NEURAL INFORMATION
PROCESSING SYSTEMS

HELMHOLTZ
MUNICH

Roche

Training Transitive and Commutative Multimodal Transformers with LoReTTa

NeurIPS 2023 Presentation

Manuel Tran, Yashin Dicente Cid, Amal Lahiani, Fabian J. Theis, Tingying Peng, Eldad Klaiman
Roche Diagnostics, Helmholtz Munich, Technical University of Munich

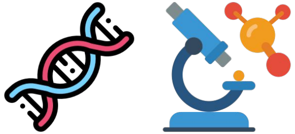
10.12.2023

Fully aligned modalities are only partially observed

Many domains collect only modalities (A, B) or (B, C), but never (A, C) or (A, B, C) together

Train: Clinic 1

Train: Clinic 2

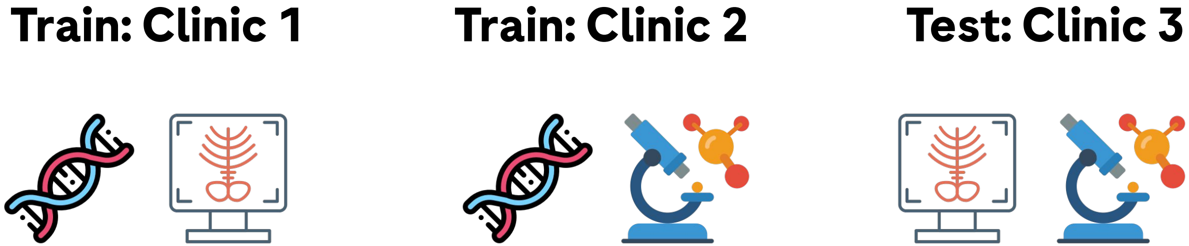


This cover has been designed using images from Flaticon.com

Tran et al., Training Transitive and Commutative Multimodal Transformers with LoReTTa, NeurIPS 2023

Fully aligned modalities are only partially observed

Many domains collect only modalities (A, B) or (B, C), but never (A, C) or (A, B, C) together



Train: Clinic 1

Train: Clinic 2

Test: Clinic 3

This cover has been designed using images from Flaticon.com

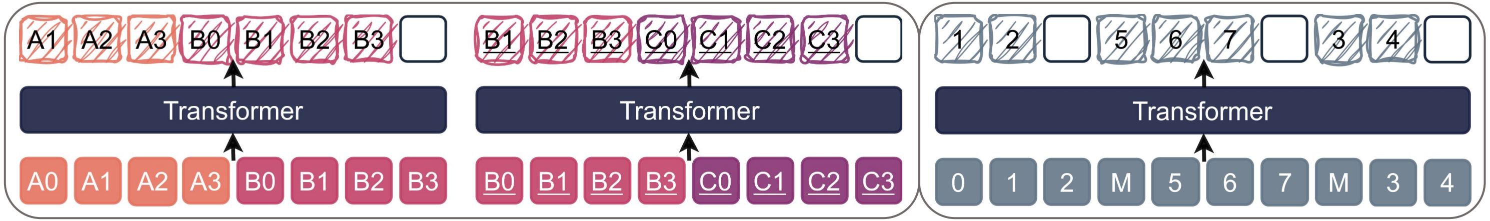
Tran et al., Training Transitive and Commutative Multimodal Transformers with LoReTTa, NeurIPS 2023

Causal, commutative, and masked modeling

We first start with generative modeling to generate modality A from B, B from A, and so on

Causal & Commutative Modeling

Causal Masked Modeling



Tokenize modalities A, B, and C. Predict the next token. Switch input order.

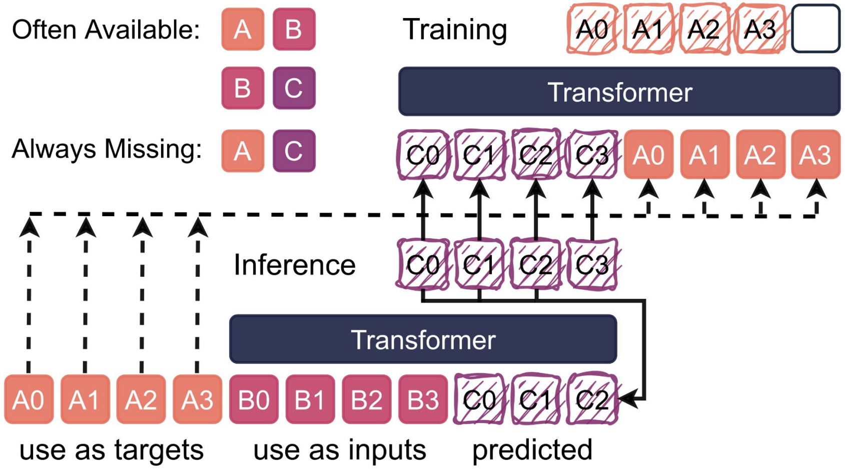
Move masked tokens to the end.

▨ Predicted tokens

Tran et al., Training Transitive and Commutative Multimodal Transformers with LoReTTa, NeurIPS 2023

Transitive modeling

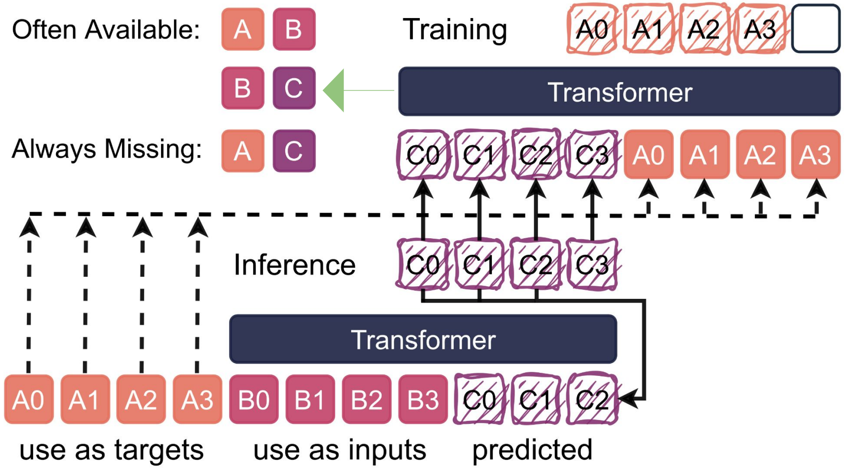
Given (A, B), we generate B ⟶ C ⟶ A and get (A, C)



Tran et al., Training Transitive and Commutative Multimodal Transformers with LoReTTa, NeurIPS 2023

Transitive modeling

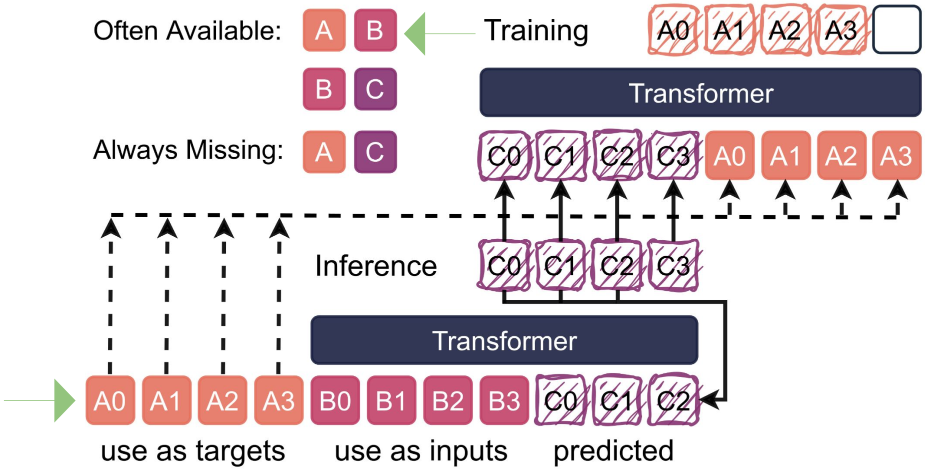
Given (A, B), we generate B ⟶ C ⟶ A and get (A, C)



Tran et al., Training Transitive and Commutative Multimodal Transformers with LoReTTa, NeurIPS 2023

Transitive modeling

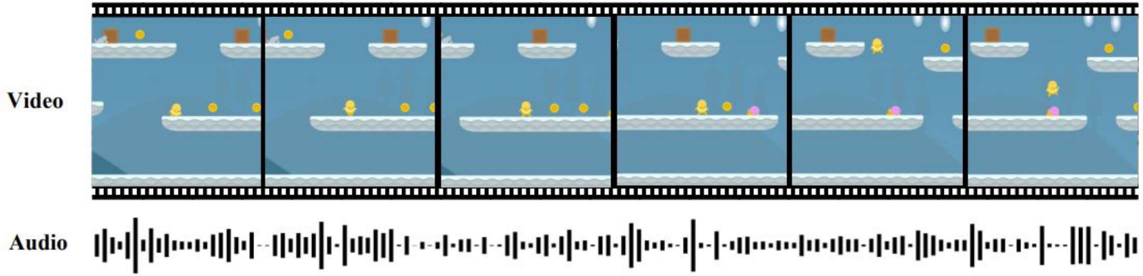
Given (A, B), we generate B ⟶ C ⟶ A and get (A, C)



Tran et al., Training Transitive and Commutative Multimodal Transformers with LoReTTa, NeurIPS 2023

Experiments

We have extensively evaluated our method on various datasets, but here we focus on MUGEN-GAME

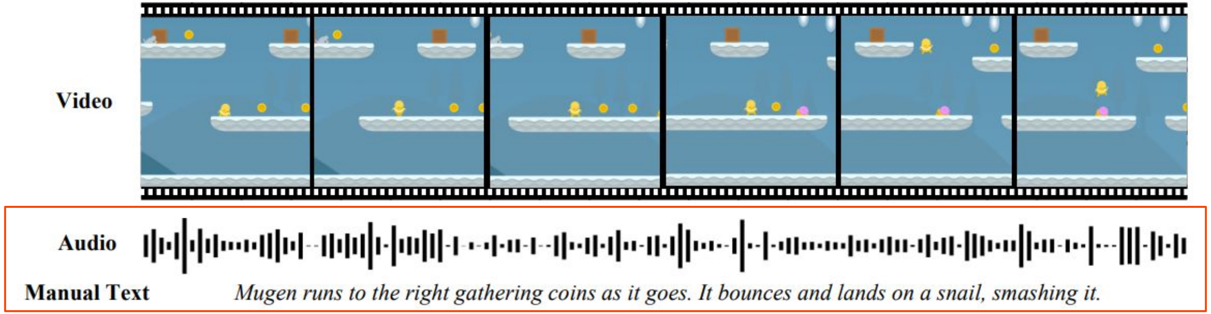


Manual Text *Mugen runs to the right gathering coins as it goes. It bounces and lands on a snail, smashing it.*

Hayes et al. (2022)

Experiments

We have extensively evaluated our method on various datasets, but here we focus on MUGEN-GAME

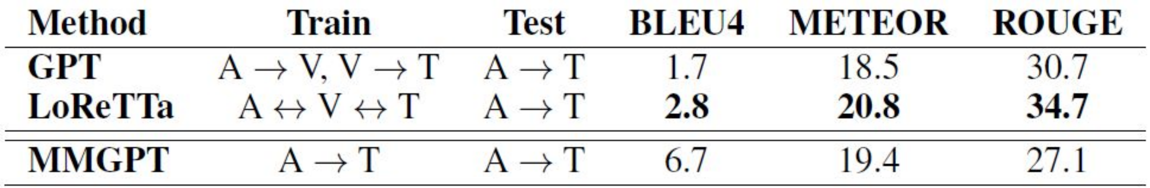


Mugen runs to the right gathering coins as it goes. It bounces and lands on a snail, smashing it.

Hayes et al. (2022)

Results on MUGEN-GAME

Train on disjoint (**A**udio, **V**ideo) and (**T**ext, **V**ideo), but test on (**A**udio, **T**ext)



Method	**Train**	**Test**	**BLEU4**	**METEOR**	**ROUGE**
GPT	A → V, V → T	A → T	1.7	18.5	30.7
LoReTTa	A ↔ V ↔ T	A → T	**2.8**	**20.8**	**34.7**
MMGPT	A → T	A → T	6.7	19.4	27.1

Tran et al., Training Transitive and Commutative Multimodal Transformers with LoReTTa, NeurIPS 2023

We introduced ***LoReTTa****, a powerful self-supervised method for combining missing mixtures of input modalities.*