How Intermodal Interaction Affects the Performance of Deep Multimodal Fusion for Mixed-Type Time Series

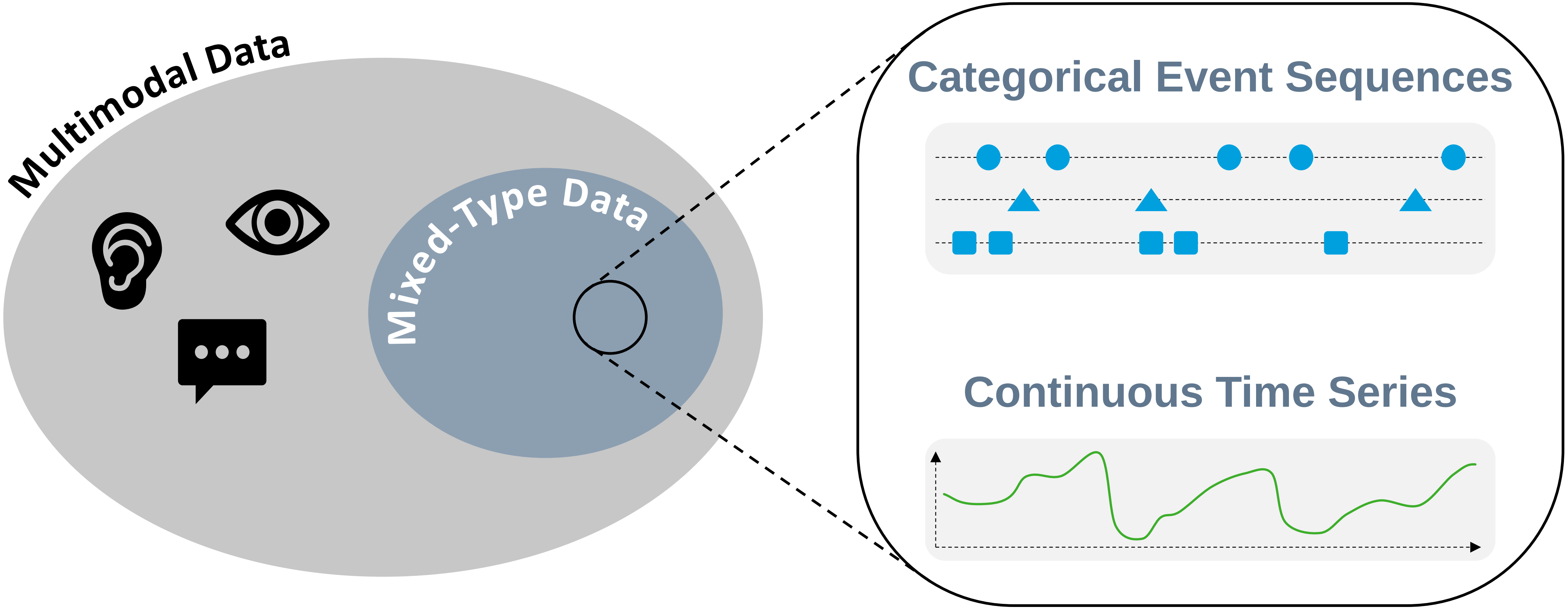
Simon Dietz, Thomas Altstidl, Dario Zanca, Bjorn Eskofier, An Nguyen

FAU Erlangen-Nurnberg, Germany

{simon.j.dietz, thomas.r.altstidl, dario.zanca, bjoern.eskofier, an.nguyen}@fau.de

Multimodal Data

What is mixed-type time series?



Multimodal Data

Where can we find it?

Healthcare

- Medical Examinations
- Clinical Measurements

Human-Computer Interaction

- Button clicks
- Accelerometers

Industry

- Event logs
- Sensor measurements

Multimodal Data

Why is it important?

The presence of one modality can influence the perception of another

E.g. **MCGurck-Effect**

Letter | Published: 23 December 1976

Hearing lips and seeing voices

HARRY MCGURK & JOHN MACDONALD

Nature **264**, 746–748 (1976) | Cite this article

44k Accesses | **184** Altmetric | Metrics

2 Videos, same audio + **Different lip movements** → **Different perceived audio**

Fusion Types and Methods

Fusion Type

When are modalities fused

- **Late Fusion**
- **Intermediate Fusion**
- **Early Fusion**

Fusion Method

How are modalities fused

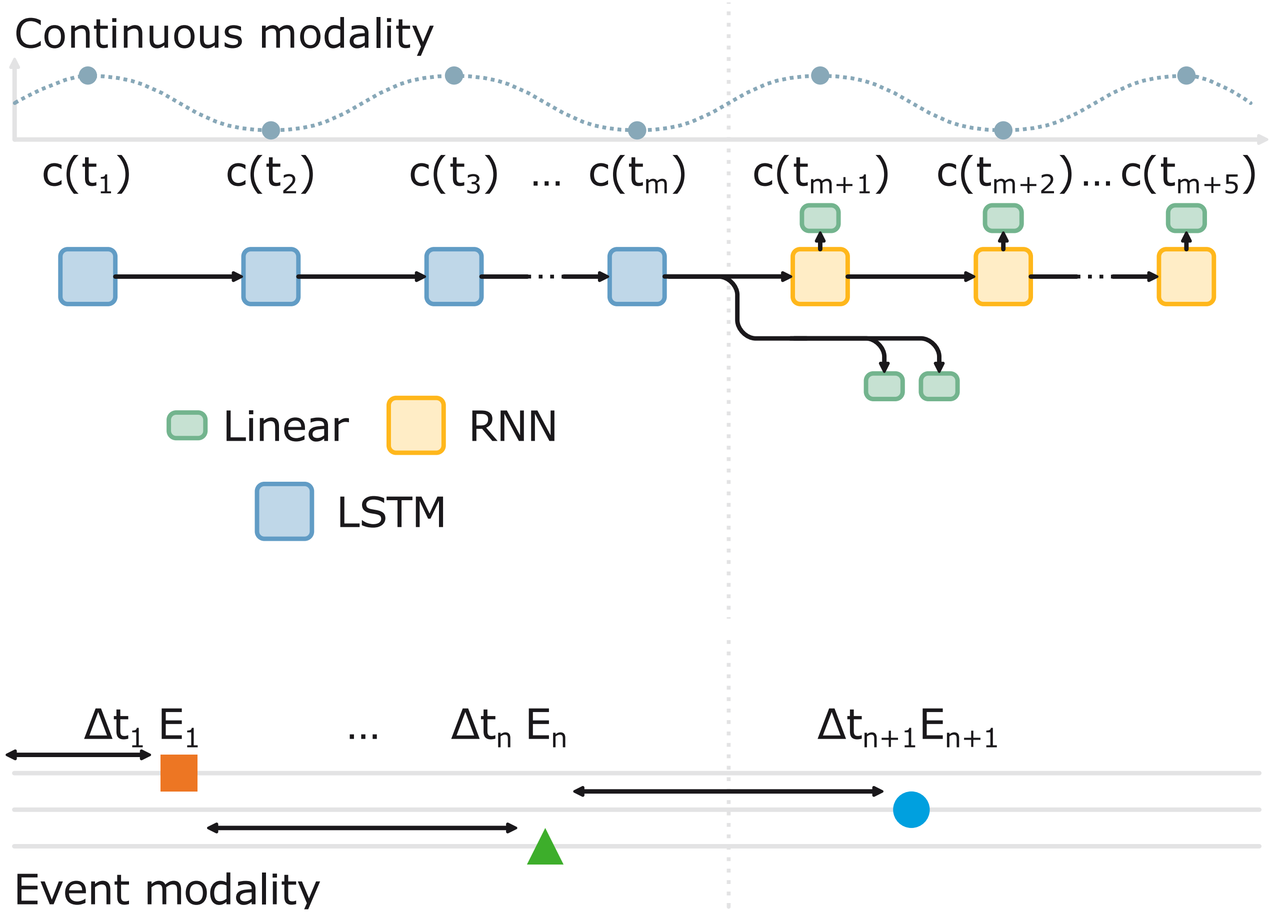
- **Concatenation**
- **Weighted Mean**
- **Gating**
- **…**

Main Topics

How does the nature of the data
influece the optimal fusion approach

Fusion Types

Unimodal (no fusion)

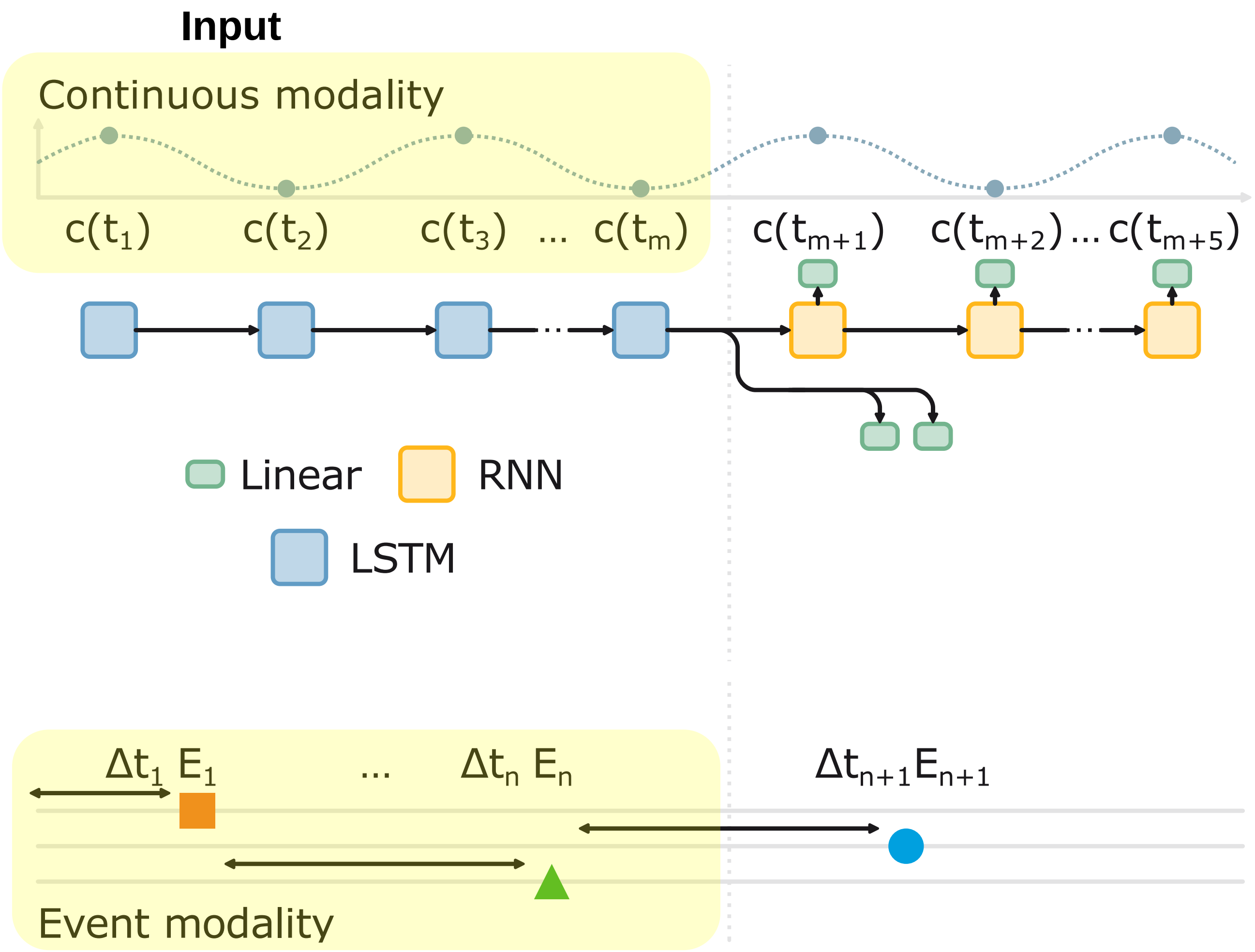


Unimodal Baselines

- No Fusion
- Forecast based on a single modality

Fusion Types

Unimodal (no fusion)

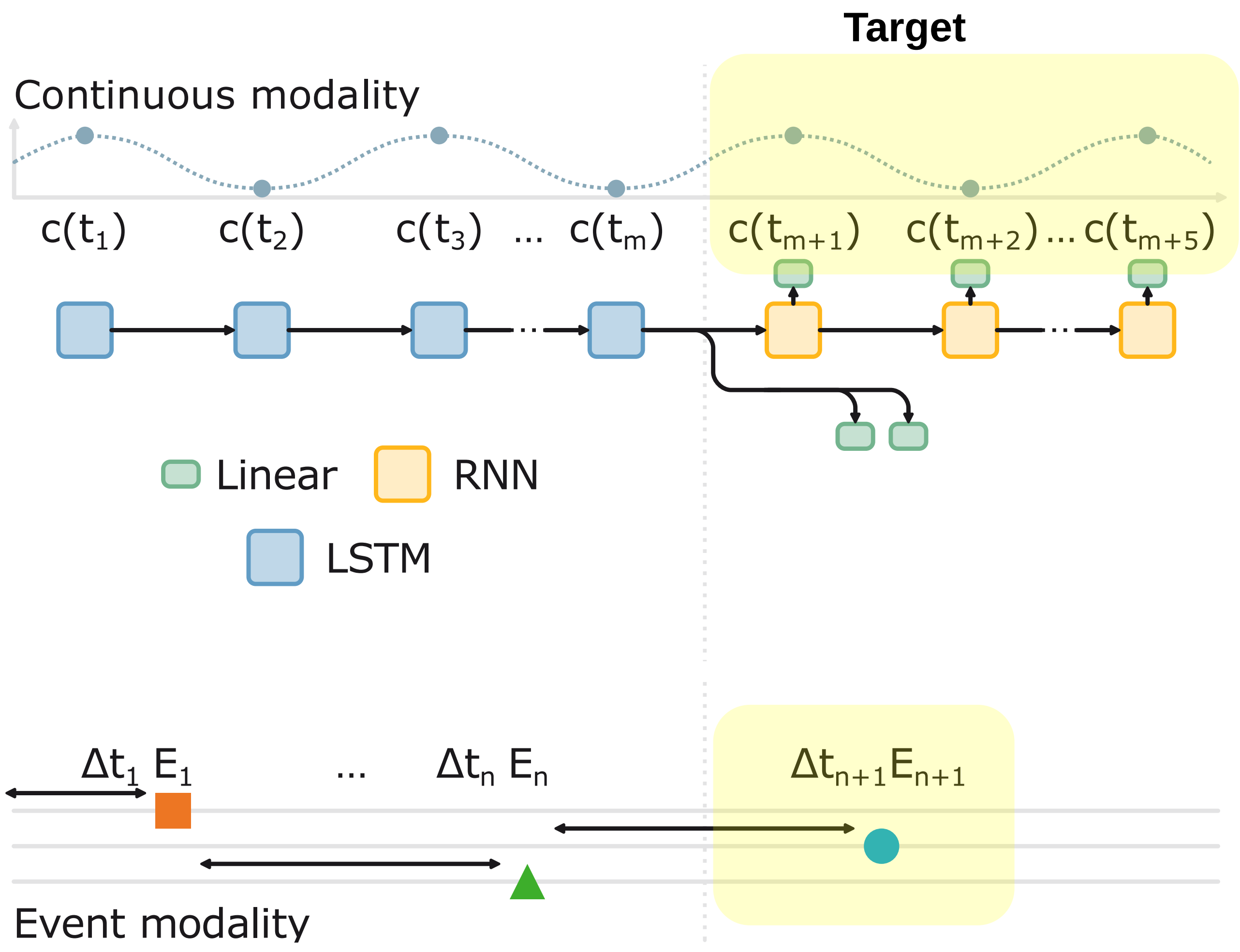


Unimodal Baselines

- No Fusion
- Forecast based on a single modality

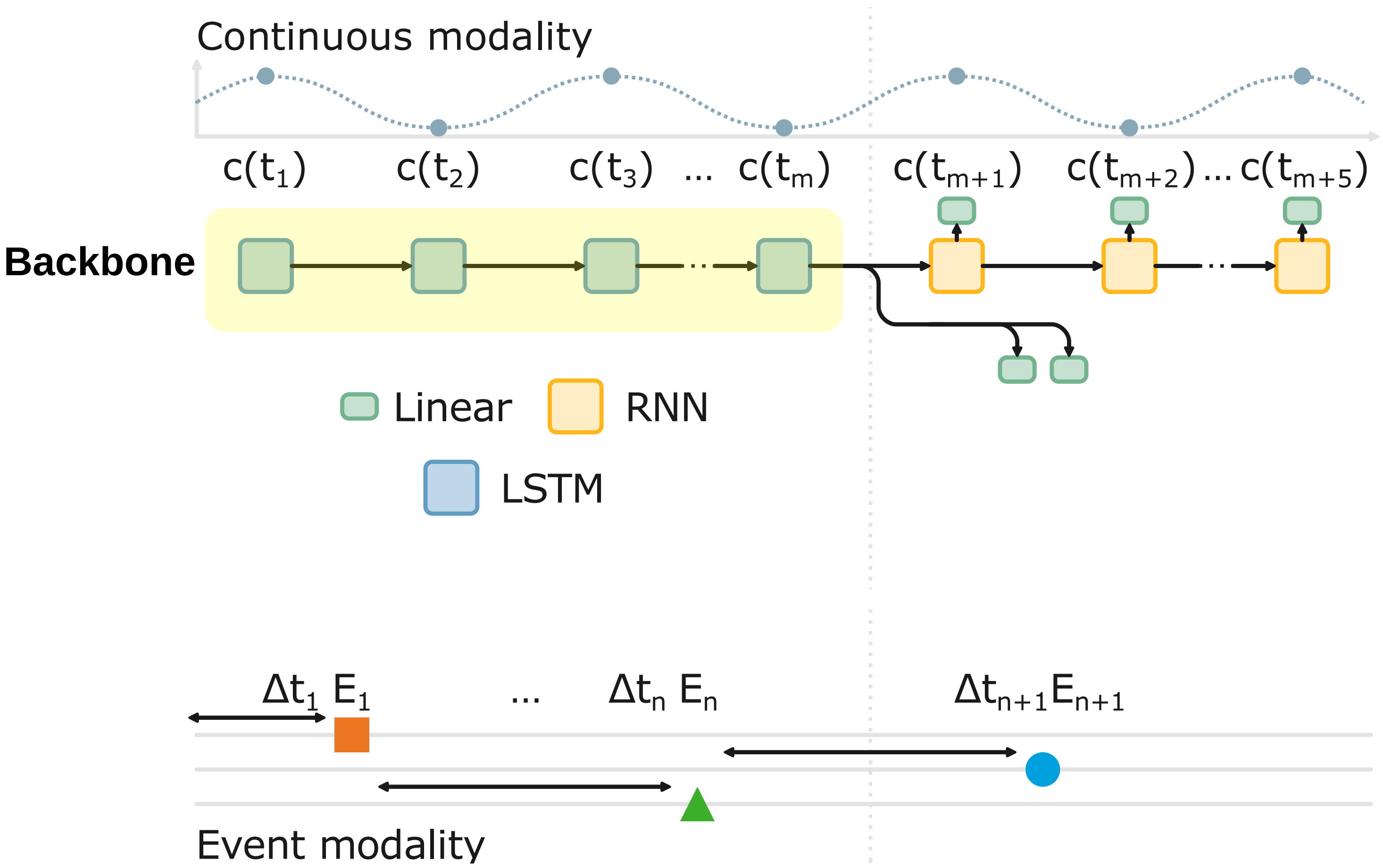
Fusion Types

Unimodal (no fusion)



Unimodal Baselines

- No Fusion
- Forecast based on a single modality

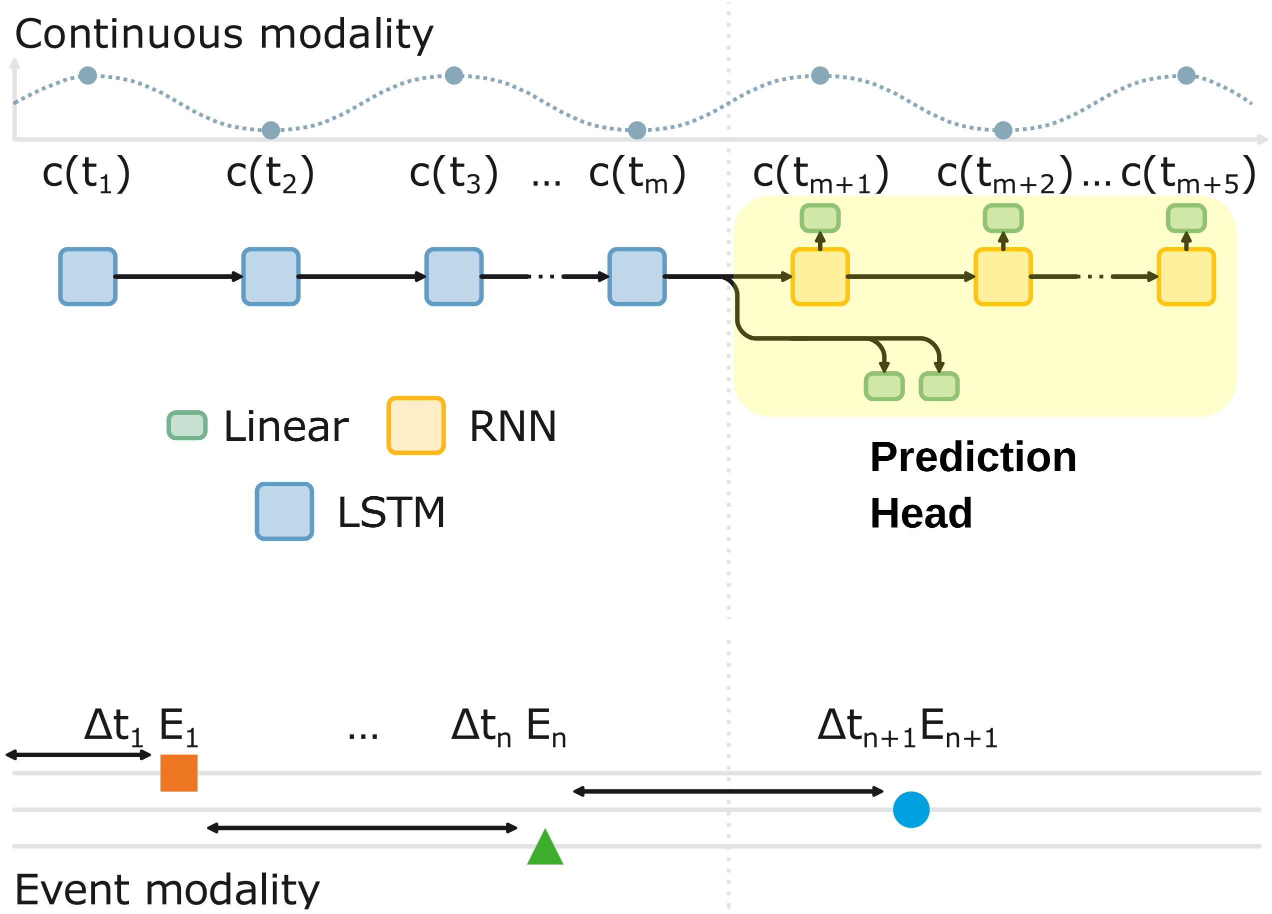


Unimodal Baselines

- No Fusion
- Forecast based on a single modality

Fusion Types

Unimodal (no fusion)

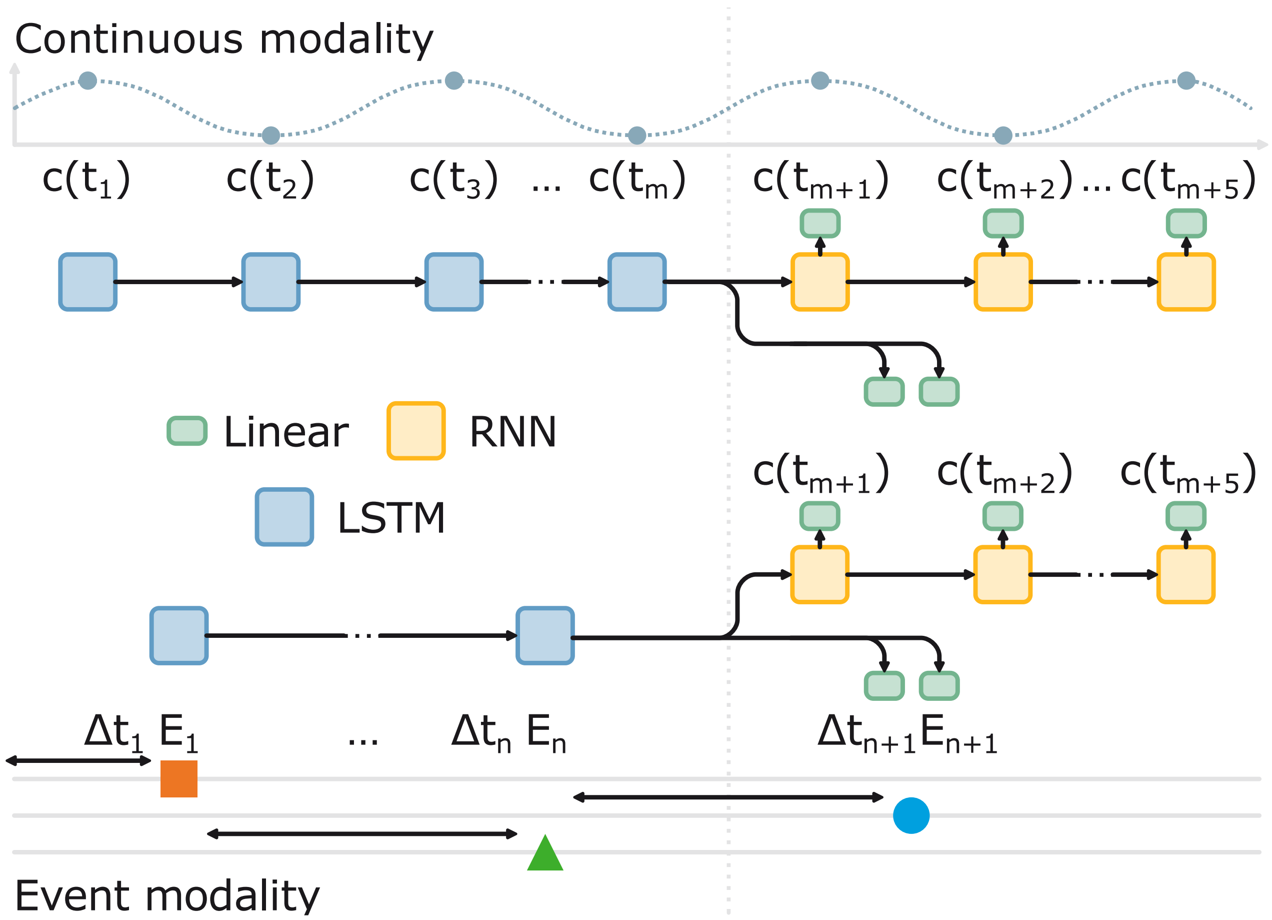


Unimodal Baselines

- No Fusion
- Forecast based on a single modality

Fusion Types

Unimodal (no fusion)

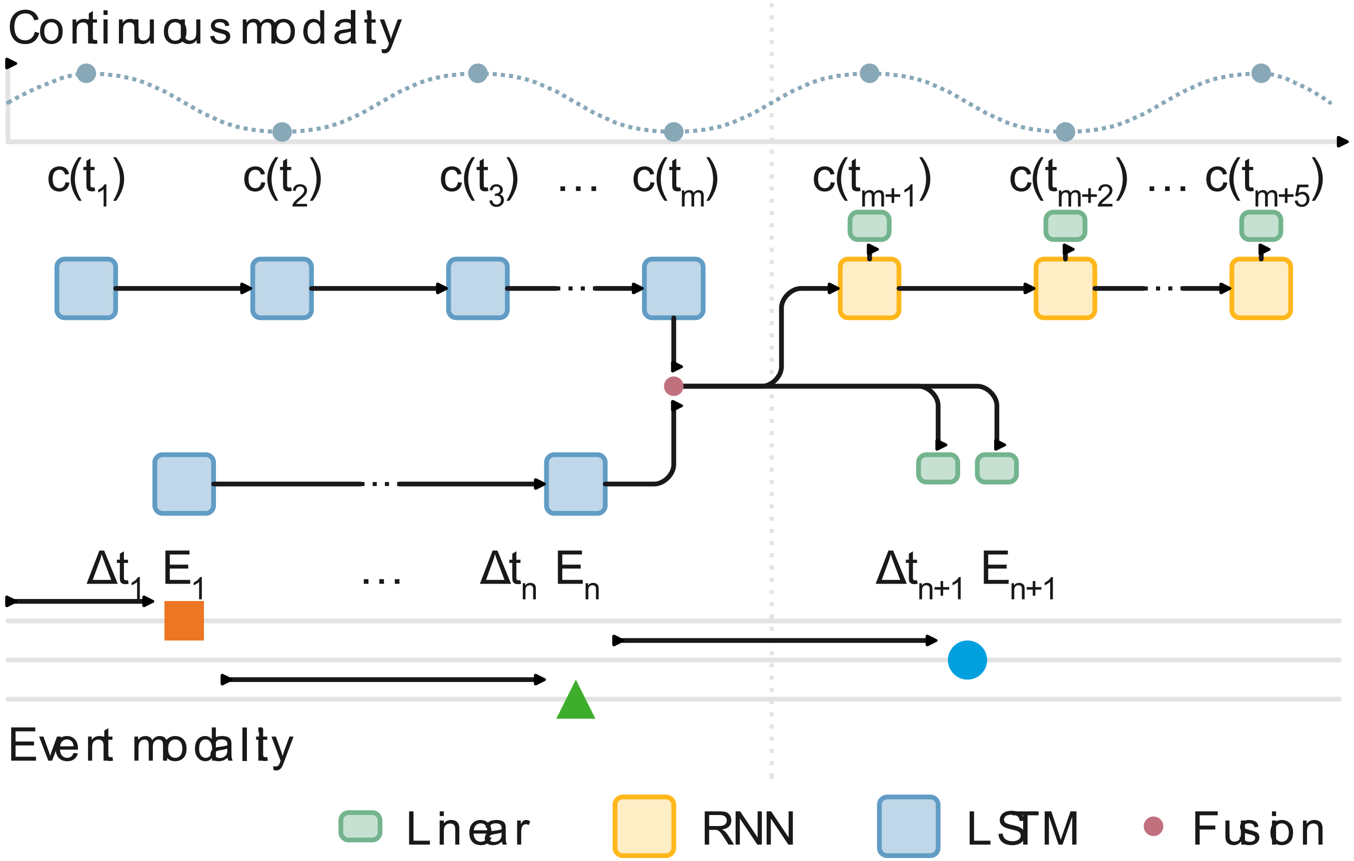


Unimodal Baselines

- No Fusion
- Forecast based on a single modality

Fusion Types

Intermediate Fusion

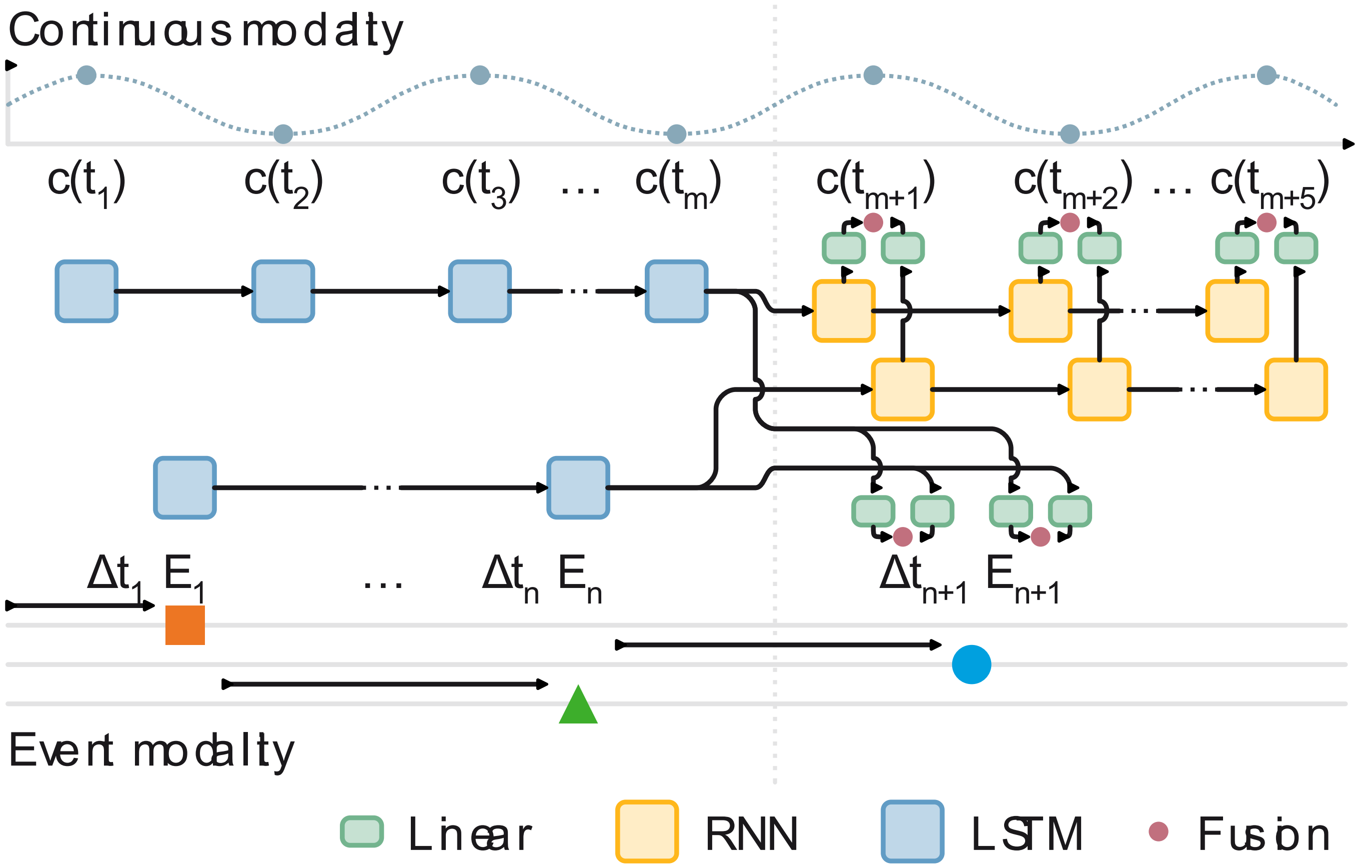


Intermediate Fusion

- 2 Backbones + 1 Prediction Head
- Fuse final unimodal representations
- One forecast based on both modalities

Fusion Types

Late Fusion

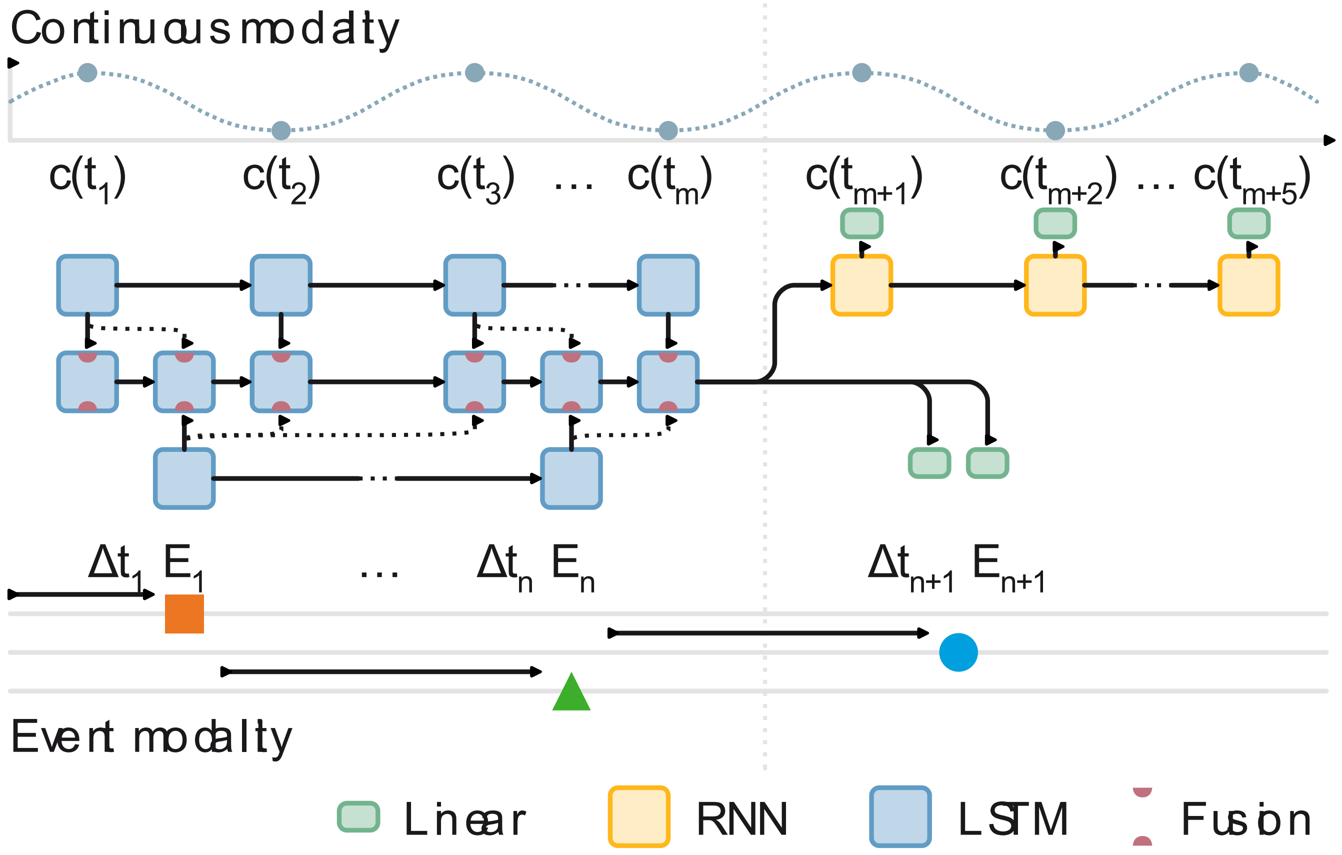


Late Fusion

- 2 Backbones + 2 Prediction Heads
- One model per modality
- Unimodal predictions are fused

Fusion Types

Early Fusion



Early Fusion

- 1 common Backbone + 1 Prediction Head
- 2 unimodal + 1 multimodal LSTM
- Unimodal representations are fused at each timestep

Fusion Methods

- Concatenation
- Weighted mean
- Weighted mean with correlation [Yang 2017]
- Gating [Arevalo 2017, Narayanan 2020]
- Feature sharing [Wang 2015]

Deep Multimodal Representation Learning from Temporal Data

Xitong Yang*[1], Palghat Ramesh[2], Radha Chitta*[3], Sriganesh Madhvanath*[3], Edgar A. Bernal*[4] and Jiebo Luo[5]

[1]University of Maryland, College Park [2]PARC [3]Conduent Labs US
[4]United Technologies Research Center [5]University of Rochester

[1]xyang35@cs.umd.edu, [2]Palghat.Ramesh@parc.com, [3]{Radha.Chitta, Sriganesh.Madhvanath}@conduent.com, [4]bernalea@utrc.utc.com, [5]jluo@cs.rochester.edu

GATED MULTIMODAL UNITS FOR INFORMATION FUSION

Arevalo, John
Dept. of Computing Systems and Industrial Engineering
Universidad Nacional de Colombia
Cra 30 No 45 03-Ciudad Universitaria
jearevaloo@unal.edu.co

Solorio, Thamar
Dept. of Computer Science
University of Houston
Houston, TX 77204-3010
solorio@cs.uh.edu

Montes-y-Gómez, Manuel
Instituto Nacional de Astrofísica, Óptica y Electrónica
Computer Science Department
Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla
C.P. 72840 Puebla, Mexico
smmontesg@inaoep.mx

González, Fabio A.
Dept. of Computing Systems and Industrial Engineering
Universidad Nacional de Colombia
Cra 30 No 45 03-Ciudad Universitaria
fagonzalezo@unal.edu.co

MMSS: Multi-modal Sharable and Specific Feature Learning for RGB-D Object Recognition

Anran Wang[1], Jianfei Cai[1], Jiwen Lu[2], and Tat-Jen Cham[1]
[1] School of Computer Engineering, Nanyang Technological University, Singapore
[2] Department of Automation, Tsinghua University, Beijing, China

IEEE ROBOTICS AND AUTOMATION LETTERS, VOL. 5, NO. 2, APRIL 2020 1287

Gated Recurrent Fusion to Learn Driving Behavior from Temporal Multimodal Data

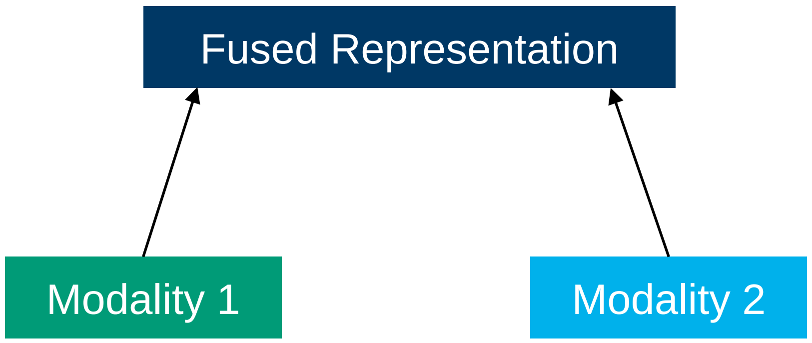
Athma Narayanan, Avinash Siravuru, and Behzad Dariush

Fusion Methods

1. Concatenation

- Simple approach that propagates the maximum amount of information.

- Can result in large representations with redundant features.

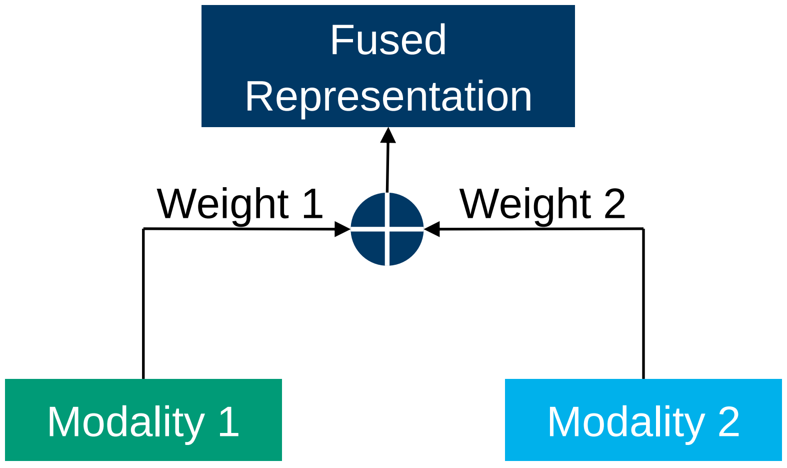


Fusion Methods

1. Concatenation

2. **Weighted Mean**

- Representations need to be of equal size.
- Smaller feature size compared to concatenation.
- Weighting can account for modality importance.



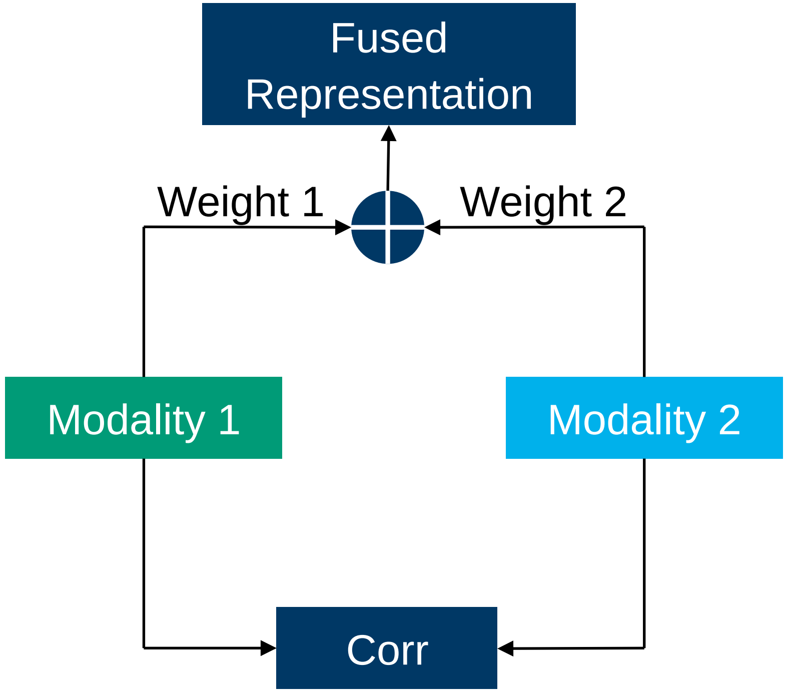
Fusion Methods

1. Concatenation
2. Weighted Mean
3. **Weighted Mean with coordinated representations**

- Calculate correlation and subtract from prediction loss.

$$L = L_{prediction} - \tau \, Corr$$

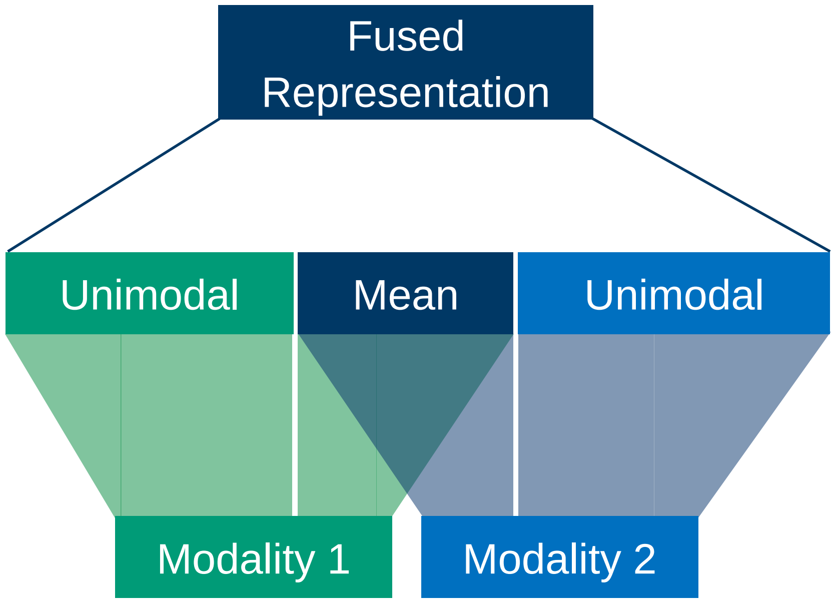
- Coordinating representations can improve the performance [Yan17].



Fusion Methods

1. Concatenation
2. Weighted Mean
3. Weighted Mean with coordinated representations.
4. **Shared Features**

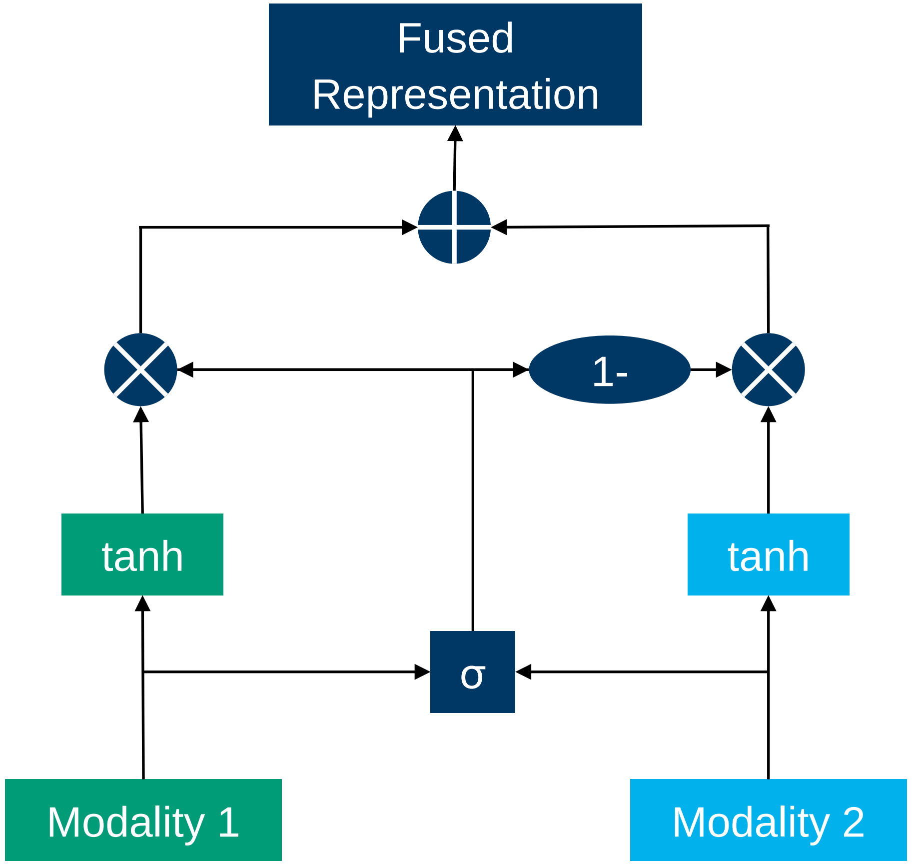
- Middle ground between concatenation and averaging.
- Based on Wan et. al. [Wan15].



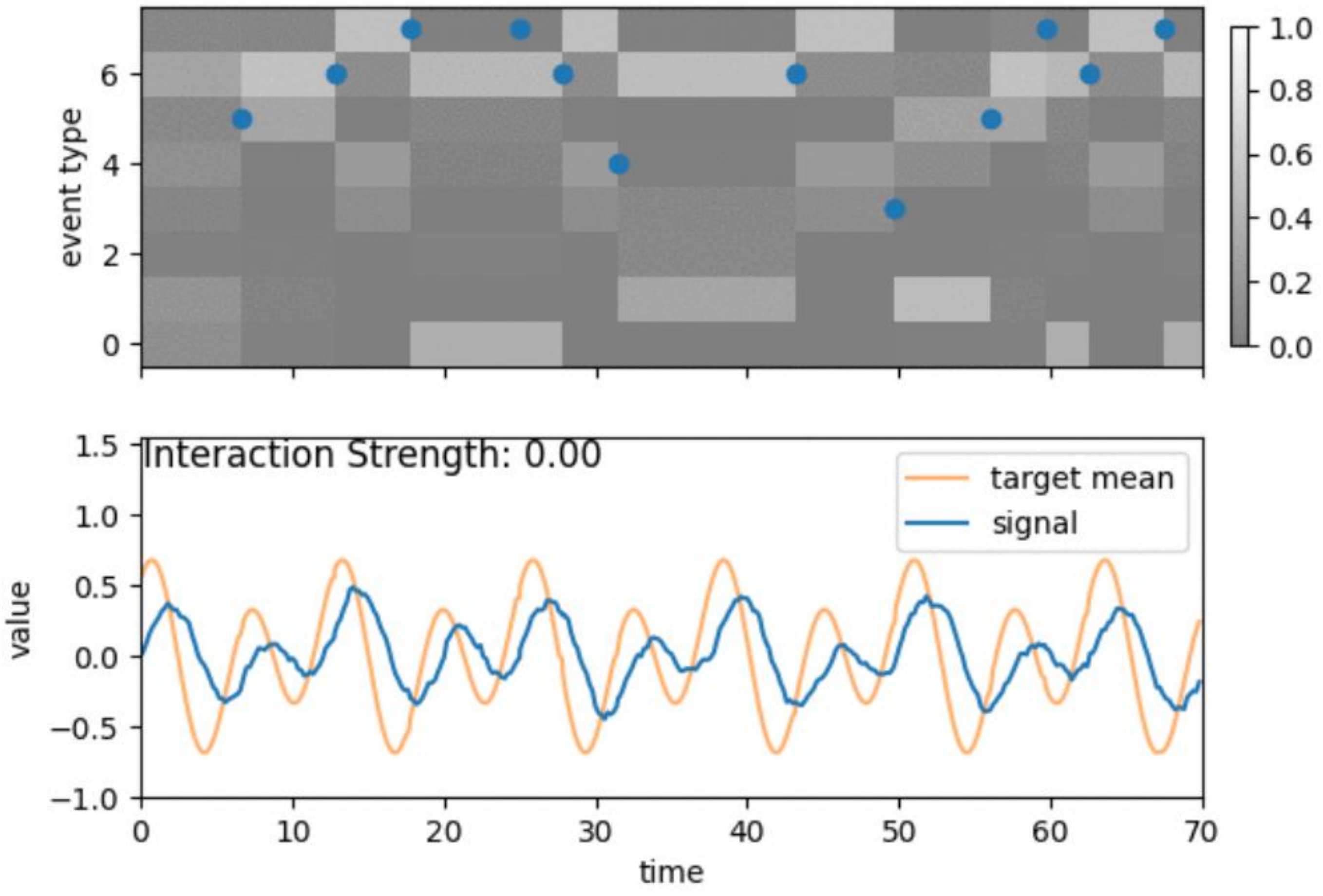
Fusion Methods

1. Concatenation
2. Weighted Mean
3. Weighted Mean with coordinated representations.
4. Shared Features
5. **Gating**

- Flexible weighting of each feature.
- Can help with model explainability.
- Proposed by Arevalo et. al. [Are17].



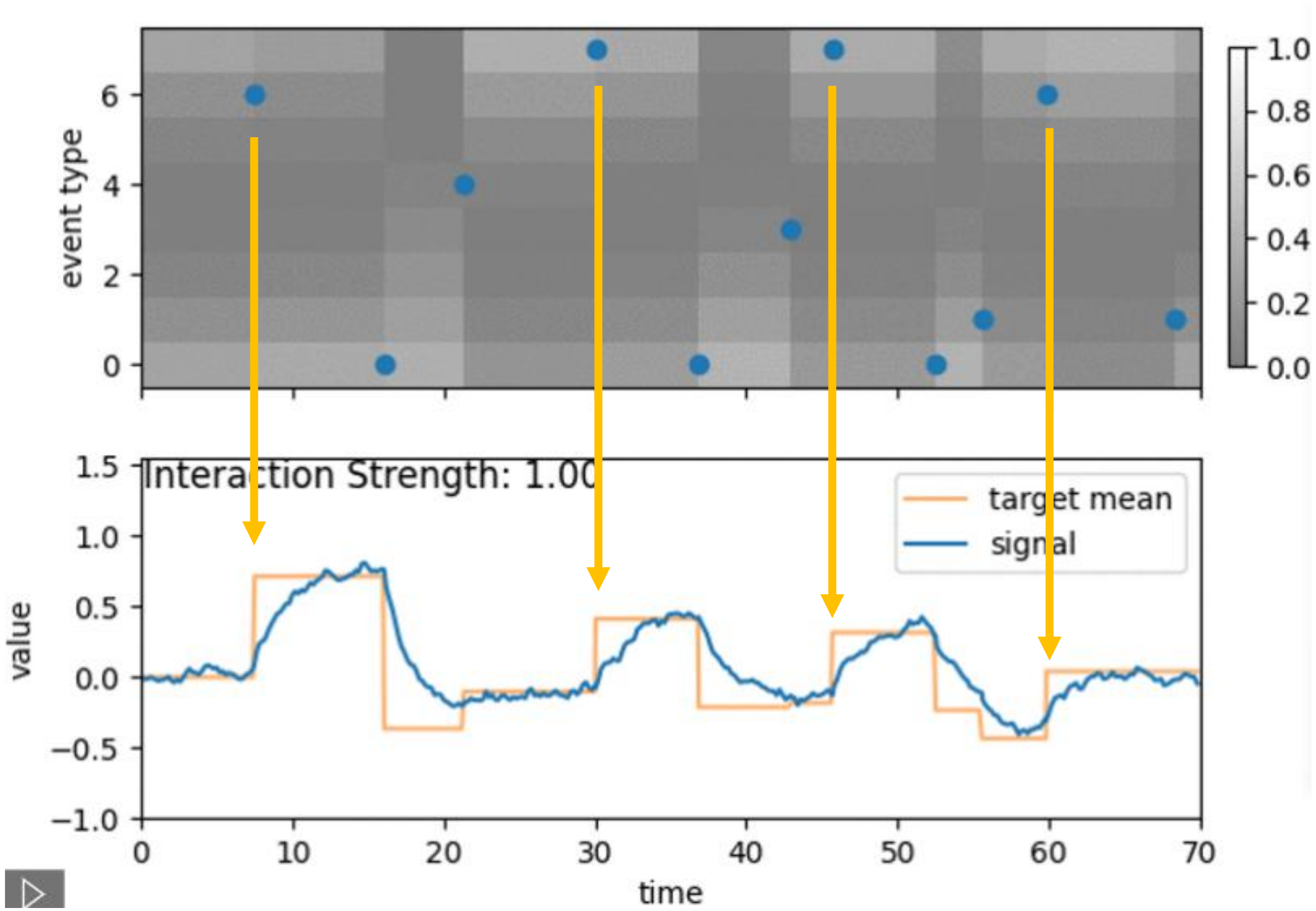
Synthetic Data



No Interaction

- Fixed transition prop.
  most likely transitions:
  - 5->6, 6->7, 7->6
- Sinusoidal base Signal

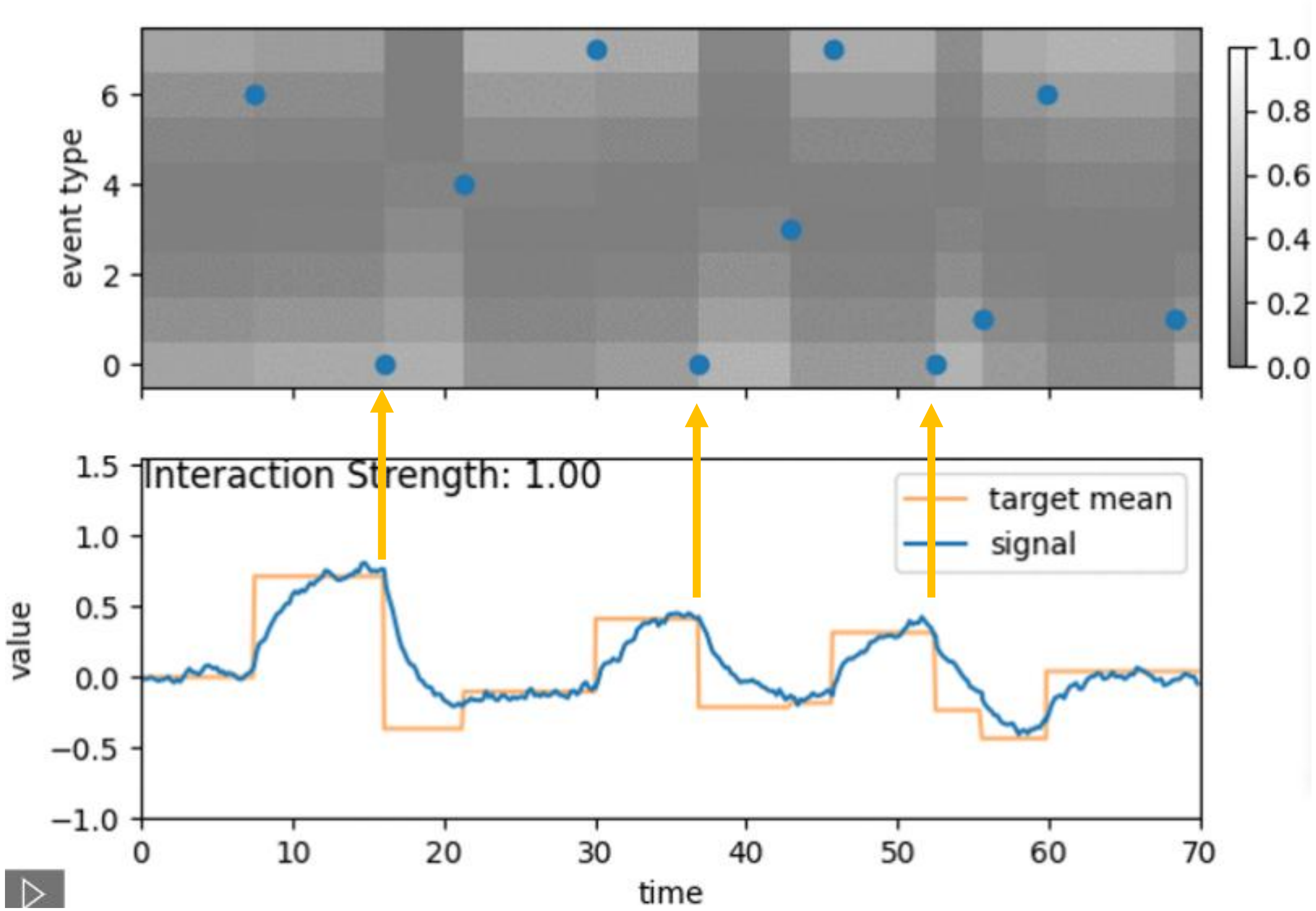
Synthetic Data



Full Interaction

- Event 7&6 → increase in target mean

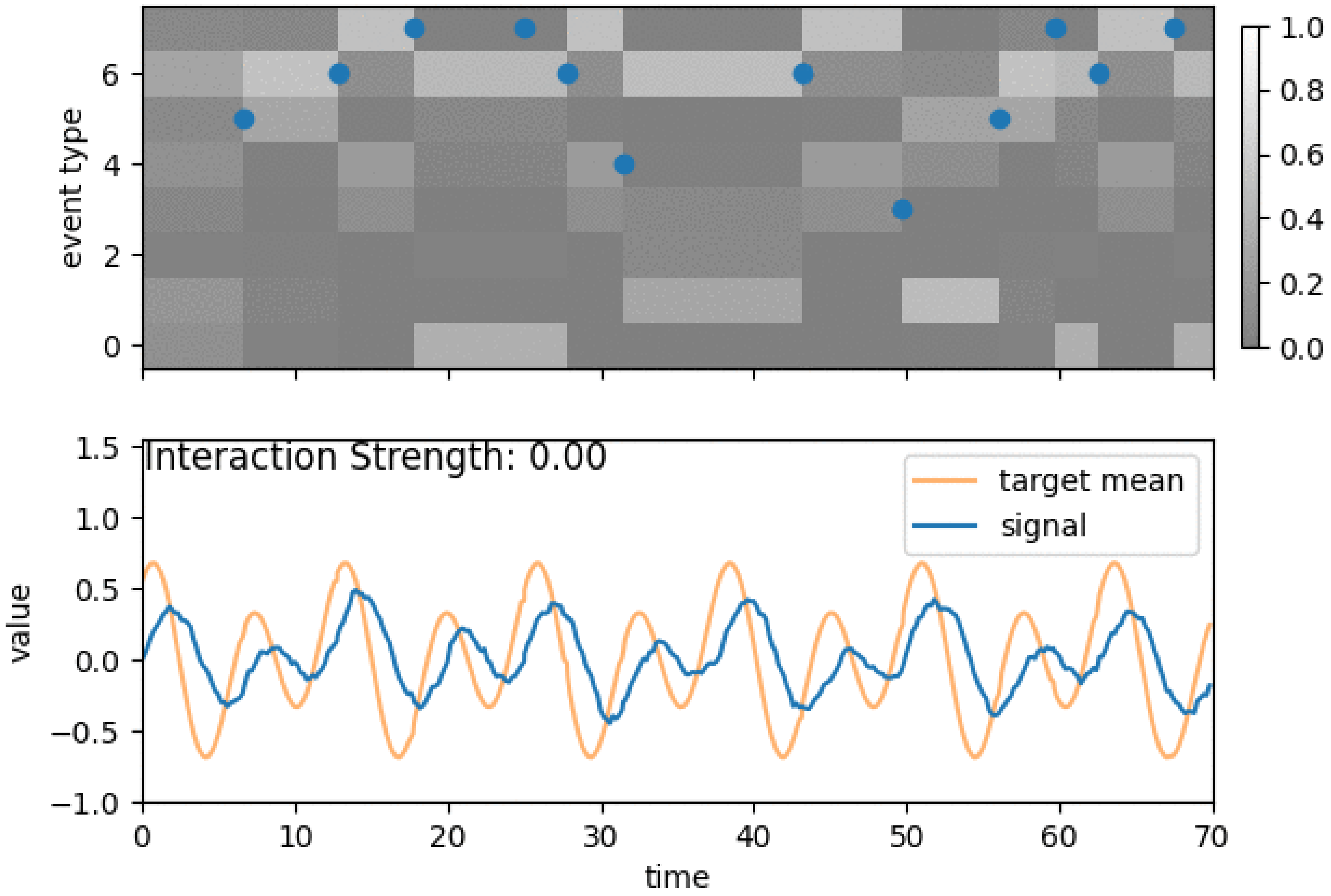
Synthetic Data



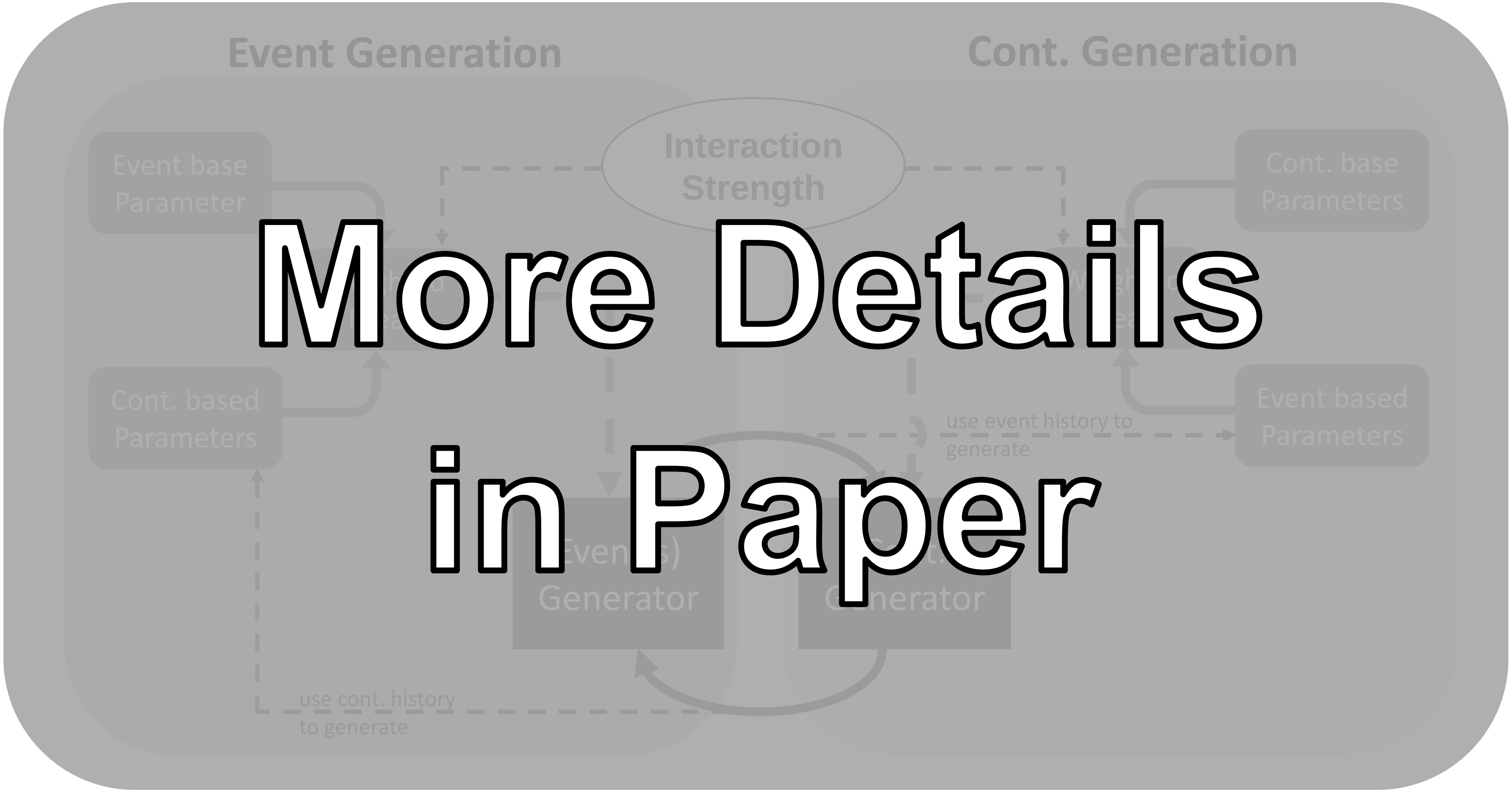
Full Interaction

- Event 7&6 → increase in target mean
- High cont. Signal → Event 0 is likely

Synthetic Data



Synthetic Data



More Details in Paper

Mixed-Type Time Series Forecasting

Results

type	method	synthetic MTTS			electrical grid			MazeBall		
		cont.	Δt	event	cont.	Δt	event	cont.	Δt	event
early	cat.	*0.044*	**0.067**	**0.899**	1.623	*0.056*	0.728	<u>2.867</u>	**0.105**	**0.717**
	mean	**0.040**	<u>0.082</u>	*0.887*	1.603	0.057	0.735	3.912	<u>0.109</u>	0.702
	corr.	0.046	0.093	0.887	1.886	0.059	0.710	3.348	0.115	0.696
	gating	<u>0.041</u>	*0.084*	<u>0.887</u>	*1.574*	<u>0.056</u>	*0.738*	3.463	*0.109*	0.703
	share	0.050	0.099	0.878	1.592	0.057	0.737	3.106	0.111	0.705
inter.	cat.	0.049	0.098	0.876	<u>1.573</u>	0.057	<u>0.740</u>	3.747	0.115	0.704
	mean	0.055	0.111	0.855	1.591	0.057	0.731	3.193	0.115	0.692
	corr.	0.045	0.085	0.880	1.738	0.059	0.638	4.296	0.112	0.704
	gating	0.049	0.087	0.872	1.723	0.056	0.710	*2.909*	0.110	0.702
	share	0.050	0.099	0.873	**1.552**	**0.055**	**0.742**	2.924	0.113	0.704
late	mean	0.059	0.171	0.759	1.751	0.062	0.719	**2.626**	0.113	0.705
	corr.	0.058	0.162	0.763	1.823	0.062	0.704	3.486	0.113	*0.706*
uni.	cont.	0.049	0.472	0.680	1.826	0.074	0.579	3.763	0.154	0.664
	event	0.146	0.445	0.622	2.692	0.059	0.737	5.060	0.109	<u>0.708</u>

Mixed-Type Time Series Forecasting

Results

Datasets

type	method	synthetic MTTS			electrical grid			MazeBall		
		cont.	Δt	event	cont.	Δt	event	cont.	Δt	event
early	cat.	*0.044*	**0.067**	**0.899**	1.623	*0.056*	0.728	2.867	**0.105**	**0.717**
	mean	**0.040**	0.082	*0.887*	1.603	0.057	0.735	3.912	0.109	0.702
	corr.	0.046	0.093	0.887	1.886	0.059	0.710	3.348	0.115	0.696
	gating	0.041	*0.084*	0.887	*1.574*	0.056	*0.738*	3.463	*0.109*	0.703
	share	0.050	0.099	0.878	1.592	0.057	0.737	3.106	0.111	0.705
inter.	cat.	0.049	0.098	0.876	1.573	0.057	0.740	3.747	0.115	0.704
	mean	0.055	0.111	0.855	1.591	0.057	0.731	3.193	0.115	0.692
	corr.	0.045	0.085	0.880	1.738	0.059	0.638	4.296	0.112	0.704
	gating	0.049	0.087	0.872	1.723	0.056	0.710	*2.909*	0.110	0.702
	share	0.050	0.099	0.873	**1.552**	**0.055**	**0.742**	2.924	0.113	0.704
late	mean	0.059	0.171	0.759	1.751	0.062	0.719	**2.626**	0.113	0.705
	corr.	0.058	0.162	0.763	1.823	0.062	0.704	3.486	0.113	*0.706*
uni.	cont.	0.049	0.472	0.680	1.826	0.074	0.579	3.763	0.154	0.664
	event	0.146	0.445	0.622	2.692	0.059	0.737	5.060	0.109	0.708

Mixed-Type Time Series Forecasting

Results

Forecasting Metric   Datasets

type	method	synthetic MTTS			electrical grid			MazeBall		
		cont.	Δt	event	cont.	Δt	event	cont.	Δt	event
early	cat.	*0.044*	**0.067**	**0.899**	1.623	*0.056*	0.728	2.867	**0.105**	**0.717**
	mean	**0.040**	0.082	*0.887*	1.603	0.057	0.735	3.912	0.109	0.702
	corr.	0.046	0.093	0.887	1.886	0.059	0.710	3.348	0.115	0.696
	gating	0.041	*0.084*	0.887	*1.574*	0.056	*0.738*	3.463	*0.109*	0.703
	share	0.050	0.099	0.878	1.592	0.057	0.737	3.106	0.111	0.705
inter.	cat.	0.049	0.098	0.876	1.573	0.057	0.740	3.747	0.115	0.704
	mean	0.055	0.111	0.855	1.591	0.057	0.731	3.193	0.115	0.692
	corr.	0.045	0.085	0.880	1.738	0.059	0.638	4.296	0.112	0.704
	gating	0.049	0.087	0.872	1.723	0.056	0.710	*2.909*	0.110	0.702
	share	0.050	0.099	0.873	**1.552**	**0.055**	**0.742**	2.924	0.113	0.704
late	mean	0.059	0.171	0.759	1.751	0.062	0.719	**2.626**	0.113	0.705
	corr.	0.058	0.162	0.763	1.823	0.062	0.704	3.486	0.113	*0.706*
uni.	cont.	0.049	0.472	0.680	1.826	0.074	0.579	3.763	0.154	0.664
	event	0.146	0.445	0.622	2.692	0.059	0.737	5.060	0.109	0.708

Mixed-Type Time Series Forecasting

Results

Fusion Types

type	method	synthetic MTTS cont.	synthetic MTTS Δt	synthetic MTTS event	electrical grid cont.	electrical grid Δt	electrical grid event	MazeBall cont.	MazeBall Δt	MazeBall event
early	cat.	*0.044*	**0.067**	**0.899**	1.623	*0.056*	0.728	<u>2.867</u>	**0.105**	**0.717**
	mean	**0.040**	<u>0.082</u>	*0.887*	1.603	0.057	0.735	3.912	<u>0.109</u>	0.702
	corr.	0.046	0.093	0.887	1.886	0.059	0.710	3.348	0.115	0.696
	gating	<u>0.041</u>	*0.084*	<u>0.887</u>	*1.574*	<u>0.056</u>	*0.738*	3.463	*0.109*	0.703
	share	0.050	0.099	0.878	1.592	0.057	0.737	3.106	0.111	0.705
inter.	cat.	0.049	0.098	0.876	<u>1.573</u>	0.057	<u>0.740</u>	3.747	0.115	0.704
	mean	0.055	0.111	0.855	1.591	0.057	0.731	3.193	0.115	0.692
	corr.	0.045	0.085	0.880	1.738	0.059	0.638	4.296	0.112	0.704
	gating	0.049	0.087	0.872	1.723	0.056	0.710	*2.909*	0.110	0.702
	share	0.050	0.099	0.873	**1.552**	**0.055**	**0.742**	2.924	0.113	0.704
late	mean	0.059	0.171	0.759	1.751	0.062	0.719	**2.626**	0.113	0.705
	corr.	0.058	0.162	0.763	1.823	0.062	0.704	3.486	0.113	*0.706*
uni.	cont.	0.049	0.472	0.680	1.826	0.074	0.579	3.763	0.154	0.664
	event	0.146	0.445	0.622	2.692	0.059	0.737	5.060	0.109	<u>0.708</u>

Mixed-Type Time Series Forecasting

Results

Fusion Types

Fusion Methods

type	method	synthetic MTTS cont.	synthetic MTTS Δt	synthetic MTTS event	electrical grid cont.	electrical grid Δt	electrical grid event	MazeBall cont.	MazeBall Δt	MazeBall event
early	cat.	*0.044*	**0.067**	**0.899**	1.623	*0.056*	0.728	2.867	**0.105**	**0.717**
	mean	**0.040**	0.082	*0.887*	1.603	0.057	0.735	3.912	0.109	0.702
	corr.	0.046	0.093	0.887	1.886	0.059	0.710	3.348	0.115	0.696
	gating	0.041	*0.084*	0.887	*1.574*	0.056	*0.738*	3.463	*0.109*	0.703
	share	0.050	0.099	0.878	1.592	0.057	0.737	3.106	0.111	0.705
inter.	cat.	0.049	0.098	0.876	1.573	0.057	0.740	3.747	0.115	0.704
	mean	0.055	0.111	0.855	1.591	0.057	0.731	3.193	0.115	0.692
	corr.	0.045	0.085	0.880	1.738	0.059	0.638	4.296	0.112	0.704
	gating	0.049	0.087	0.872	1.723	0.056	0.710	*2.909*	0.110	0.702
	share	0.050	0.099	0.873	**1.552**	**0.055**	**0.742**	2.924	0.113	0.704
late	mean	0.059	0.171	0.759	1.751	0.062	0.719	**2.626**	0.113	0.705
	corr.	0.058	0.162	0.763	1.823	0.062	0.704	3.486	0.113	*0.706*
uni.	cont.	0.049	0.472	0.680	1.826	0.074	0.579	3.763	0.154	0.664
	event	0.146	0.445	0.622	2.692	0.059	0.737	5.060	0.109	0.708

Mixed-Type Time Series Forecasting

Results

type	method	synthetic MTTS			electrical grid			MazeBall		
		cont.	Δt	event	cont.	Δt	event	cont.	Δt	event
early	cat.	*0.044*	**0.067**	**0.899**	1.623	*0.056*	0.728	2.867	**0.105**	**0.717**
	mean	**0.040**	0.082	*0.887*	1.603	0.057	0.735	3.912	0.109	0.702
	corr.	0.046	0.093	0.887	1.886	0.059	0.710	3.348	0.115	0.696
	gating	0.041	*0.084*	0.887	*1.574*	0.056	*0.738*	3.463	*0.109*	0.703
	share	0.050	0.099	0.878	1.592	0.057	0.737	3.106	0.111	0.705
inter.	cat.	0.049	0.098	0.876	1.573	0.057	0.740	3.747	0.115	0.704
	mean	0.055	0.111	0.855	1.591	0.057	0.731	3.193	0.115	0.692
	corr.	0.045	0.085	0.880	1.738	0.059	0.638	4.296	0.112	0.704
	gating	0.049	0.087	0.872	1.723	0.056	0.710	*2.909*	0.110	0.702
	share	0.050	0.099	0.873	**1.552**	**0.055**	**0.742**	2.924	0.113	0.704
late	mean	0.059	0.171	0.759	1.751	0.062	0.719	**2.626**	0.113	0.705
	corr.	0.058	0.162	0.763	1.823	0.062	0.704	3.486	0.113	*0.706*
uni.	cont.	0.049	0.472	0.680	1.826	0.074	0.579	3.763	0.154	0.664
	event	0.146	0.445	0.622	2.692	0.059	0.737	5.060	0.109	0.708

Conclusion

- The **strength** and **direction** of intermodal interactions affect the optimal fusion strategy

- **Early Fusion**
  - Good at capturing low-level bidirectional interactions
  - Best performance when the interaction strength is medium (both modalities carry relevant information)

- **Concatenation** often beats more sophisticated fusion methods

The best forecasting approach depends on the nature of the time series

Thank you for your attention

Marginalization