

# Leveraging Modality Tags for Enhanced Cross-Modal Video Retrieval

Adriano Fragomeni

Dima Damen

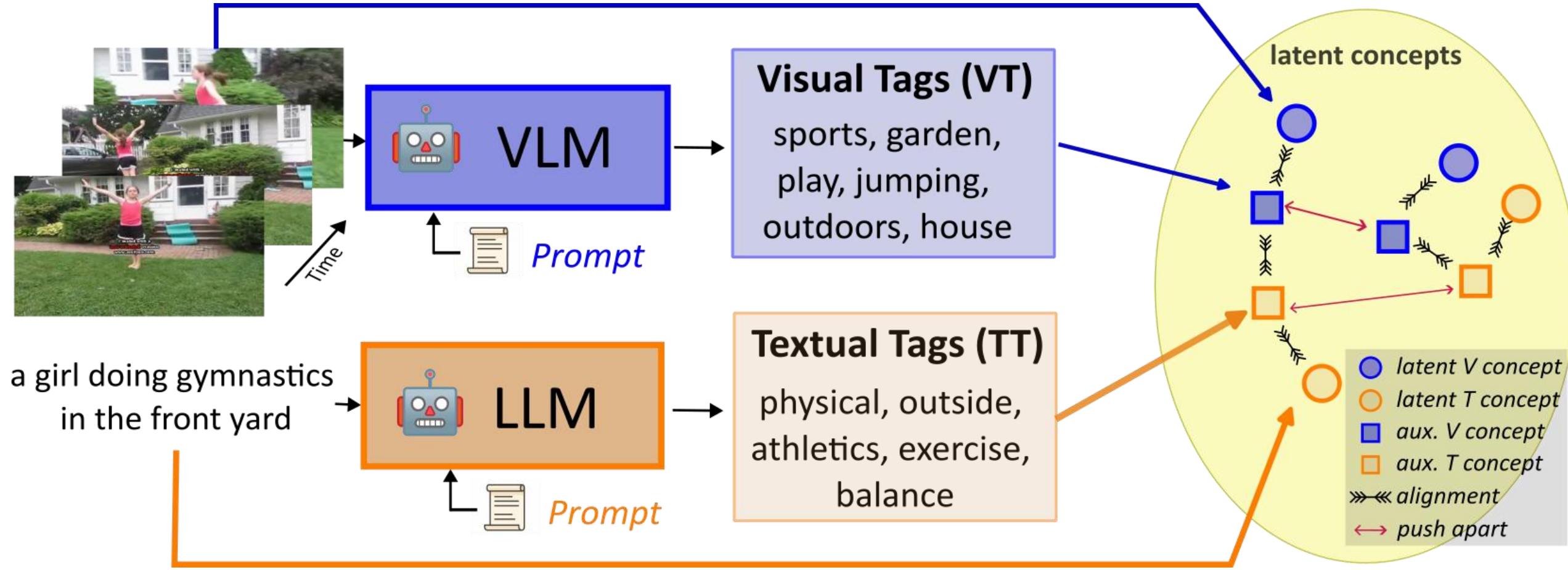
Michael Wray

University of Bristol



## Overview

Leveraging auxiliary modality-specific tags extracted from foundation models to learn auxiliary latent concepts and improving cross-modal alignment.

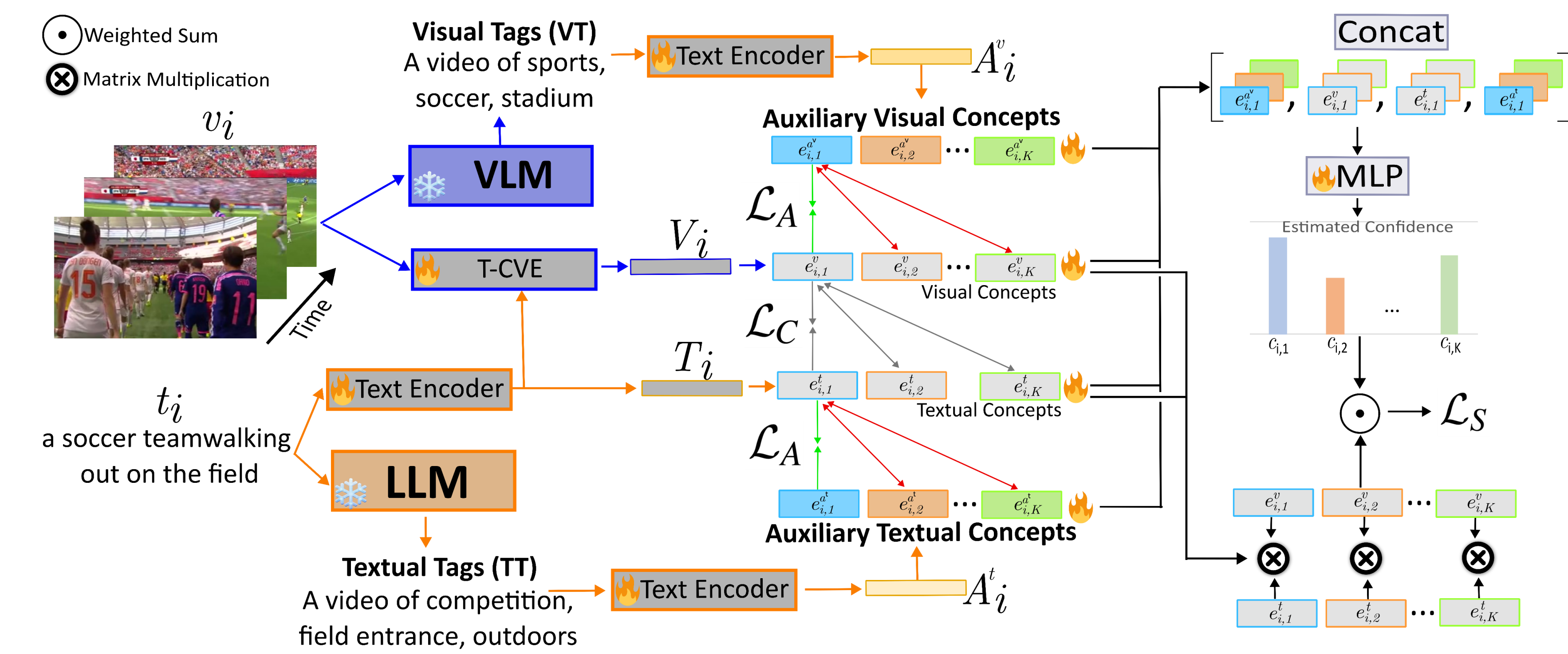


## Tag Examples

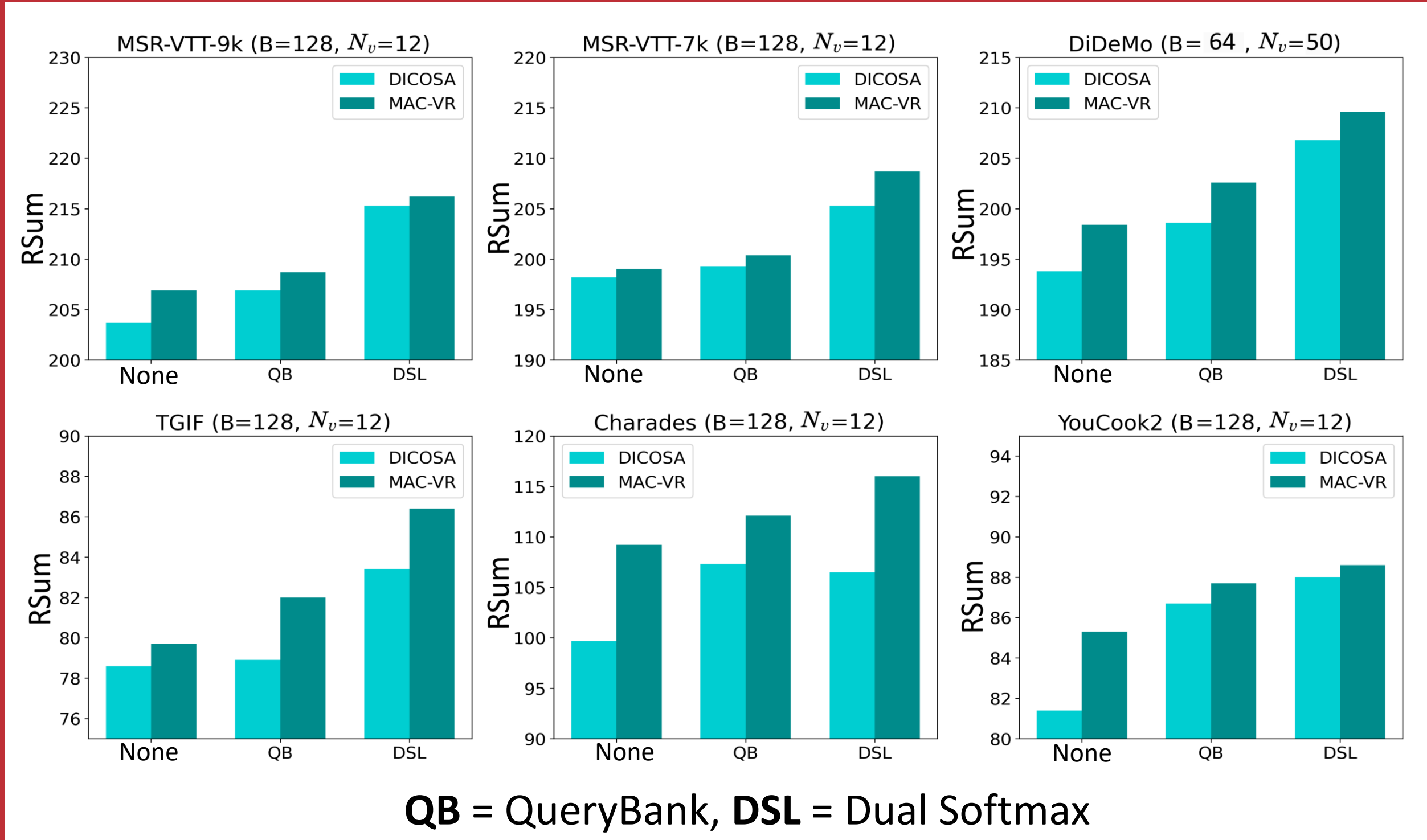
| Video | Visual Tags (VT)                                                                                                | Textual Tags (TT)                                                                                              | Caption                                                      |
|-------|-----------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
|       | design, racing, road, <b>driving</b> , movement, control, speed, engine <b>car</b> , transportation...          | product showcase, style, brand differentiation, advertising technique, <b>automotive marketing</b> ...         | a commercial for the mazda 3 the car sliding around a corner |
|       | <b>birthday celebration</b> , family gathering, cake, candle, child, <b>birthday</b> , making wishes, family... | <b>event</b> , reaction, harmony, applause, social, emotion, entertainment, collective, <b>celebration</b> ... | the people begin to clap.                                    |
|       | <b>crispy exterior</b> , hot, meat, <b>crispy</b> , chicken, frying chicken, deep fry meat-based dish...        | <b>crisping</b> , texture change, browning, golden, high heat, <b>crust</b> , alter food state...              | place chicken in hot oil and fry until golden brown          |

Blue and orange highlight semantic overlap that might be used for cross-modal alignment.

## Architecture



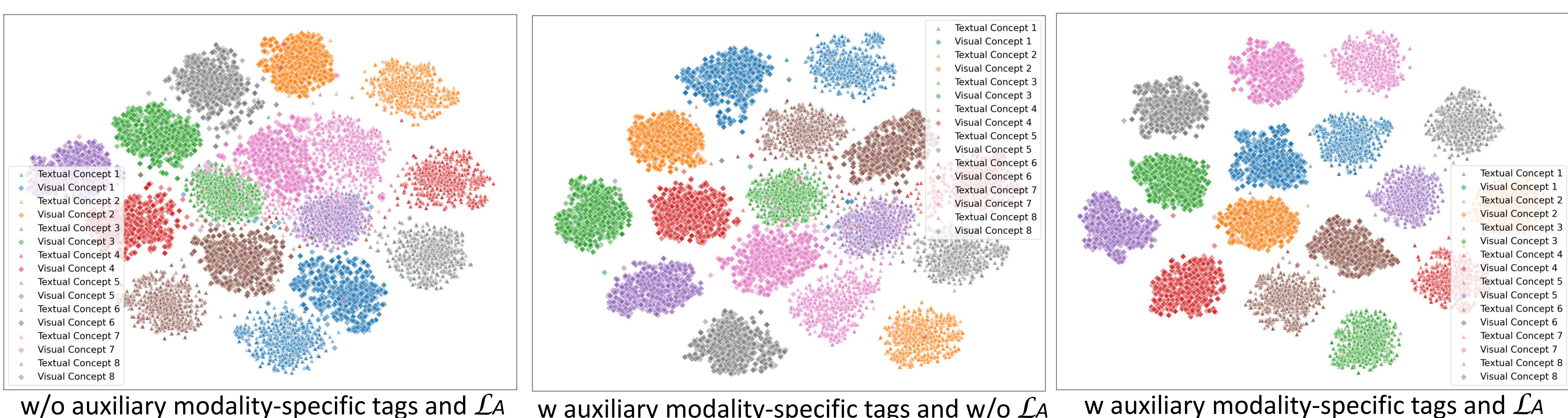
## Results



## Qualitative Results

| Query: a class is being introduced to a digital reading device                                                      | Video | Rank | + VT                                                                                                                                            | + TT                                                                                                             | Rank |
|---------------------------------------------------------------------------------------------------------------------|-------|------|-------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|------|
|                                                                                                                     |       | 12   | young, student, technology, many, help, book, learning                                                                                          | initial setup, engagement, feedback, reading materials, assistant, introduction, training, interactive features  | 2    |
| Query: a person is cooking on stage                                                                                 | Video | Rank | + VT                                                                                                                                            | + TT                                                                                                             | Rank |
|                                                                                                                     |       | 1    | singing, audience, dressing, television, performance, talent, humor, cooking                                                                    | exhibition cooking, creativity, theater food, skill, showcasing talents, artistry, entertainment, display skills | 4    |
| Query: a woman takes off her jacket ... all the people sitting. group of people playing soccer woman picks up shirt | Video | Rank | + VT                                                                                                                                            | + TT                                                                                                             | Rank |
|                                                                                                                     |       | 4    | playing soccer, gathering, socializing friends, team sports, friendship camaraderie, picknicking, drinking, lazing                              | sports, interaction, connection, gathering, friendship, community, friends, social                               | 1    |
| Query: woman with purple suit is dancing on scenery                                                                 | Video | Rank | + VT                                                                                                                                            | + TT                                                                                                             | Rank |
|                                                                                                                     |       | 7    | skill, dance, beauty, talent, skillful, athleticism, artistic expression, performance                                                           | artistic, artistry, feminine, elegance, dancing, music, dance, spectacle                                         | 1    |
| Query: A person is cooking some food. Another person is sitting in a chair and tying a pile of mail.                | Video | Rank | + VT                                                                                                                                            | + TT                                                                                                             | Rank |
|                                                                                                                     |       | 31   | cooking food, shared space, performing task together, home, using kitchen utensils, woman, food, sharing chores                                 | domestic life, household, food, domesticity, chores, organizing mail, cooking, meal, maintenance                 | 2    |
| Query: shred some cabbage and prepare some cauliflowers                                                             | Video | Rank | + VT                                                                                                                                            | + TT                                                                                                             | Rank |
|                                                                                                                     |       | 7    | piece food, vegetable processing, hand-held knife, cutting board, ingredient preparations, knife skills, cauliflower recipe, cooking techniques | ingredients, shredding, vegetarian, cutting, vegetable, healthy, food prep., cooking                             | 1    |

## Concept Alignment



## Conclusion

- Auxiliary tags provide semantic cues that help bridge the gap between video and text modalities.
- The novel Alignment loss  $\mathcal{L}_A$  ensures consistent mapping between modality representations.
- Reliance on foundation models can lead to *hallucinated* tags.