LEARNING FROM UNLABELED MULTIMODAL DATA

Representation Learning from Multimodal Sensor Data

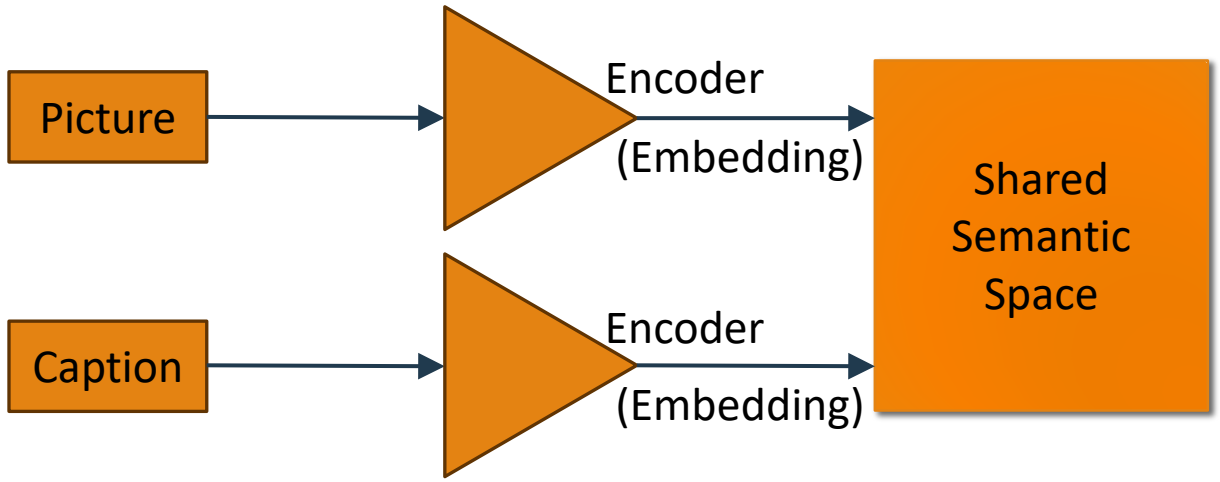
Reminders

- Look for HW1 on the class webpage (will be out by 11:59pm).
- Student-led talks:
  - Your presentation topics have been assigned. Each project group is assigned one topic (50 min talk + 10 min for Q&A). Please go to the “schedule” page to find out your topic.
  - Each group is free to survey the topic on their own and choose approximately 6 papers to present. (AI help is allowed.)
  - I will contact your group roughly two weeks before your presentation to recommend papers and finalize your paper selection.
  - Selected papers are shared on the website roughly one week before the talk.

Multimodal Data

The general idea is to convert multimodal data into a single latent representation.

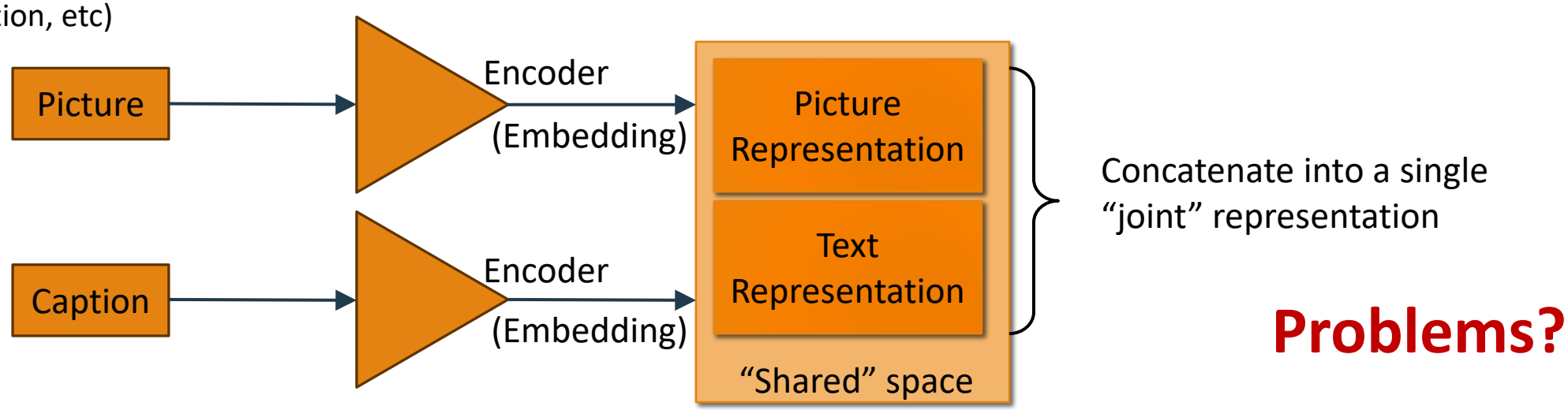
- Early solutions were inherited from NLP, where sentences in different languages (e.g., English and French) that conveyed similar meanings were mapped to the same semantic representation in the latent space, allowing for translation.
- The approach was generalized to translation among common co-existing modalities such as vision (pictures) and text (captions), where visual concepts and text descriptions were mapped to the same latent space.
- Current examples of multimodal representation learning come from video/audio, VR (video+sensor), robotics (vision+tactile), autonomous driving (vision+radar), and IoT (various modalities including acoustic, seismic, motion, etc)



Multimodal Data

The general idea is to convert multimodal data into a single latent representation.

- Early solutions were inherited from NLP, where sentences in different languages (e.g., English and French) that conveyed similar meanings were mapped to the same semantic representation in the latent space, allowing for translation.
- The approach was generalized to translation among common co-existing modalities such as vision (pictures) and text (captions), where visual concepts and text descriptions were mapped to the same latent space.
- Current examples of multimodal representation learning come from video/audio, VR (video+sensor), robotics (vision+tactile), autonomous driving (vision+radar), and IoT (various modalities including acoustic, seismic, motion, etc)



Problems?

Why Not Just Concatenate Multimodal Embeddings?

Alignment prevents modality shortcutting

- Without alignment, a downstream model may over-rely on the dominant modality.

Alignment allows for cross-modal operations:

- Zero-shot transfer
- Missing modality robustness
- Cross-modal retrieval
- Modality translation

Alignment allows similarity-based reasoning across modalities

- Can’t express distance metrics across modalities if they are not aligned.

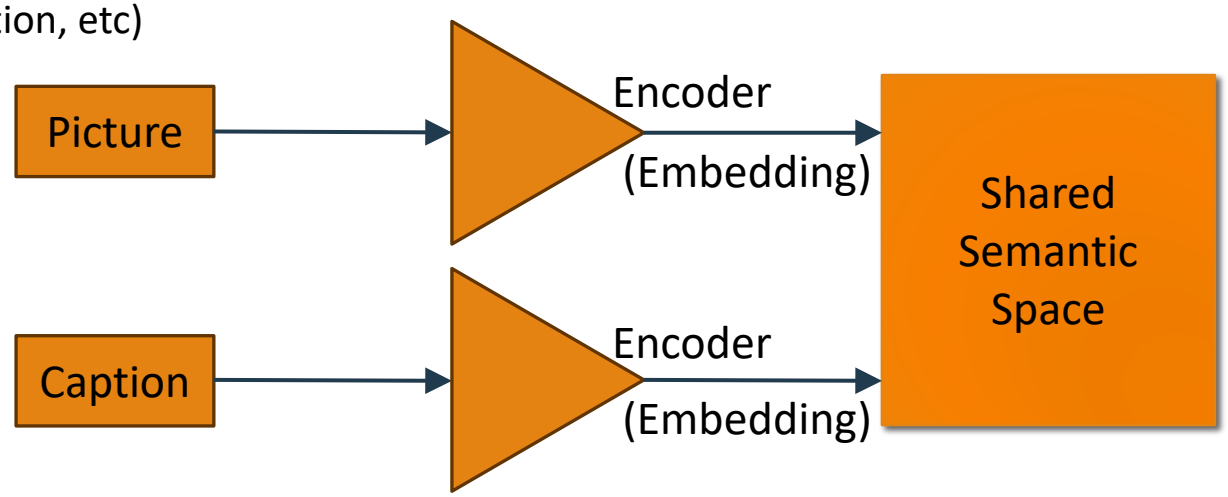
Alignment prevents redundant representation

- Without alignment, similar information may be encoded twice. Downstream models will be less efficient.

Multimodal Data

The general idea is to convert multimodal data into a single latent representation.

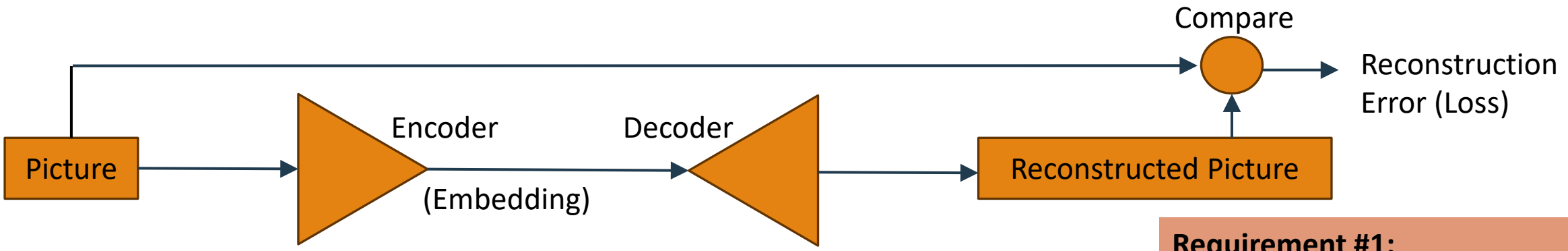
- Early solutions were inherited from NLP, where sentences in different languages (e.g., English and French) that conveyed similar meanings were mapped to the same semantic representation in the latent space, allowing for translation.
- The approach was generalized to translation among common co-existing modalities such as vision (pictures) and text (captions), where visual concepts and text descriptions were mapped to the same latent space.
- Current examples of multimodal representation learning come from video/audio, VR (video+sensor), robotics (vision+tactile), autonomous driving (vision+radar), and IoT (various modalities including acoustic, seismic, motion, etc)



Requirement #1:
Each encoder must faithfully represent the content of the respective modality

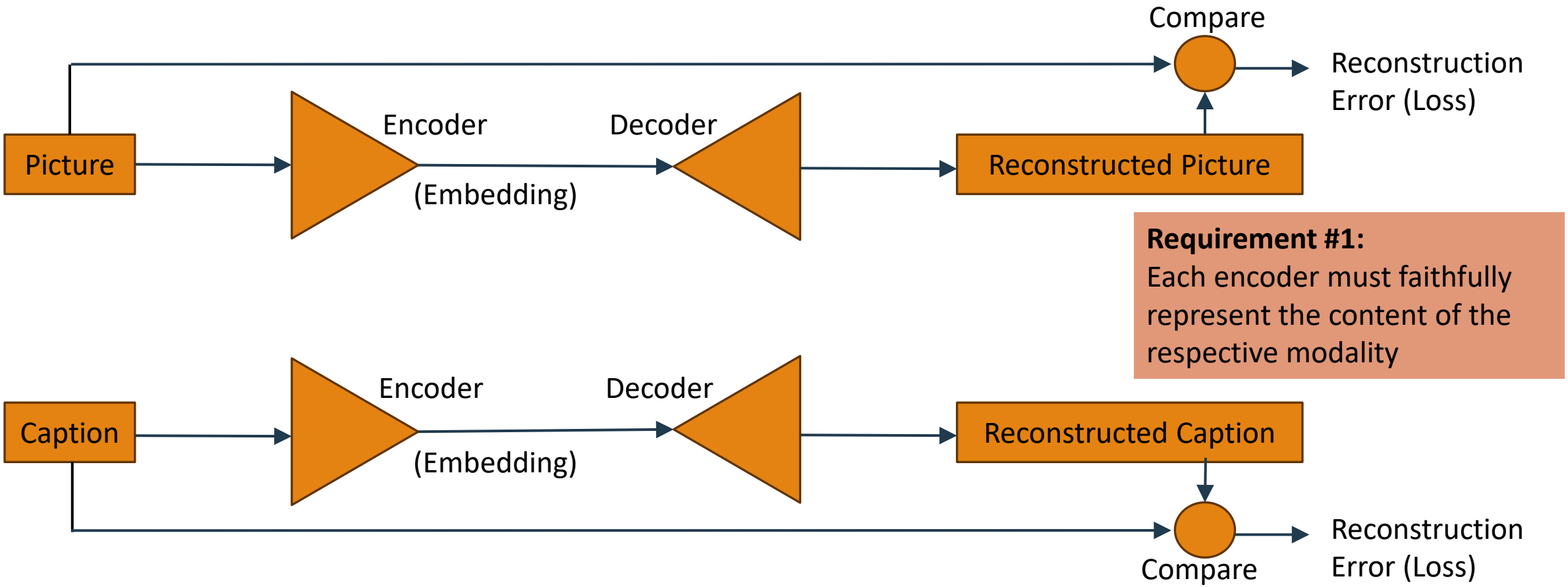
Requirement #2:
The latent representations of the different modalities must be “aligned”

Multimodal Alignment

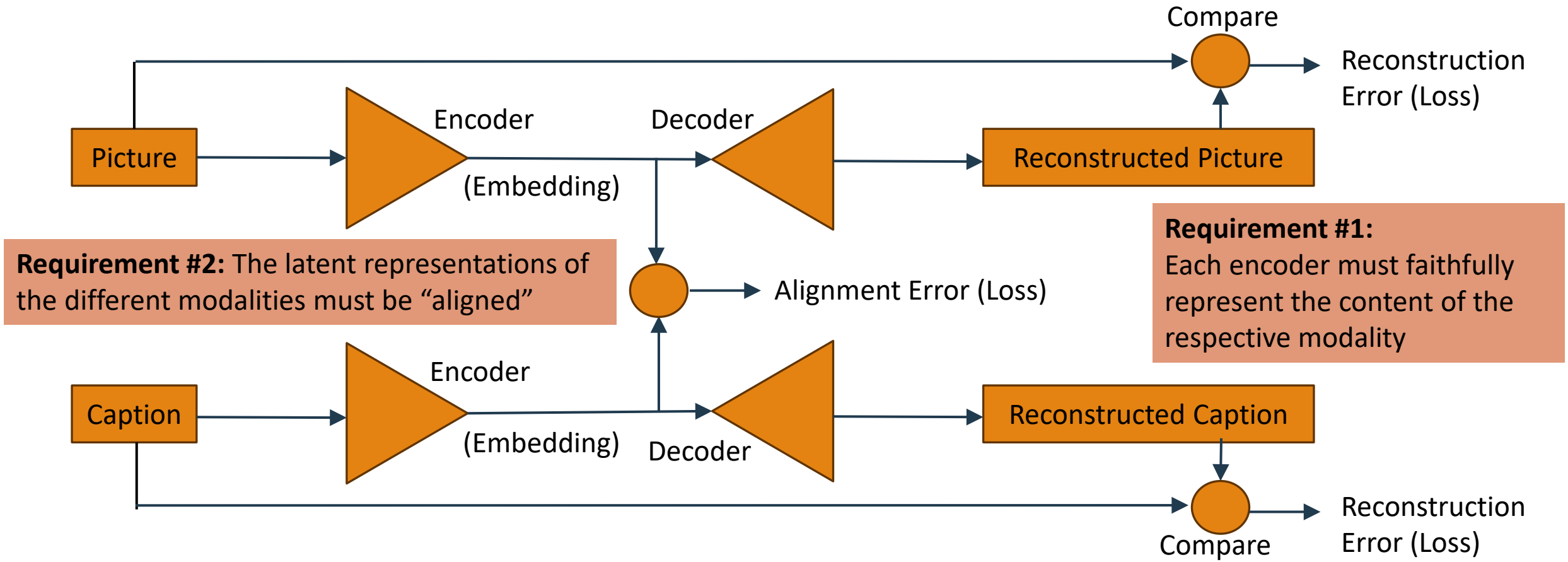


Requirement #1:
Each encoder must faithfully represent the content of the respective modality

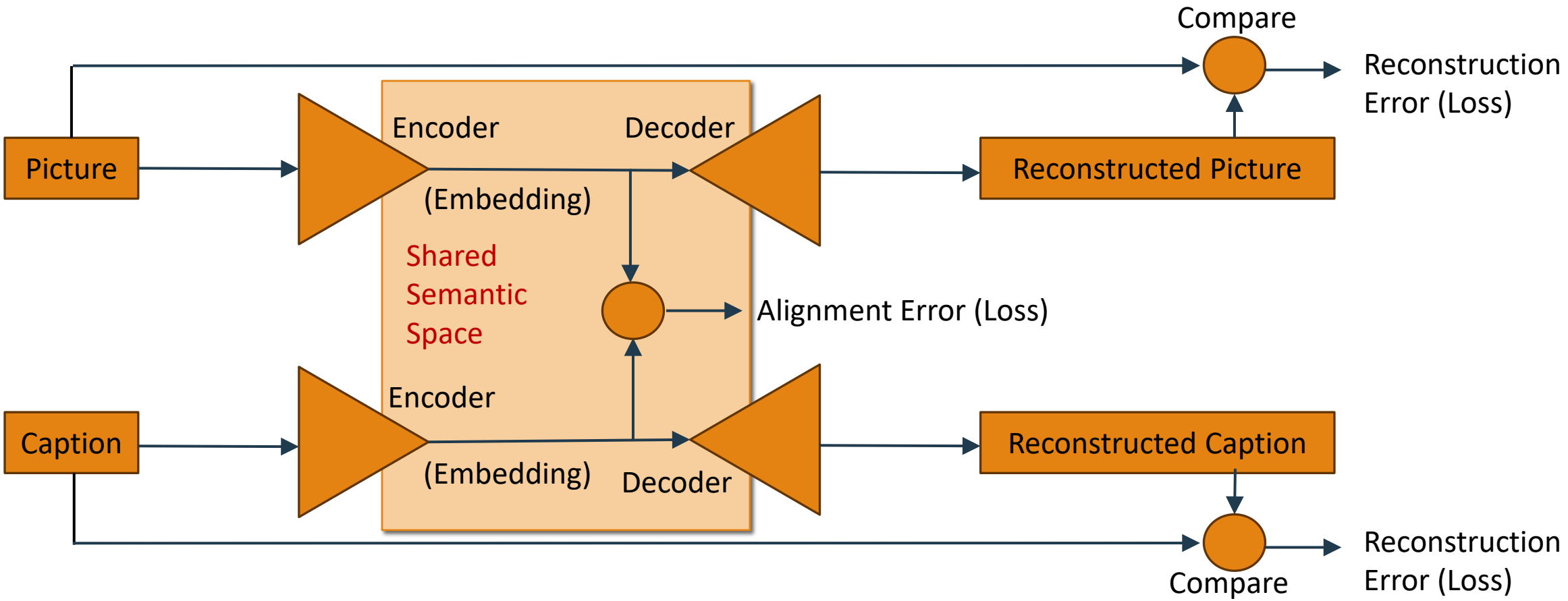
Multimodal Alignment



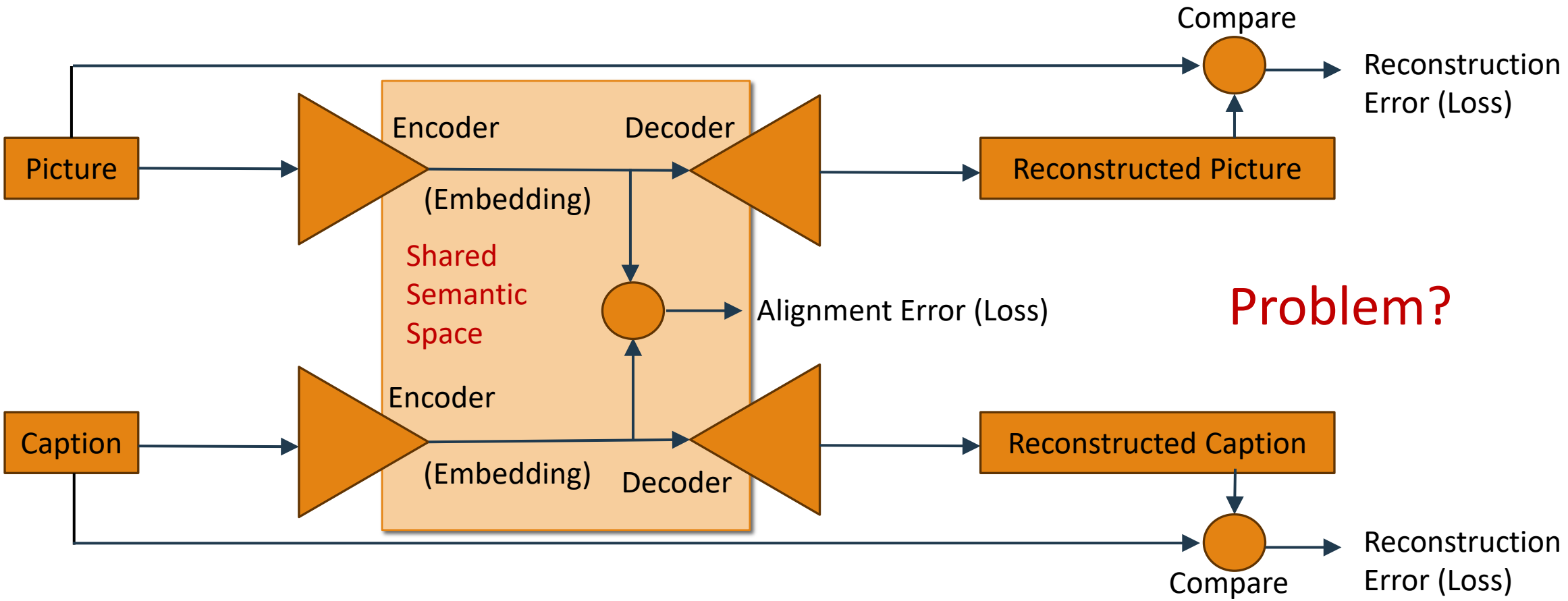
Multimodal Alignment



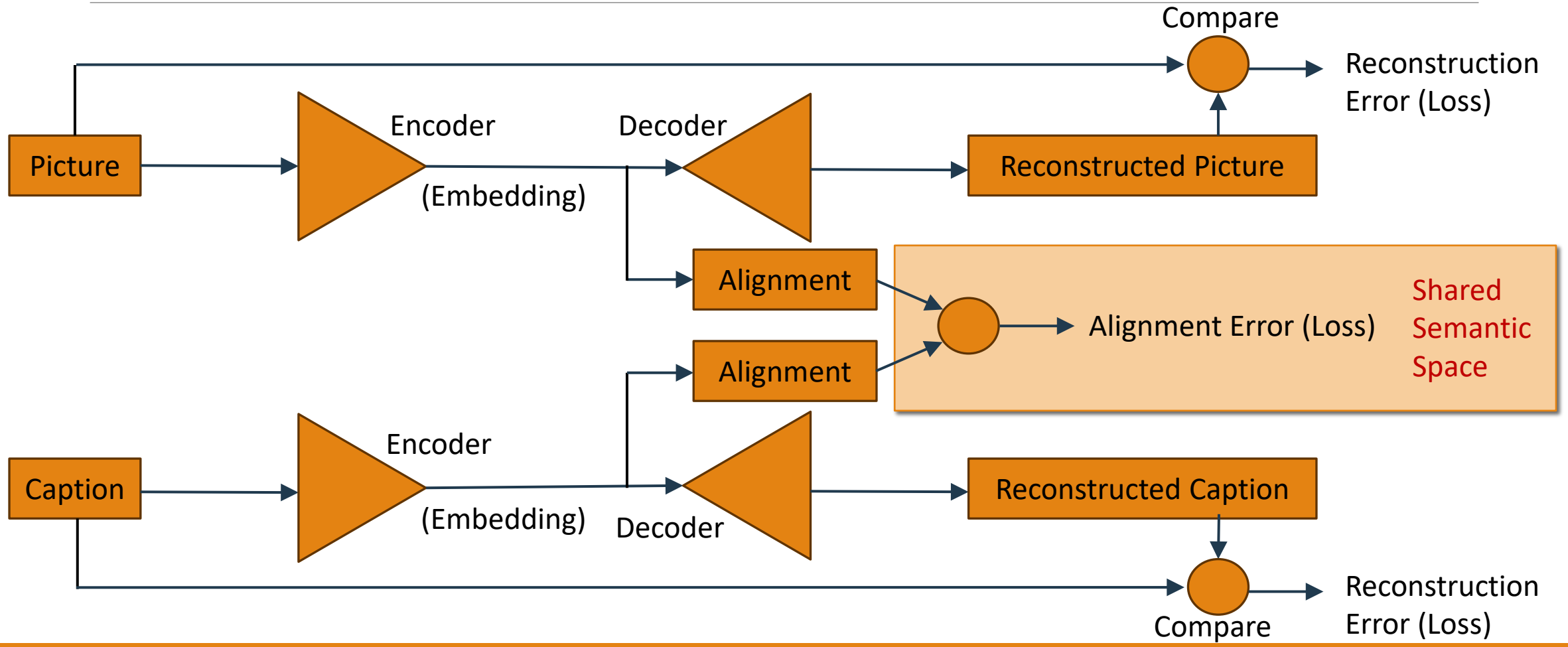
Multimodal Alignment



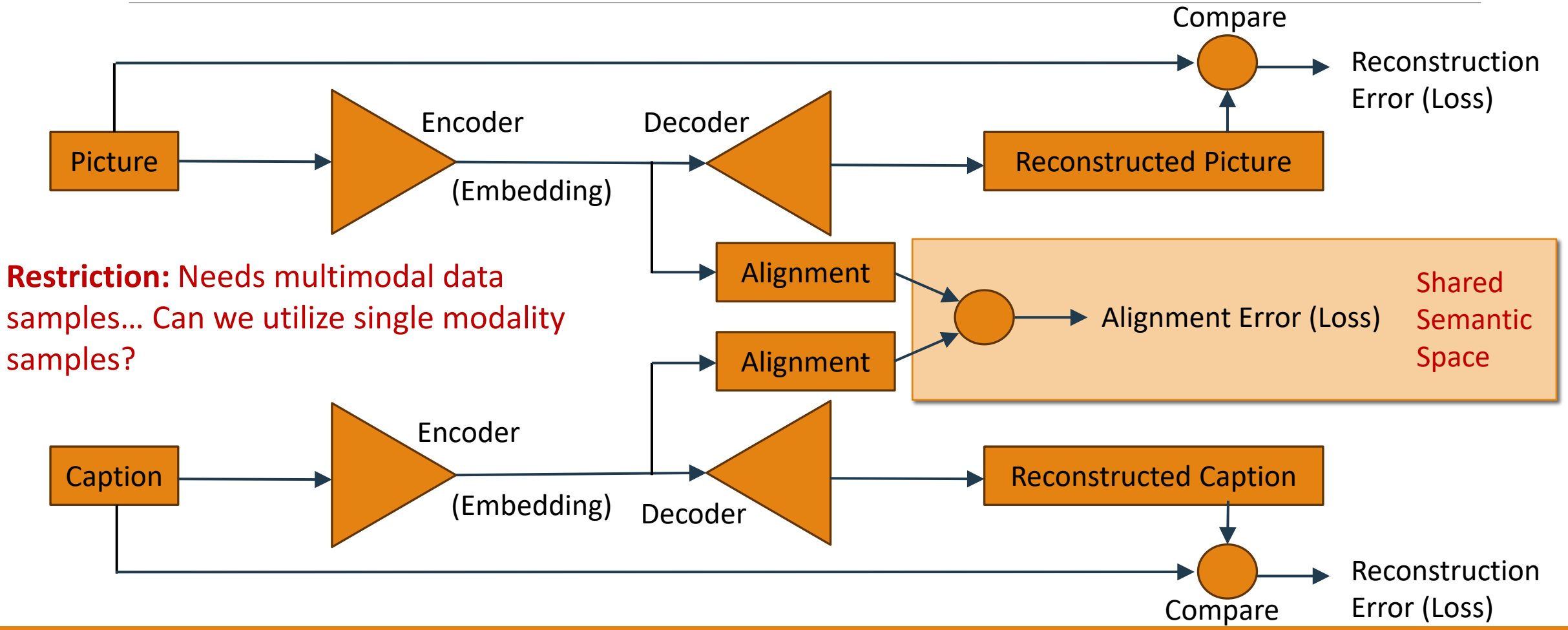
Multimodal Alignment



Multimodal Alignment

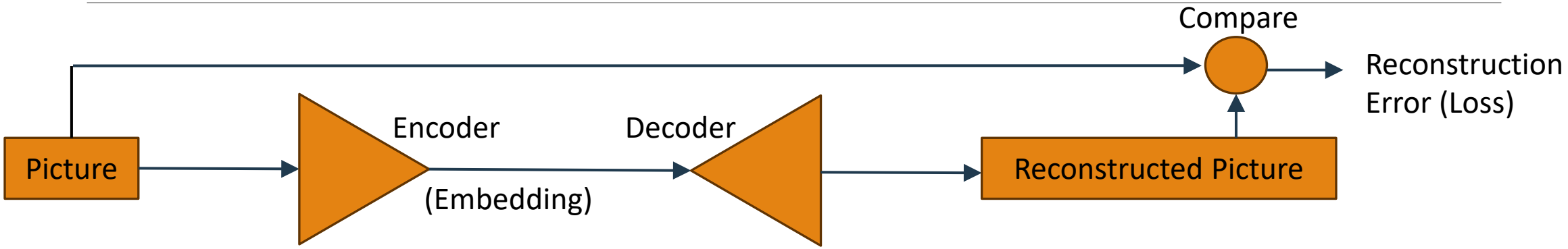


Multimodal Alignment

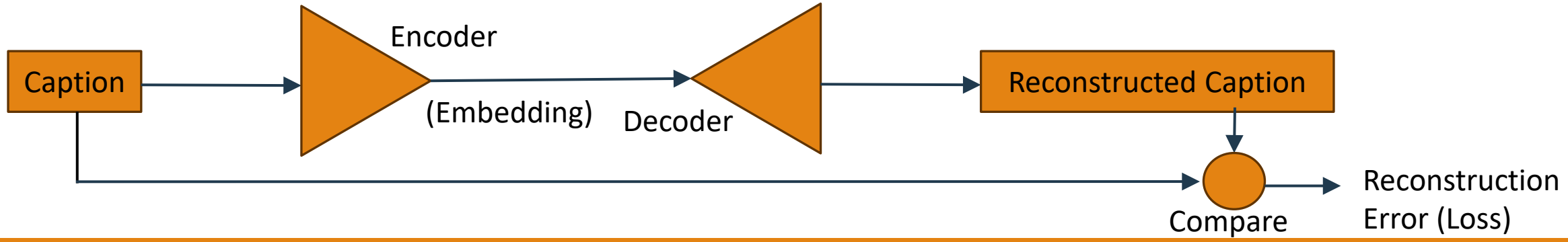


Restriction: Needs multimodal data samples… Can we utilize single modality samples?

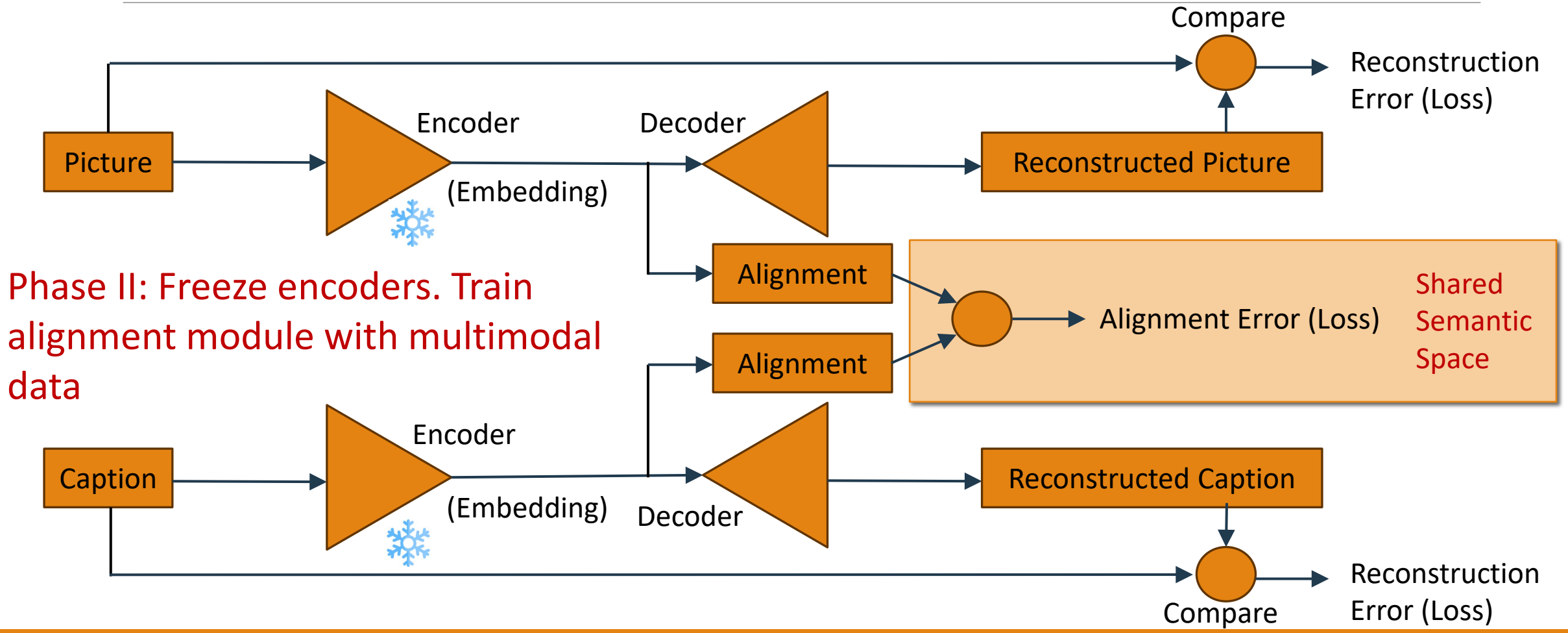
Multimodal Alignment (Two Phase)



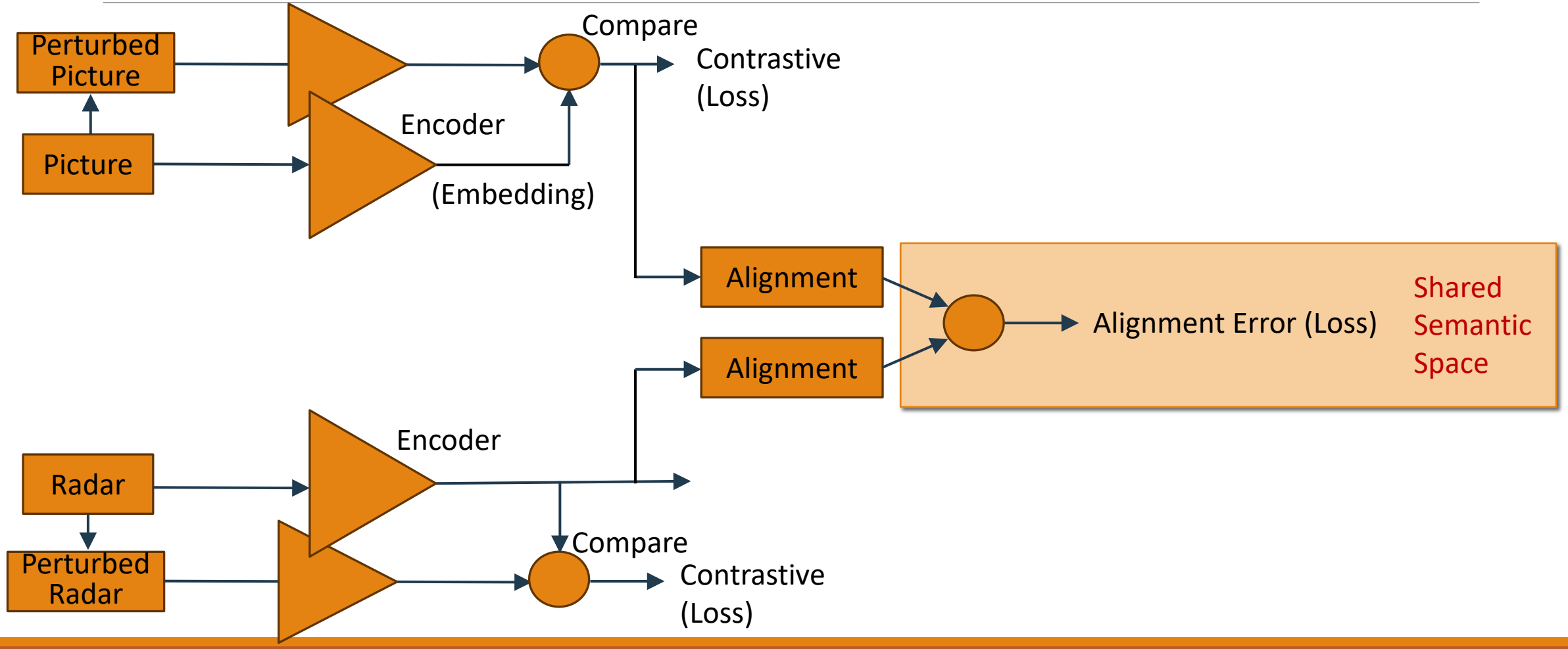
Phase I: Train single modality encoders using single-modality data



Multimodal Alignment (Two Phase)



Multimodal Alignment Variants: Contrastive Learning



Question

What learning objective is better at learning to represent individual modalities in multimodal representation learning, contrastive learning or reconstruction?

Question

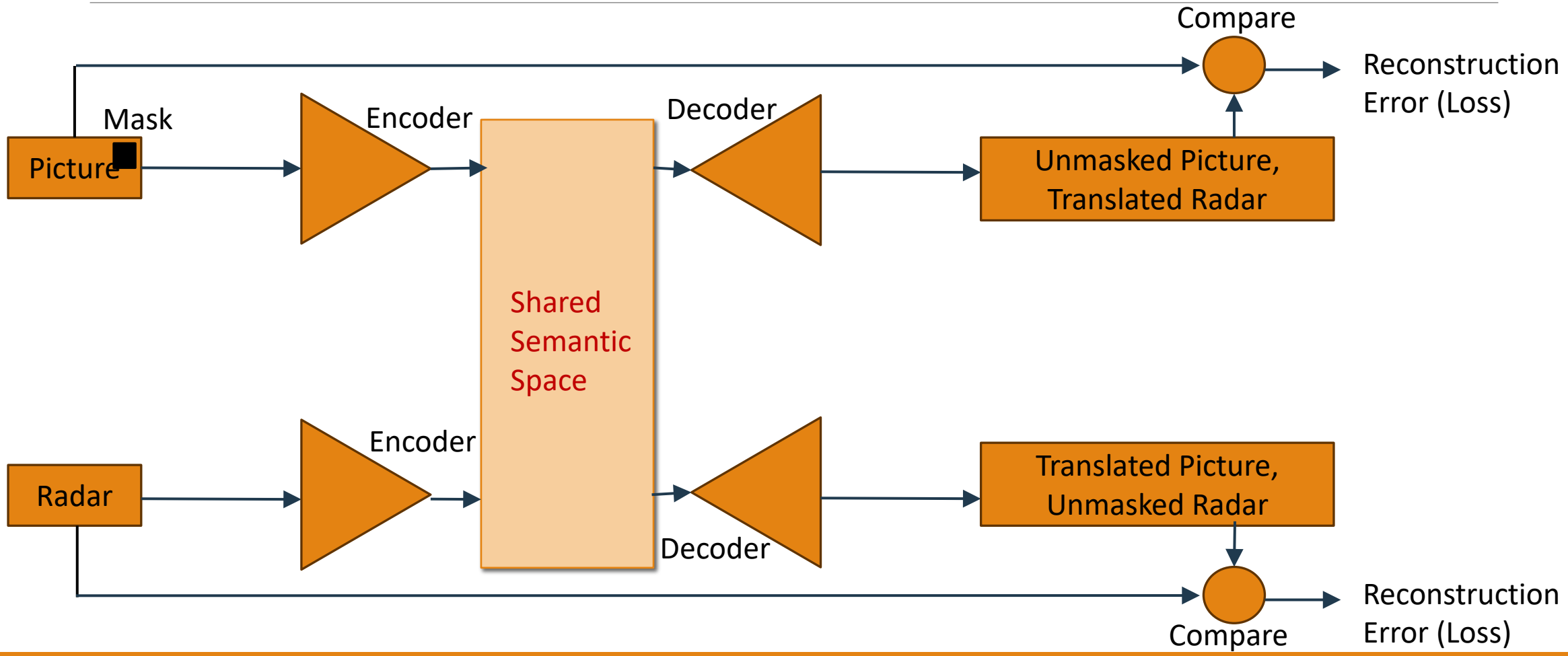
What learning objective is better at learning to represent individual modalities in multimodal representation learning, contrastive learning or reconstruction?

Contrastive learning encourages:

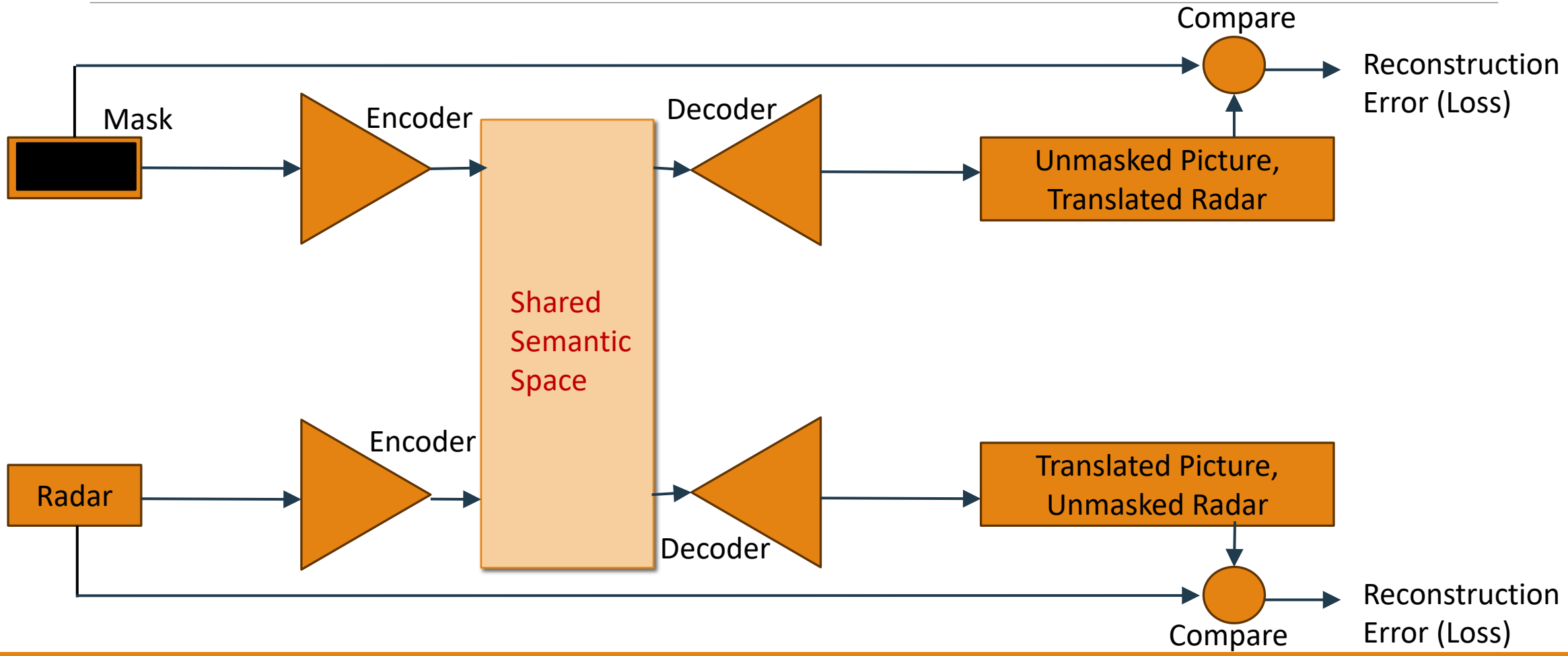
- Invariance to augmentations.
- Generalization/classification (discriminative structure).
- Clustering of semantically similar samples.
- Use of inductive bias: “Keep what distinguishes examples; discard nuisance variation.”

Reconstruction objectives: can be better at producing faithful generative tasks

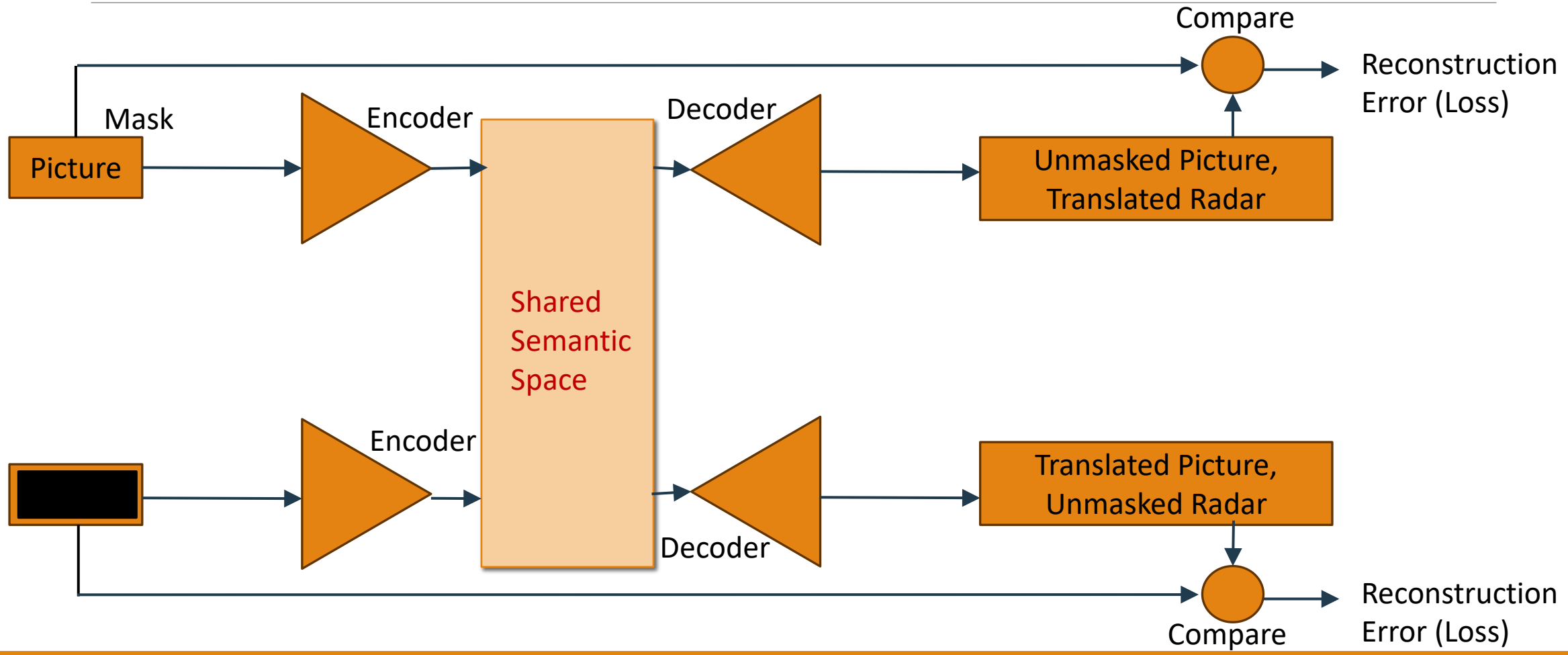
Multimodal Alignment Variants: Cross-Modal Masked Auto-Encoding



Multimodal Alignment Variants: Cross-Modal Masked Auto-Encoding



Multimodal Alignment Variants: Cross-Modal Masked Auto-Encoding



Dave Vedant, Fotios Lygerakis, and Elmar Rueckert. "Multimodal visual-tactile representation learning through self-supervised contrastive pre-training." In 2024 IEEE ICRA, pp. 8013-8020. IEEE, 2024.

Example: Visual + Tactile Representation Learning (for Robotics)

Motivation:

- Visual representations are not always enough to guide robotic manipulation of various objects
- Tactile information, such as texture, can help improve object manipulation (e.g., slippery objects should be handled differently from rough objects, etc)

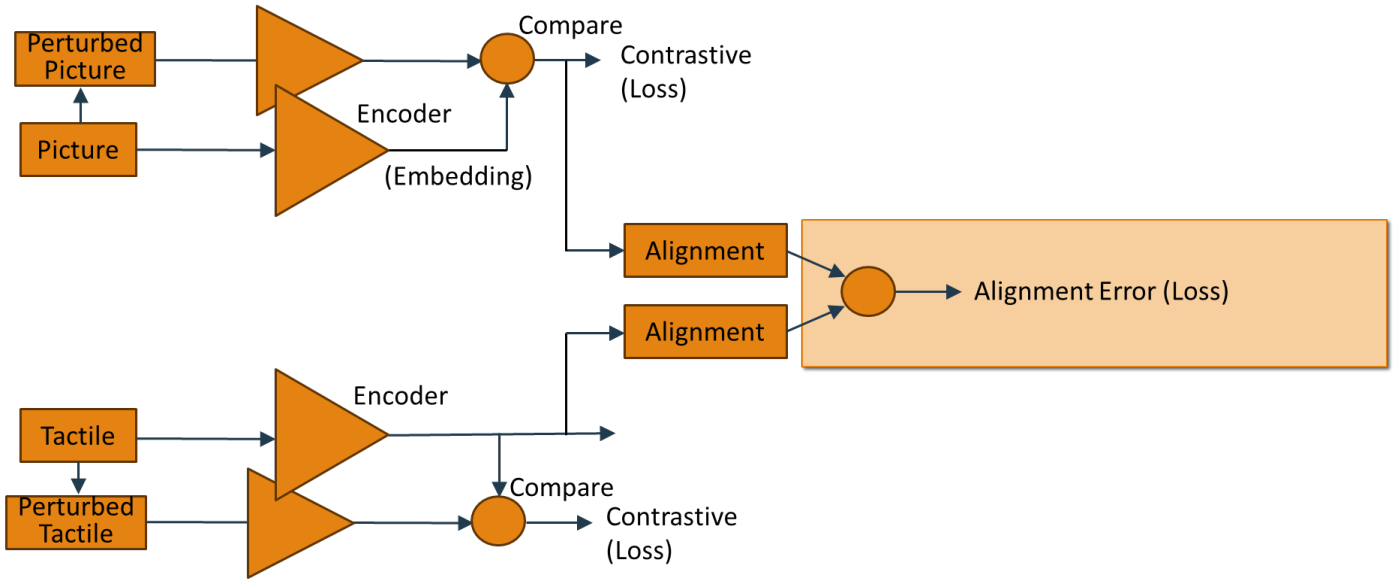
Example: Visual + Tactile Representation Learning (for Robotics)

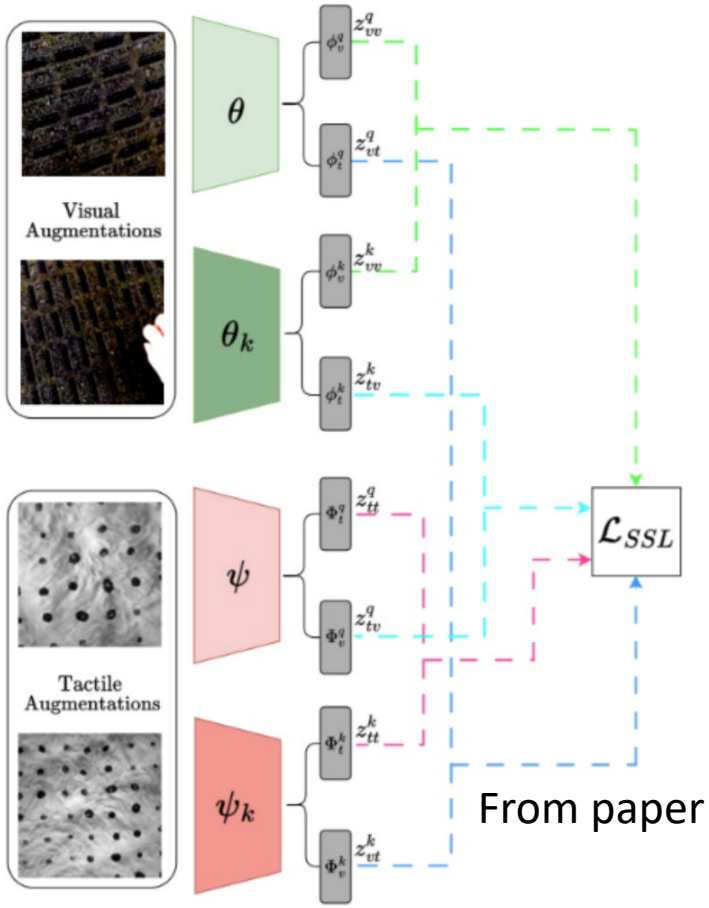
Contrastive loss (vision)

Contrastive loss (tactile)

Alignment loss

$$\mathcal{L}_{mm} = \mathcal{L}_{vv} + \mathcal{L}_{tt} + \lambda_{inter}(\mathcal{L}_{vt} + \mathcal{L}_{tv})$$





From paper

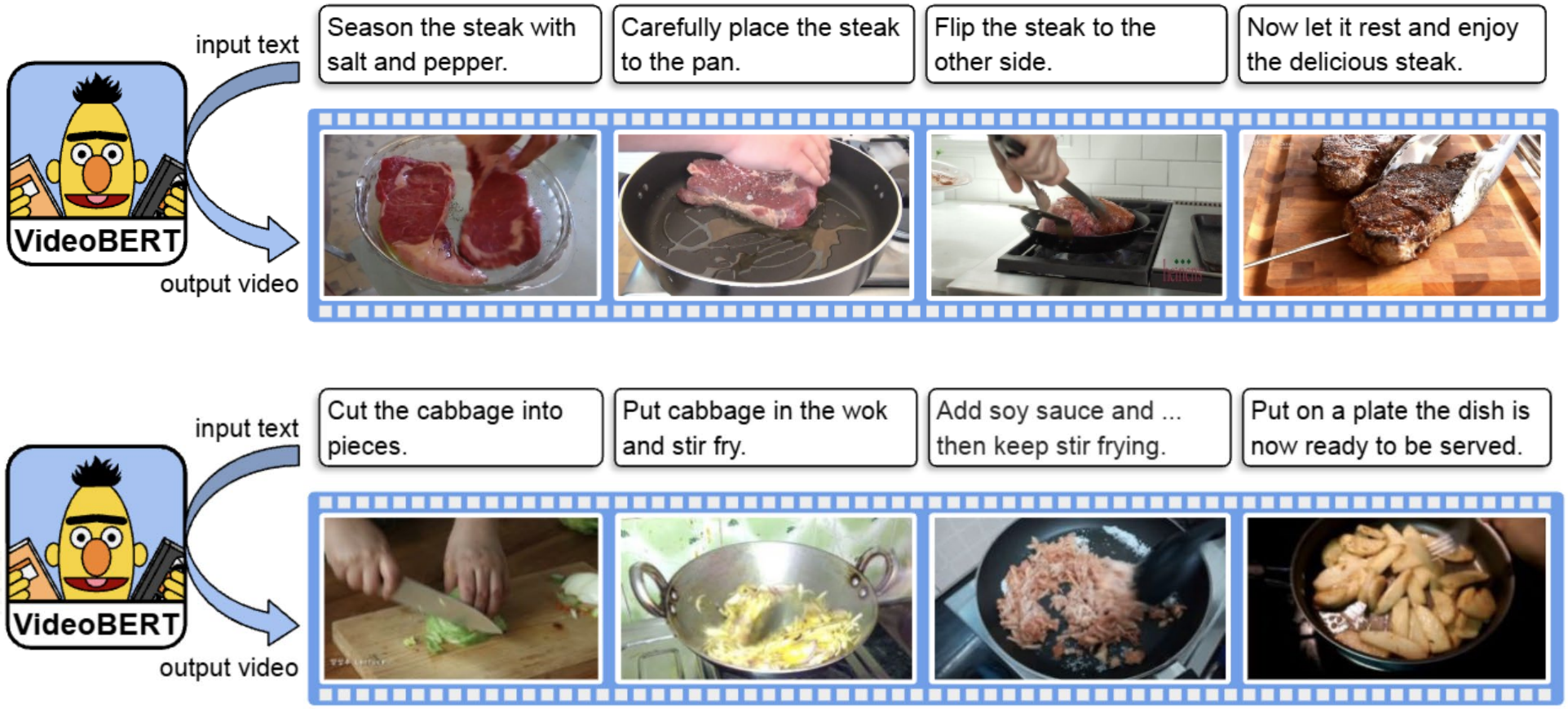
Example: Visual + Tactile Representation Learning (for Robotics)

Evaluation Results: Improved classification accuracy when using aligned vision and tactile information

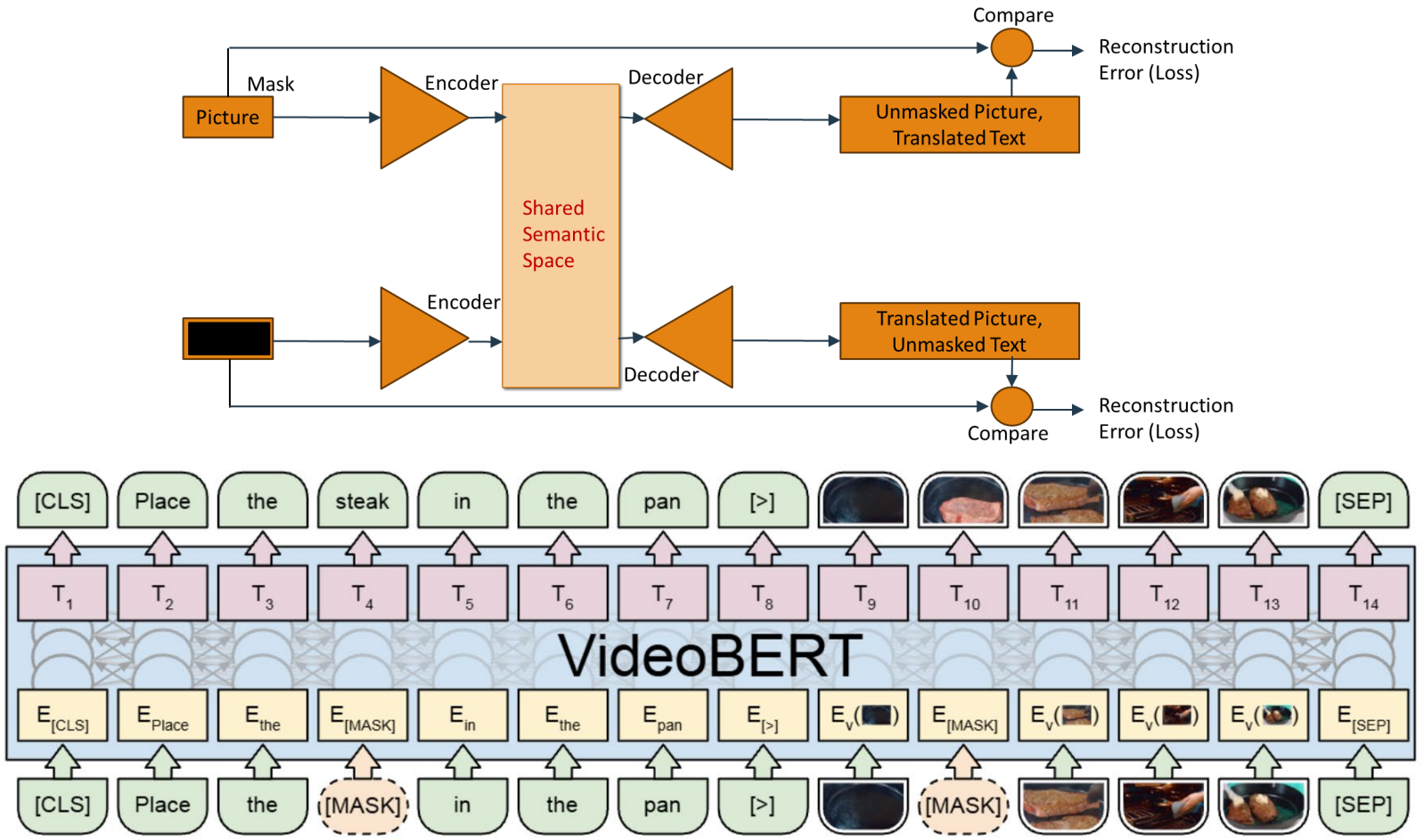
Dataset	Method	Modality	Category Accuracy%	Hard/Soft Accuracy (%)	Rough/Smooth Accuracy (%)
Chance	-	Tactile	18.6	66.1	56.3
ResNet18 [43]	Supervised Learning	Tactile	57.4	**89.1**	79.3
		Tactile + Visual	48.0	85.9	80.0
TAG [11]	Contrastive Multiview Coding	Tactile	54.7	77.3	79.4
		Tactile + Visual	68.6	87.1	82.4
SSVTP [10]	InfoNCE	Tactile	46.1	79.7	75.8
		Tactile + Visual	70.7	88.6	83.6
MViTac (Ours)	Multimodal Contrastive Training	Tactile	**57.6**	86.2	**82.1**
		Tactile + Visual	**74.9**	**91.8**	**84.1**

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. "Videobert: A joint model for video and language representation learning." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 7464-7473. 2019.

Example: VideoBERT



VideoBERT



Trained with YouTube videos with cooking instructions (audio is converted to text with speech recognition software).

Training uses a multimodal masked auto-encoder (based on the BERT code-base) that masks parts of text and video then attempts to reconstruct them

VideoBERT Evaluation

"Now, let me show you how to [MASK] the [MASK]"

VideoBERT is asked to fill in the two blanks in the above sentence (typically a verb and a noun, respectively), given a visual depiction of the activity.

Top verbs: make, assemble, prepare
Top nouns: pizza, sauce, pasta

Top verbs: make, do, pour
Top nouns: cocktail, drink, glass

Top verbs: make, prepare, bake
Top nouns: cake, crust, dough

VideoBERT Evaluation


GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

S3D: cut the tomatoes into thin slices


GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan


GT: cut yu choy into diagonally medium pieces

VideoBERT: chop the cabbage

S3D: cut the roll into thin slices


GT: remove the calamari and set it on paper towel

VideoBERT: fry the squid in the pan

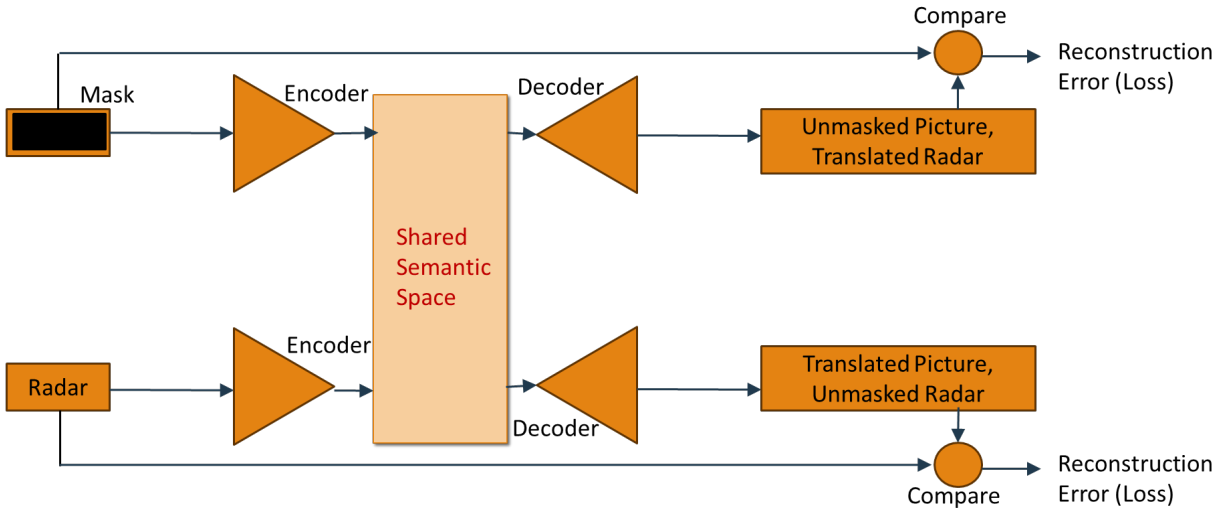
S3D: add the noodles to the pot

Differences when Applying Multimodal Learning to Sensor Data?

Differences when Applying Multimodal Learning to Sensor Data?

- **A difference in goal:**
  - We do not want to “hallucinate” features (as in generative AI that produces video from text)
  - We also do not want to “abstract” information (as in AI that generates image captions from images)
  - Rather, we want to efficiently represent all useful multimodal data to enable downstream analytics.

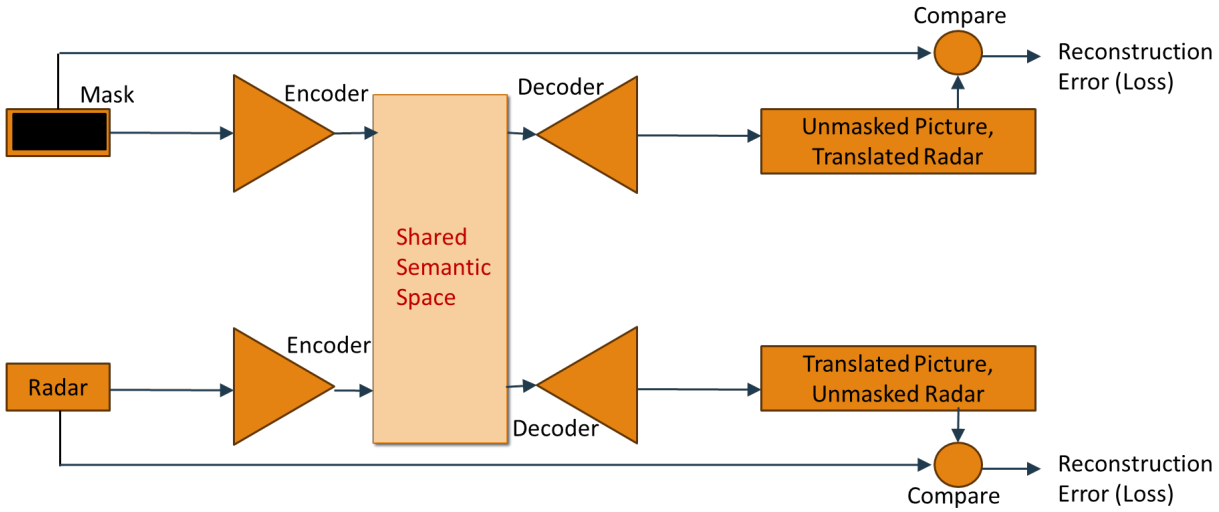
What’s the problem with learning a shared semantic space as depicted here?



Differences when Applying Multimodal Learning to Sensor Data

- **A difference in goal:**
  - We do not want to “hallucinate” features (as in generative AI that produces video from text)
  - We also do not want to “abstract” information (as in AI that generates image captions from images)
  - Rather, we want to efficiently represent all useful multimodal data to enable downstream analytics.

What’s the problem with learning a shared semantic space as depicted here?



Challenge: How to align multimodal sensor data while being able to represent both shared (cross-modal) information and unique information to each sensor?

Differences when Applying Multimodal Learning to Sensor Data

- **A difference in approach:**
  - In contrastive learning, to perform cross-modal alignment, what are notions of similarity across different modalities?
  - In masked auto-encoding, how to properly weigh cross-modal reconstruction errors?
  - How to ensure that data from different modalities are not encoded redundantly?
  - How to handle cases where not all modalities are always present?

A Note on HW1 (Will be Out Tonight)

Homework format: **Debate.**

- A proposition will be shared (e.g., “AI will kill us all”).
- Each group will adopt a view (pro/anti) regarding this proposition
- The homework for the group is to support their view with arguments
- The arguments will be debated in class
- A poll on the proposition (pro/anti) will be done before and after the debate
- The goal: recruit more people to your view (in post-debate poll, compared to pre-debate poll)
- Modification: To reduce bias (since each group will need to defend their opinion), we shall do two debate questions each time. Only half the groups are involved in preparing arguments for a given debate question. The other half are “audience” for that question. Thus, for each question, the debating groups will try to “recruit” members of the audience to their view.