Appendices

A ADDITIONAL DETAILS ON THE EXPERIMENTS

We run all inferences of LLAVA and encoder models on an NVIDIA RTX A5000 GPU, and solving Equation 2 with $35,000$ multimodal feature vectors on a 64 core Xeon Gold 6226R CPU machine takes less than 10 minutes. The implementation of CCA is from CCA Zoo Chapman & Wang (2021).

Multimodal Encoders. The multimodal image-text encoder used throughout the experiments is *laion/CLIP-ViT-bigG-14-laion2B-39B-b160k* from Huggingface. The multimodal audio-text encoder used is *laion/larger_clap_general* from Huggingface (Wu* et al., 2023).

Unimodal Encoders. We use several unimodal encoders and show the difference in performance in Section E. To encode images, we tested DINOv2-Giant (Oquab et al., 2023) and the unimodal part of the multimodal encoders previously mentioned. To encode text, we tested GTR-t5-large (Ni et al., 2022) and the unimodal part of the multimodal encoders mentioned above. To show CSA's ability to combine unimodal models, we never tried using the paired unimodal encoders of a multimodal encoder in our experiments, *i.e.*, using CLIP to encode both images and text.

Flickr30k. We trained ASIF and CSA on the Flickr validation set, which includes $145,000$ images and 5 captions for each image. We then validated the models on a test set of $5,000$ images and $25,000$ captions.

COSMOS. We trained ASIF and CSA on the COSMOS validation set, which includes $41,006$ image-caption pairs. We then validated the models using a test set of 1,700 image-caption pairs, with half of the captions labeled as misinformation by human annotators.

B TOWARDS MORE MODALITIES—AUDIO AND TEXT

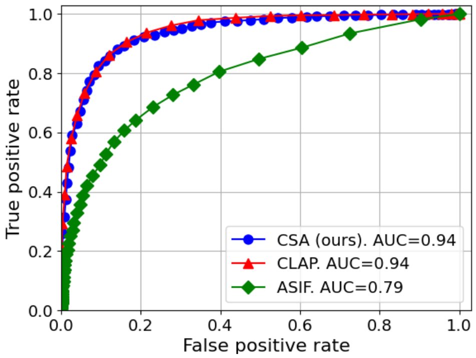


Figure 7: **Classification of YouTube audio and genre tags:** CSA (blue) performs as well as CLAP, the CLIP-inspired multimodal audio and text encoder, and outperforms ASIF in classifying genre tags of YouTube audio.

We now show CSA's generalization ability to more modalities with MusicCaps (Agostinelli et al., 2023). We use GTR and CLAP to encode YouTube audio along with the tagged genre descriptions of the audio. We conducted a classification task in which the models input the audio and a tag and output if the audio aligns with the caption. Similar to the mislabeled ImageNet experiment, we show the ROC curves and compare the AUC in Figure 7. We trained ASIF and CSA for $3,777$ data points and tested all methods on $1,625$ data points. We randomly sampled a tag for each data point during both training and inference. In Figure 7, we see that CSA performs as well as CLAP, the CLIP-inspired multimodal audio and text encoder, and outperforms ASIF. Thus, we conclude that CSA extends its capabilities beyond image and text, effectively handling audio and text as well.

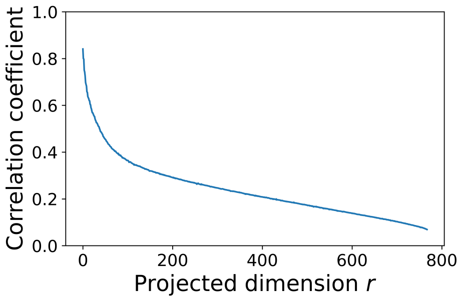


Figure 8: **Correlation coefficients of COSMOS image and caption features under CSA:** The data are inherently noisy, as indicated by the correlation coefficients of the unimodal feature spaces, which concentrate on 0.2 to 0.4. The unimodal encoders here are GTR and DINOv2.

C CORRELATION OF FEATURE SPACES

To take a deeper look into unimodal feature spaces, we show the correlation coefficients of COSMOS image and caption features under CSA in Figure 8. The data are inherently noisy, as indicated by the correlation coefficients of the unimodal feature spaces, which concentrate on 0.2 to 0.4. This distribution of correlation coefficients highlights that, despite the fact that the original multimodal data are noisy and show complex correlations, CSA can effectively map them to a multimodal space where the similarity score remains meaningful for the zero-shot downstream tasks.

D SENSITIVITY TO HYPERPARAMETER s

Method	s	mAP	Precision@1	Precision@5
CSA	10	9.5%	18.1%	13.4%
CSA	50	32.7%	53.9%	40.0%
CSA	100	36.0%	58.3%	42.9%
CSA	200	36.6%	59.3%	43.4%
CSA	500	31.8%	56.3%	38.3%
CSA	750	27.3%	50.2%	33.6%
CLIP	✗	73.8%	92.9%	77.2%
ASIF	✗	14.6%	25.6%	20.0%

(a) Image-to-text retrieval.

Method	s	Precision@1
CSA	10	15.2%
CSA	50	41.2%
CSA	100	43.8%
CSA	200	44.7%
CSA	500	41.7%
CSA	750	40.1%
CLIP	✗	79.5%
ASIF	✗	0.1%

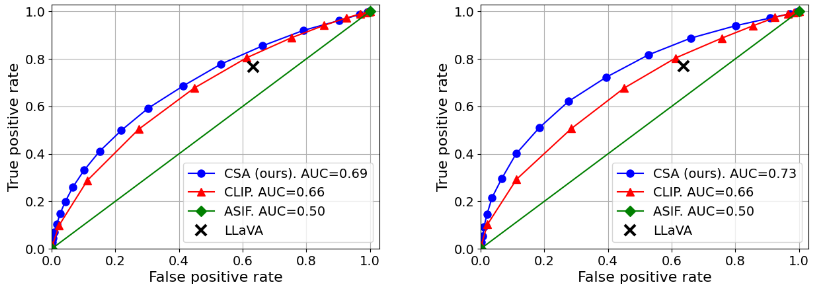
(b) Text-to-image retrieval.

Table 3: **Cross-modal retrieval on Flickr30k under different s:** CSA achieves optimal performance at $s = 200$, and its performance degrades with increases and decreases in s, illustrating the trade-off characterized in Section 5.

We show CSA's sensitivity to the hyperparameter s in terms of the end performance. In Table 3, CSA achieves optimal performance in $s = 200$, and its performance degrades with increases and decreases in s, illustrating the trade-off characterized in Section 5. However, for tasks other than retrieval, we find that a larger s improves performance in image classification, mislabeling detection, and misinformation caption detection. This phenomenon is likely due to the trade-off between distinguishability and informative embedding features, namely the distance between features. Although retrieval tasks require a more curated balance between these aspects, other tasks benefit from greater distinguishability of similarity scores.

E SENSITIVITY TO UNIMODAL ENCODERS

We change the unimodal encoders of ASIF and CSA to showcase their generalization ability to different unimodal encoders. Figure 9 shows the results on the detection of mislabeled ImageNet data with other encoders, and Figure 10 shows the results on the detection of misinformative captions



(a) Results for a training size of 5000.

(b) Results for a training size of 35000.

Figure 9: **Detecting mislabeled ImageNet images (cont'd):** CSA (blue) outperforms CLIP, ASIF, and LLaVA with a higher AUC. (a) and (b) illustrate the results for CSA and ASIF across various training set sizes, showing the superior performance of CSA with limited noisy training data. The unimodal encoders are GTR and CLIP (image) here.

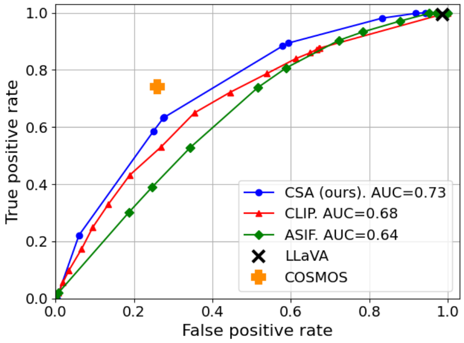


Figure 10: **Detecting misinformative COSMOS captions (cont'd):** CSA (blue) outperforms CLIP, ASIF, and LLaVA with a higher AUC. The supervised-learning method from the original COSMOS paper is the orange cross. It is the only method that outperforms CSA, though trained with supervised labels of object locations. The unimodal encoders are GTR and CLIP (image) here.

with other encoders. CSA again outperforms ASIF and CLIP while outperforming the results of the combination of GTR and DINOv2 in Section 6.