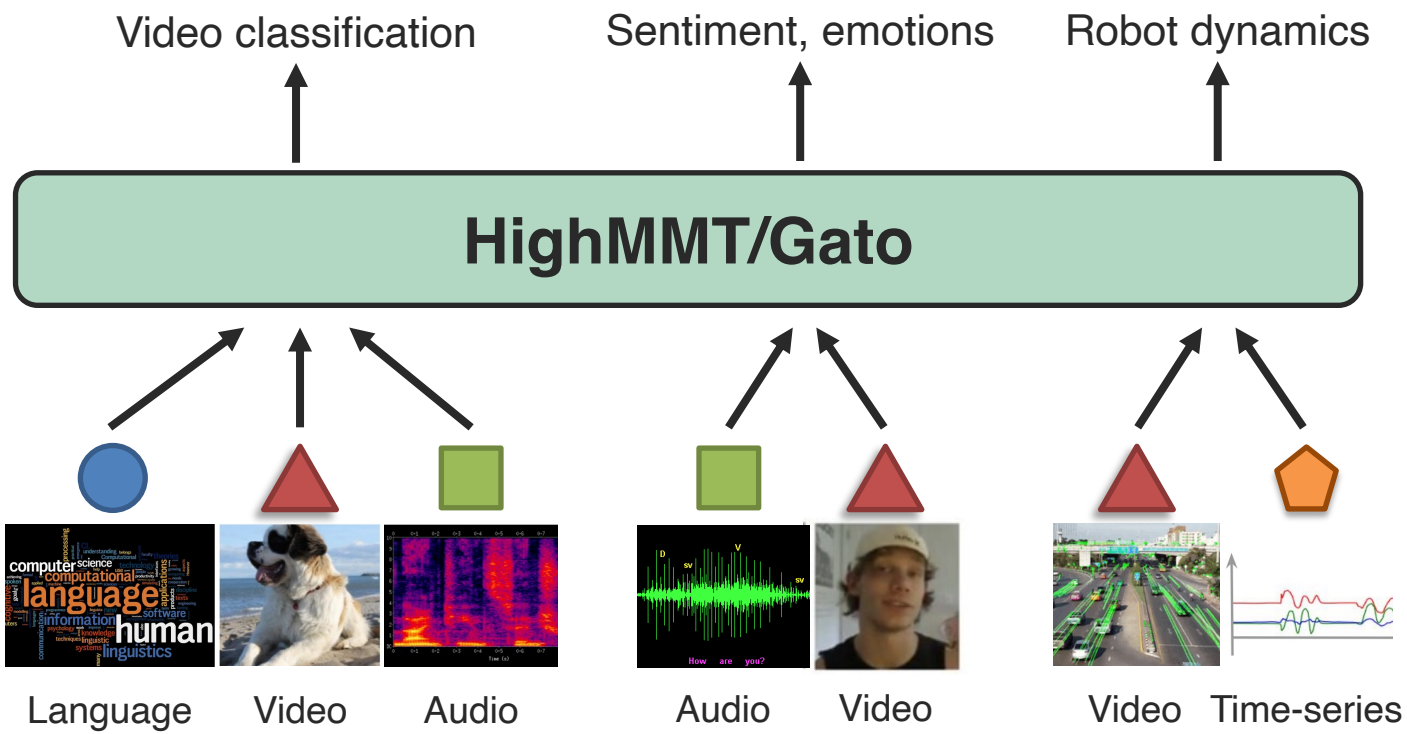
[Reed et al., A Generalist Agent. arXiv 2022]

Sub-Challenge 5a: Transfer via Pretrained Models

Open challenges

Some implicit assumptions:

- All modalities can be represented as sequences without losing information



Standardized input sequence?

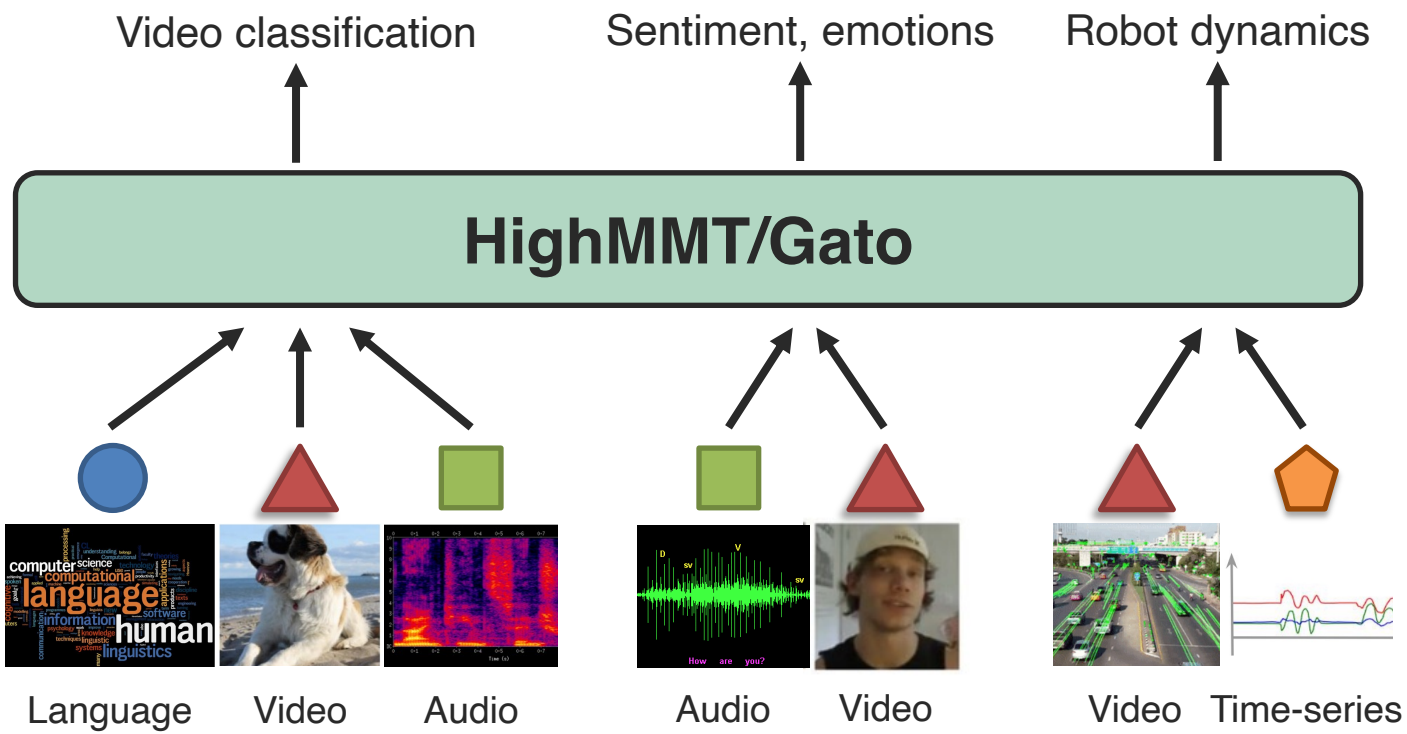
[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

Sub-Challenge 5a: Transfer via Pretrained Models

Open challenges

Some implicit assumptions:

- All modalities can be represented as sequences without losing information
- Dimensions of heterogeneity can be perfectly captured by modality-specific embeddings



Modality-specific embeddings?

Standardized input sequence?

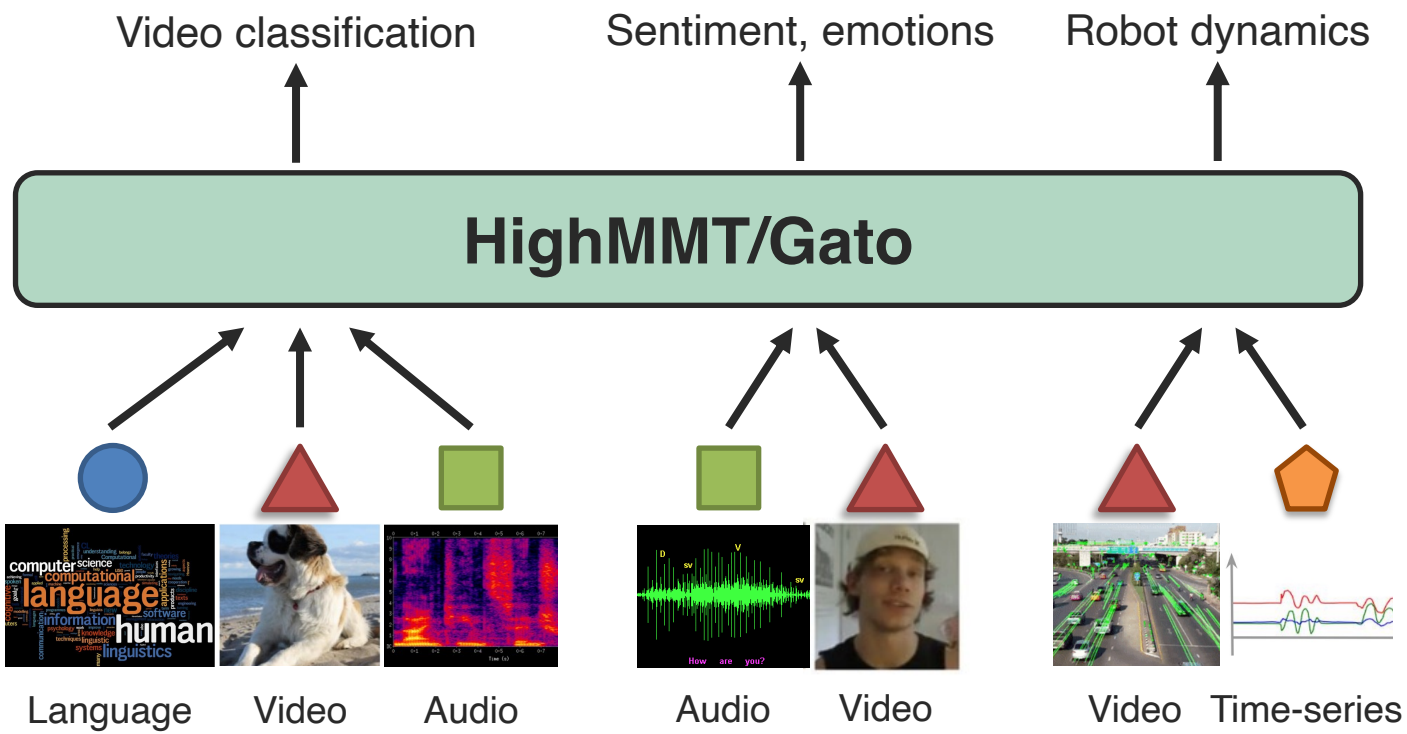
[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

Sub-Challenge 5a: Transfer via Pretrained Models

Open challenges

Some implicit assumptions:

- All modalities can be represented as sequences without losing information
- Dimensions of heterogeneity can be perfectly captured by modality-specific embeddings
- Cross-modal connections & interactions are shared across modalities and tasks



Shared multimodal model?

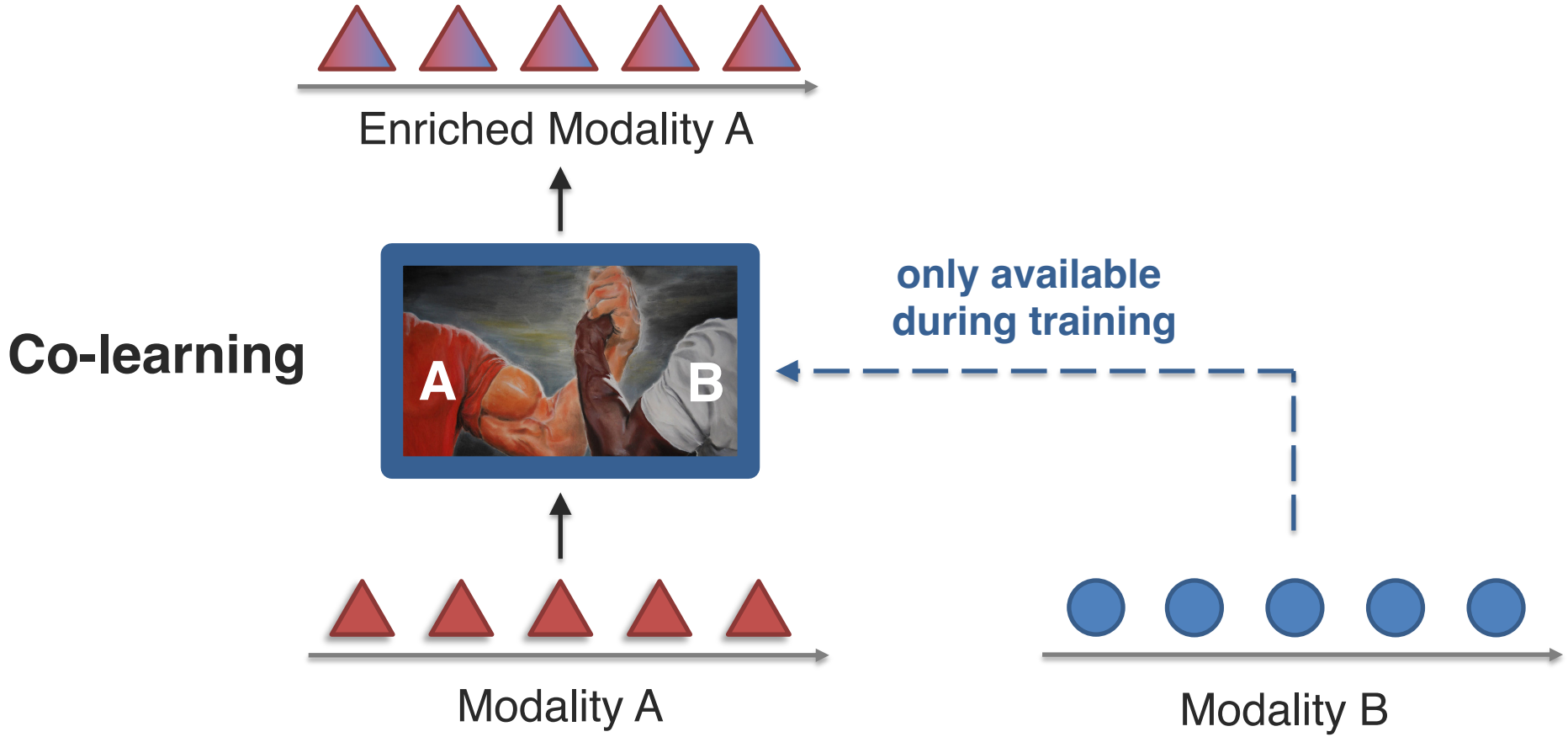
Modality-specific embeddings?

Standardized input sequence?

[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. arXiv 2022]

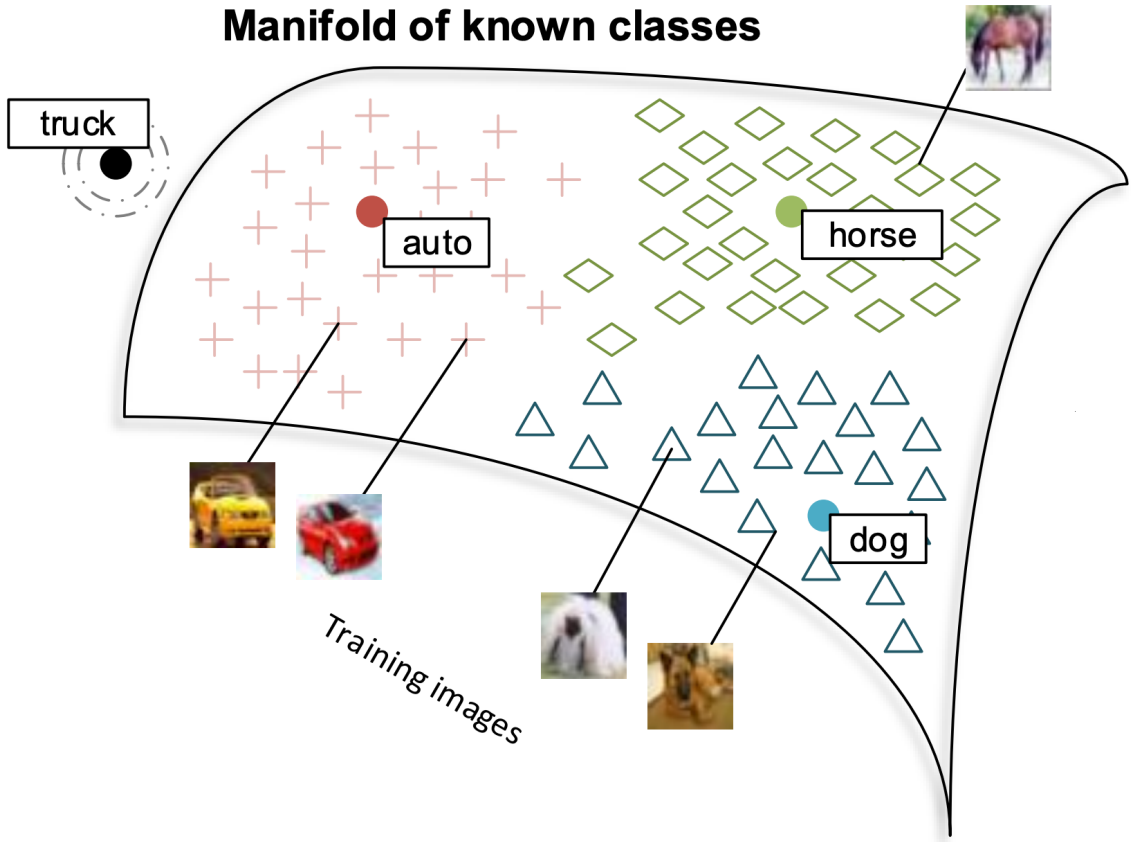
Sub-Challenge 5b: Co-learning via Representation

Definition: Transferring information from secondary to primary modality by sharing representation spaces between both modalities.

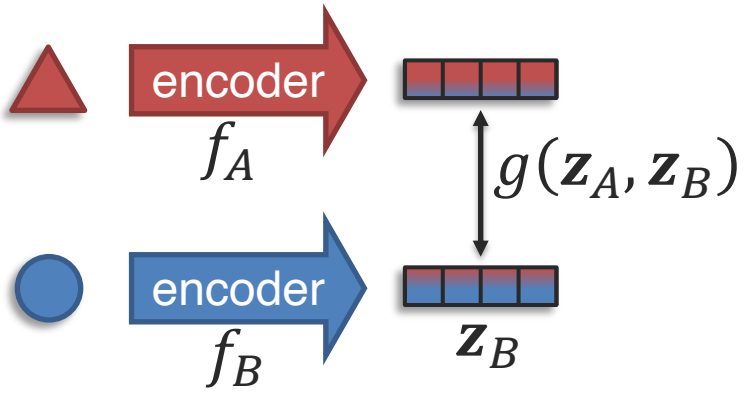


Sub-Challenge 5b: Co-learning via Representation

Representation coordination: word embedding space for zero-shot visual classification



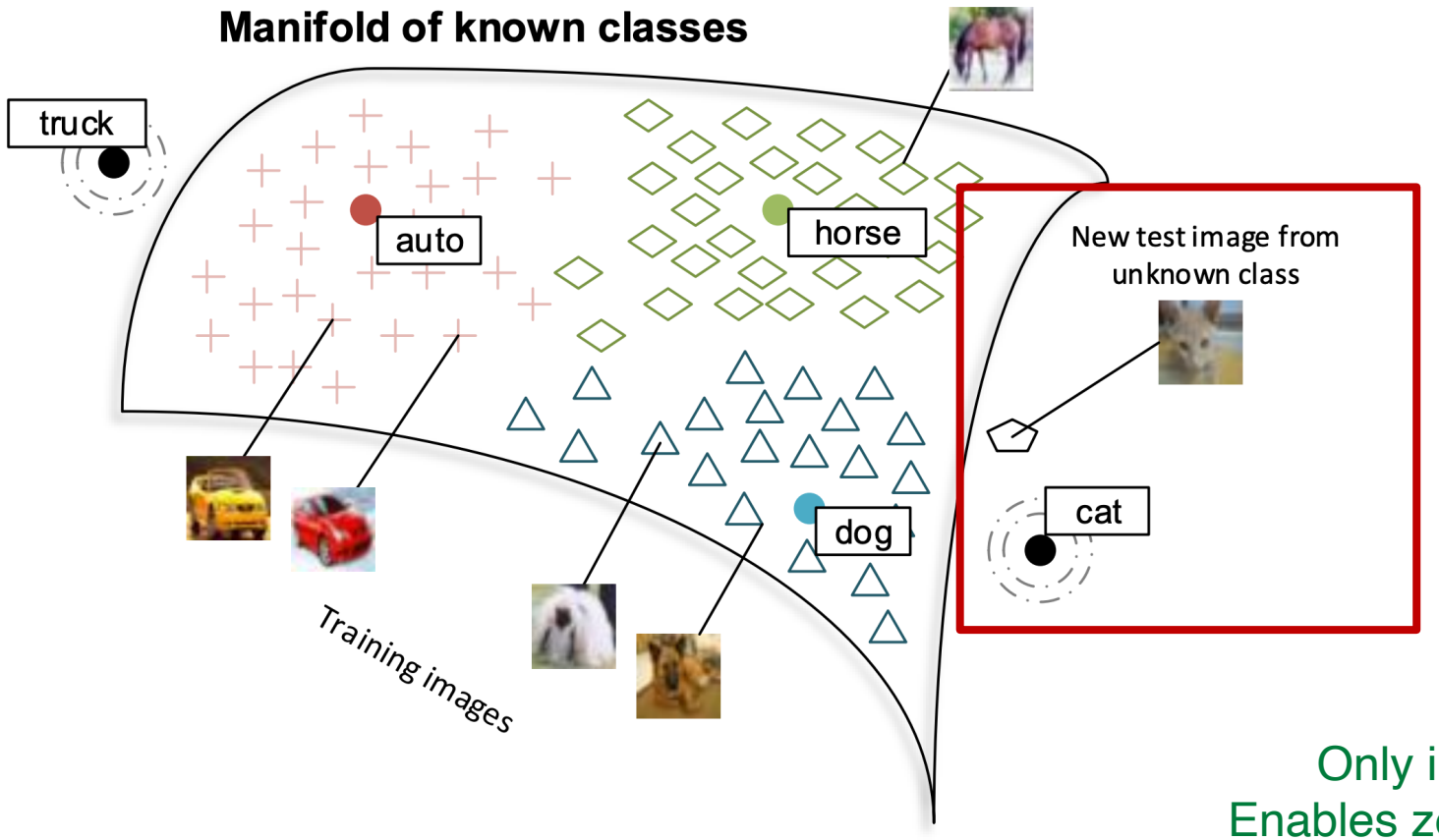
Recall representation coordination!



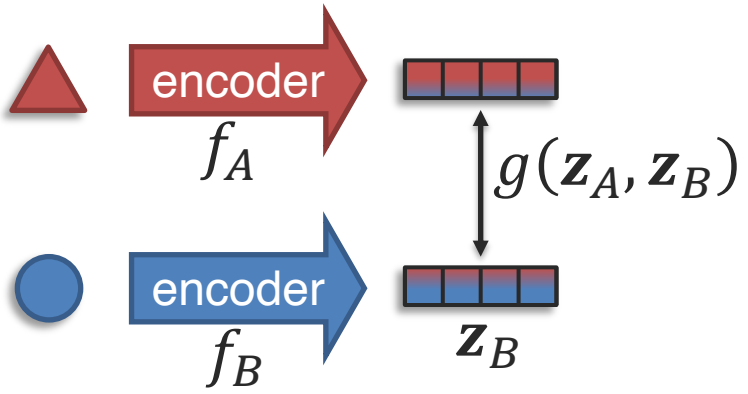
[Socher et al., Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS 2013]

Sub-Challenge 5b: Co-learning via Representation

Representation coordination: word embedding space for zero-shot visual classification



Recall representation coordination!



Only images used at test-time
Enables zero-shot image classification

[Socher et al., Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS 2013]

Sub-Challenge 5b: Co-learning via Representation

Representation fusion

Multimodal co-learning

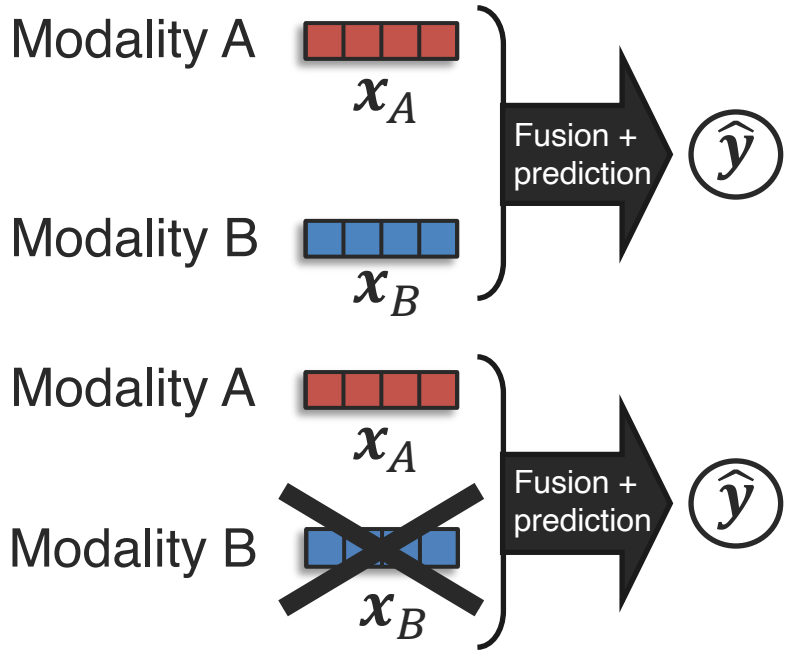
Unimodal learning

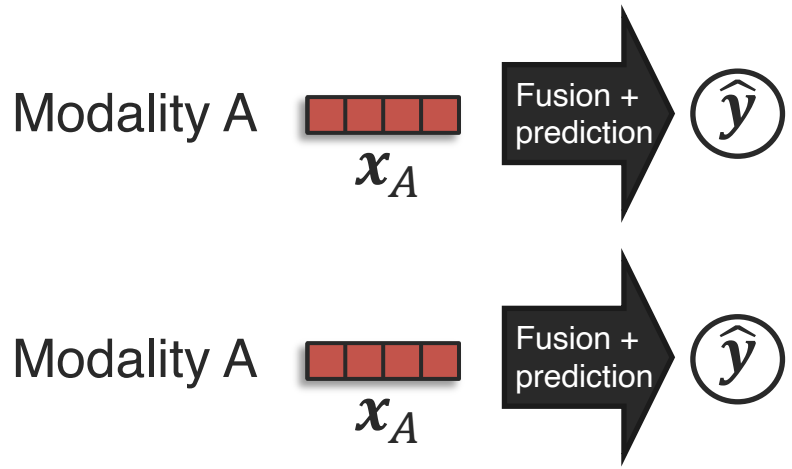
Train

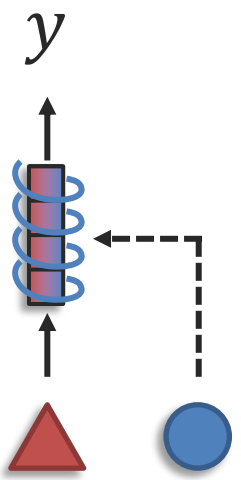
Multimodal data
Multimodal model

Test

Language-only data
Language-only model
Fill rest by 0s







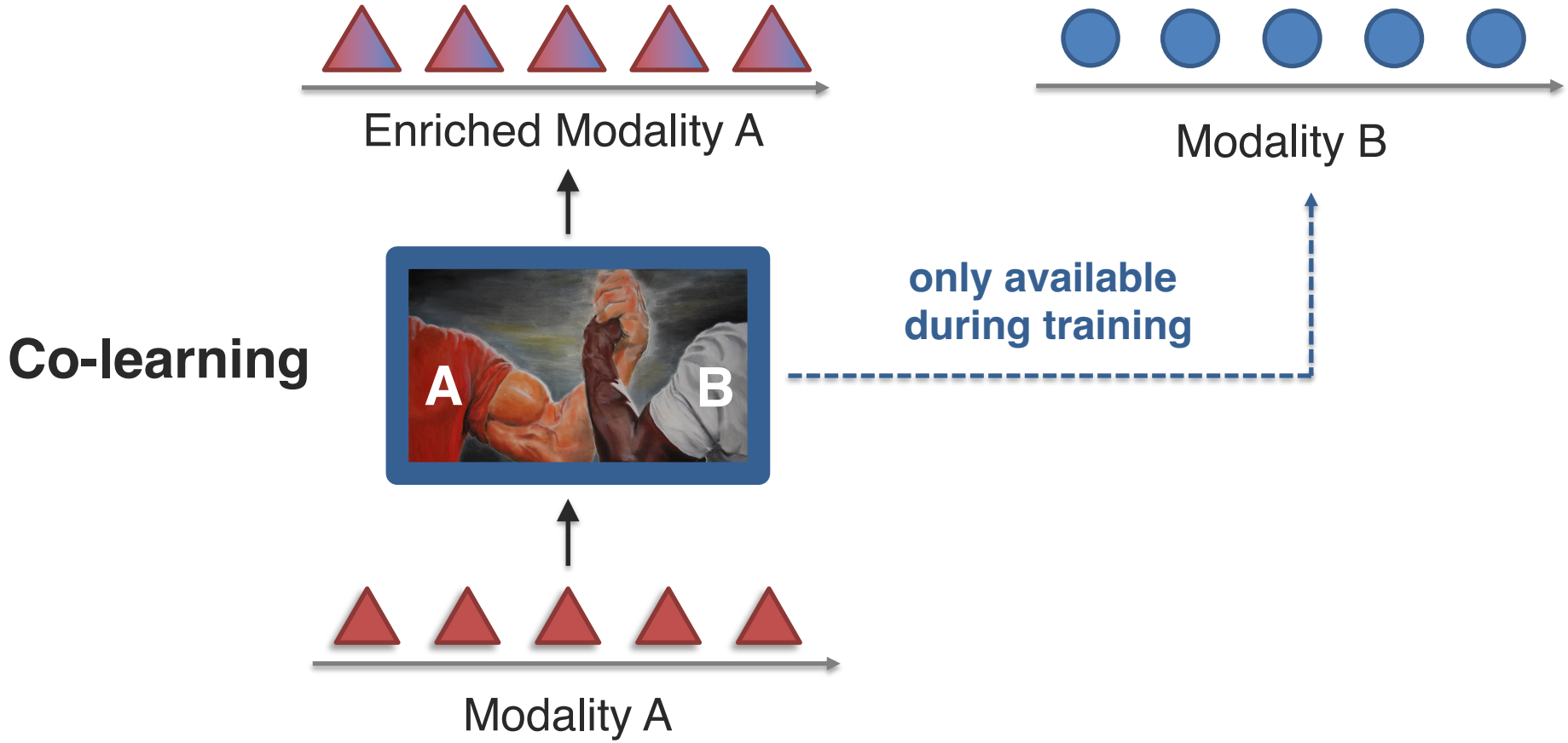
Only text used at test-time

Multimodal co-learning > language-only training

[Zadeh et al., Foundations of Multimodal Co-learning. Information Fusion 2020]

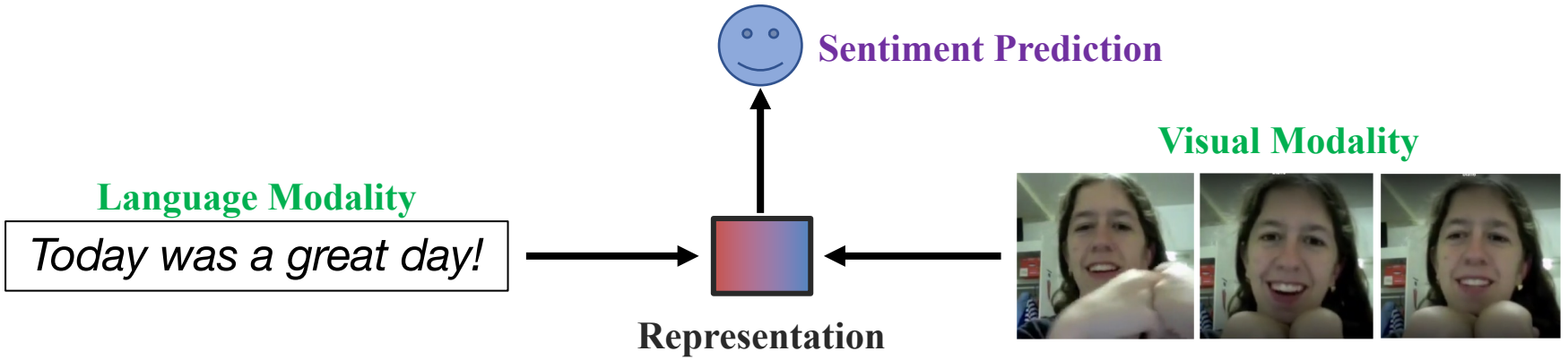
Sub-Challenge 5c: Co-learning via Generation

Definition: Transferring information from secondary to primary modality by using the secondary modality as a generation target.



Sub-Challenge 5c: Co-learning via Generation

Bimodal translations



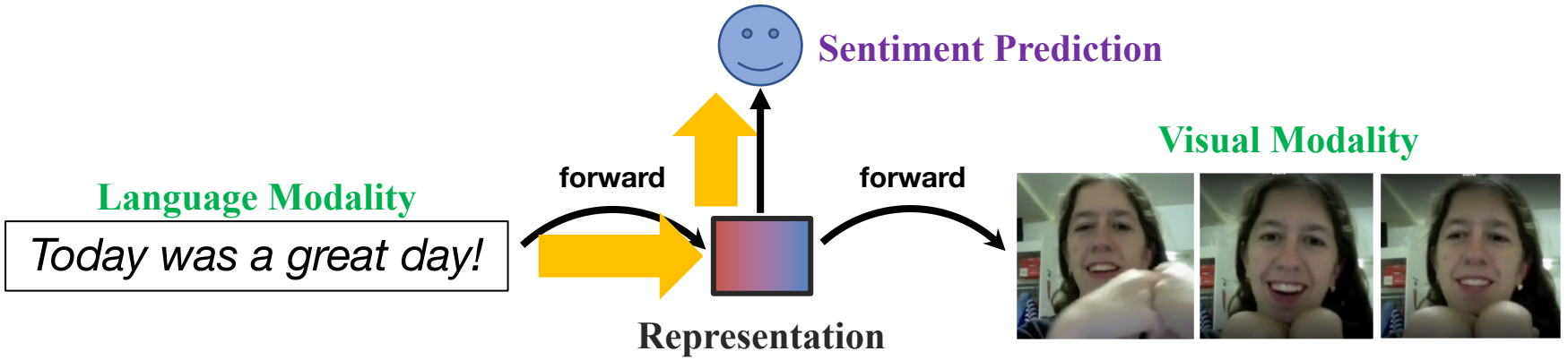
Both modalities required at test time!
Sensitive to noisy/missing visual modality.

We want to leverage information from visual modality
while being robust to it during test-time.

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Sub-Challenge 5c: Co-learning via Generation

Bimodal translations

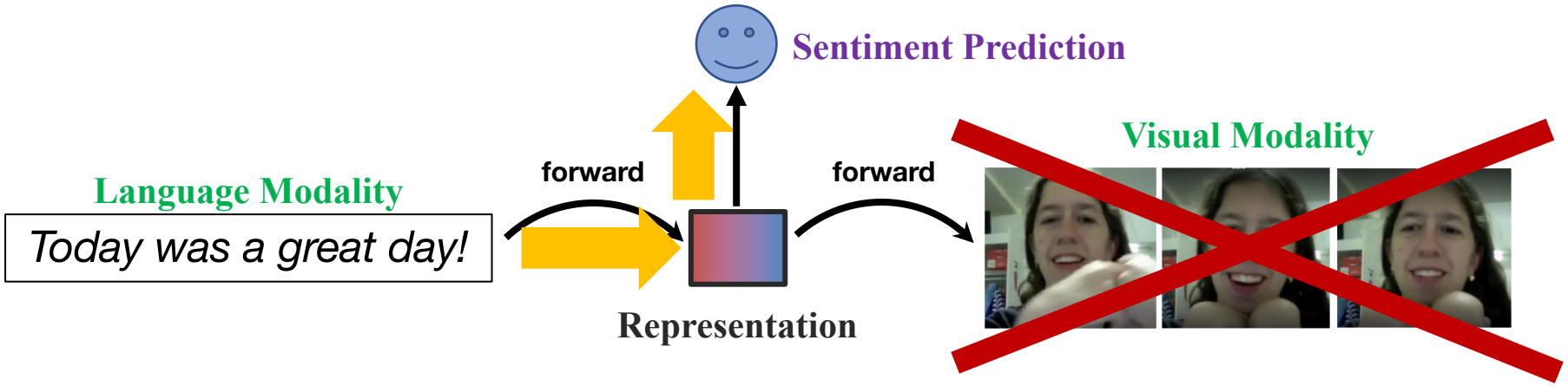


Cross-modal translation during training
Only language modality required at test time!

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Sub-Challenge 5c: Co-learning via Generation

Bimodal translations

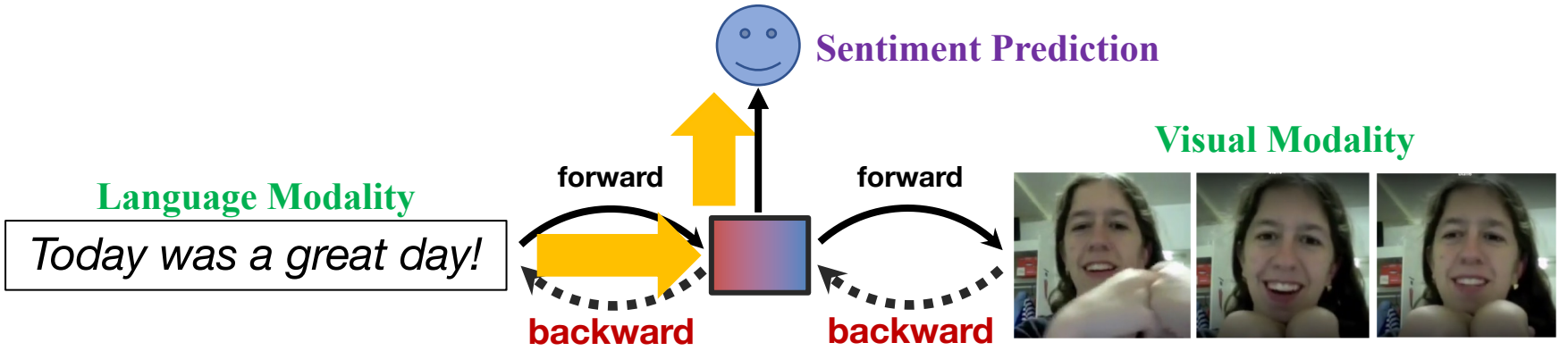


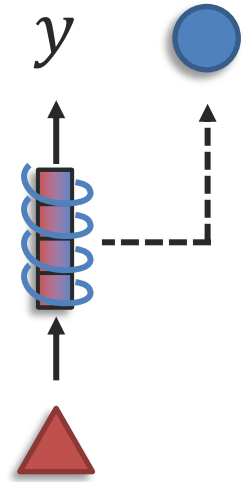
Problem: how do you ensure that both modalities are being used?

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Sub-Challenge 5c: Co-learning via Generation

Bimodal cyclic translations





Solution: cyclic translations from visual back to language

Cross-modal translation during training
Only language modality required at test time!

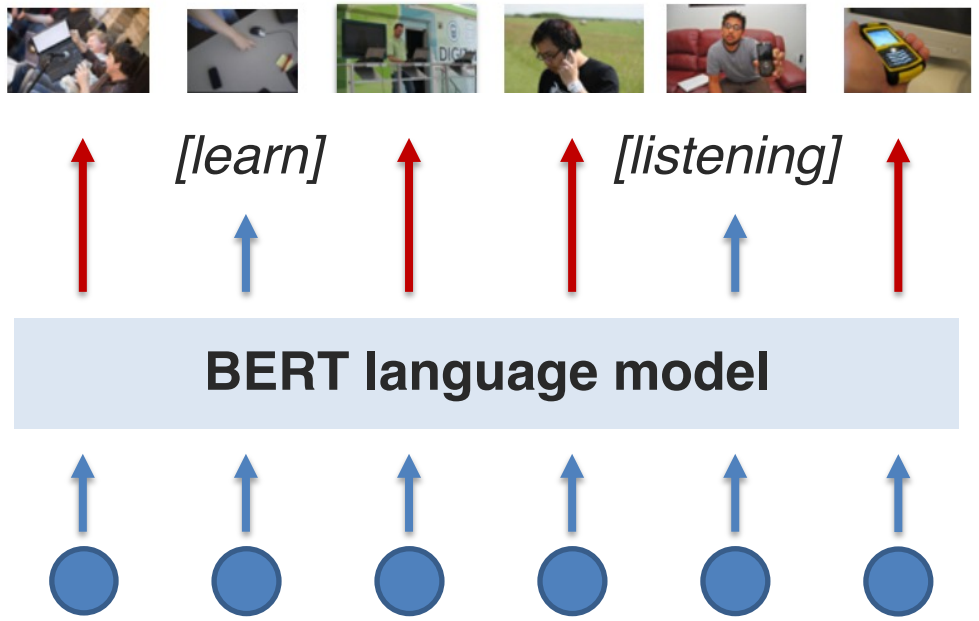
[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Sub-Challenge 5c: Co-learning via Generation

Predicting images from corresponding language

Voken (visual token) classification

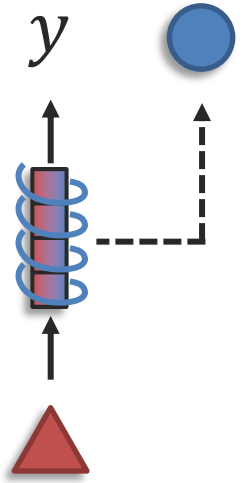
Masked language modeling



Humans [mask] language by [mask] speaking

Only text used at test-time

Multimodal co-learning > language-only training



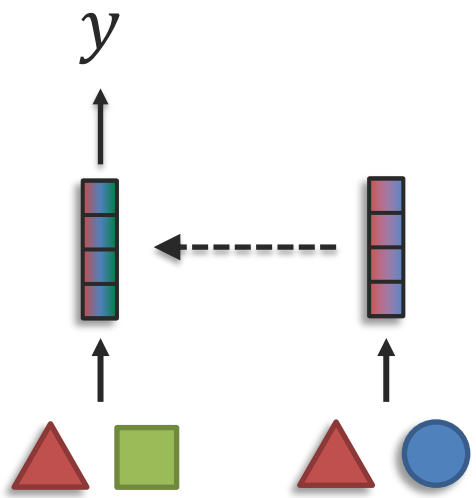
[Tan and Bansal, Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP 2020]

Summary: Transference

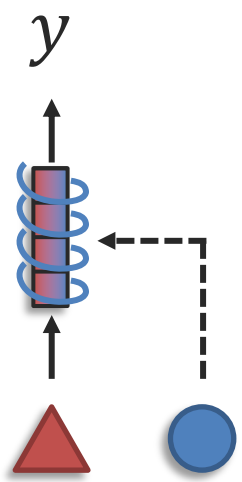
Definition: Transfer knowledge between modalities, usually to help the primary modality which may be noisy or with limited resources.

Sub-challenges:

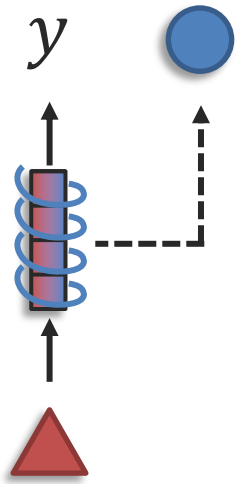
Transfer



Co-learning via representation



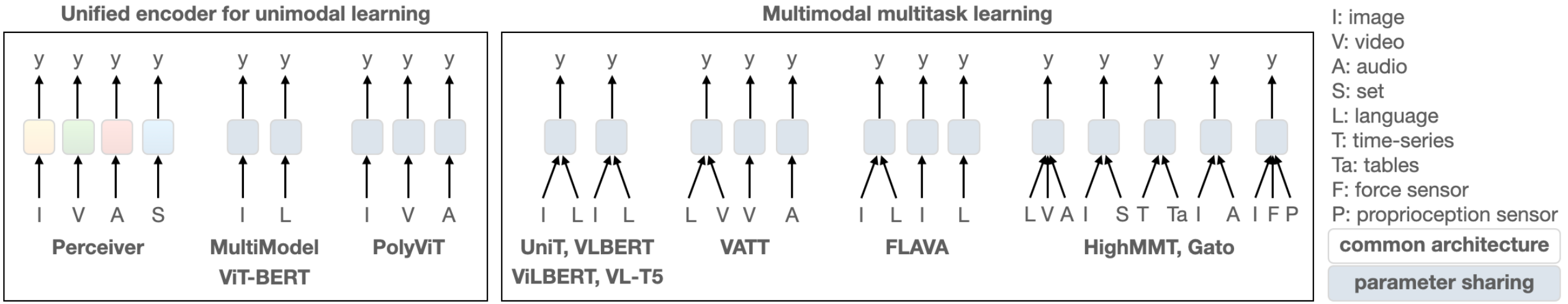
Co-learning via generation



More Transference

Open challenges

Many more dimensions of transfer



Open challenges:

- Low-resource: little downstream data, lack of paired data, robustness (next section)
- Beyond language and vision
- Settings where SOTA unimodal encoders are not deep learning e.g., tabular data
- Complexity in data, modeling, and training
- Interpretability (next section)