Improving Multimodal Data Quality with Unified Filtering Score (UF-Score)

Sangyeon Cho[1], Mingi Kim[1], JinKwon Hwang[1], Jaehoon Go[1], Minuk Ma[2], and Junyeong Kim[1],

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Department of Computer Science, UBC, Canada
whtkddus98@cau.ac.kr, mingi8233@cau.ac.kr, wlsrnjs905@cau.ac.kr, gkdwngo@cau.ac.kr, minukma@cs.ubc.ca,
junyeongkim@cau.ac.kr

Abstract

Multimodal models have made significant strides in handling diverse downstream tasks, yet the quality of the datasets they rely on remains a critical challenge. While large-scale datasets encompassing multiple modalities like image, text, and audio are crucial for training such models, these datasets often contain noisy data, which hampers their performance. Existing approaches primarily filter datasets based on pairwise modality alignment, which is insufficient for datasets with three or more modalities. To address this, we propose a novel filtering method leveraging the Unified Filtering Score (UF-Score), which evaluates data quality by considering the mean and variance of alignment scores across all possible modality pairs. Using modality-specific encoders, alignment scores are computed via cosine similarity within a shared embedding space. Our approach effectively filters low-quality data, retaining subsets that maximize alignment quality. Experiments demonstrate that this method significantly improves performance across multimodal tasks, even with reduced dataset sizes.

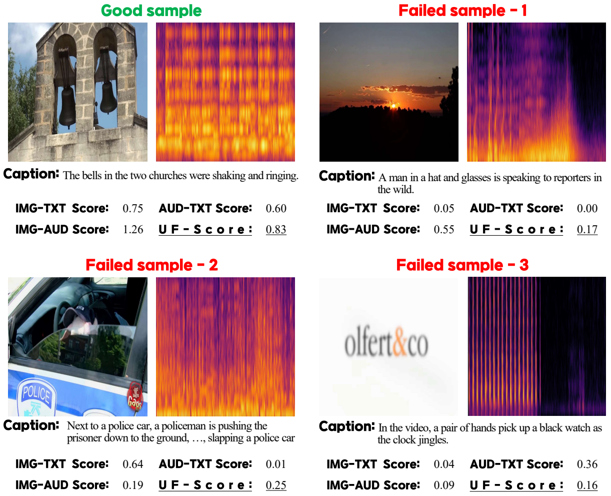

Figure 1: **Examples of Aligned and Misaligned Multimodal Samples.** Good sample exhibit well-aligned multiple modalities, whereas failed samples show misalignment in one specific modality: failed sample-1 (caption), failed sample-2 (audio), and failed sample-3 (image). "Score" represents cosine similarity, and "UF-Score" is our proposed metric, which is lower for the affected modality in failed samples.

Introduction

Recently, multimodal models such as CLIP(Radford et al. 2021), Flamingo(Alayrac et al. 2022), and PaLM-E(Driess et al. 2023) have made rapid advancements, demonstrating strong performance across various downstream tasks including retrieval and zero-shot learning. As the scale of models that process multiple modalities increases, the demand for large-scale multimodal datasets has also grown according to scaling laws(Kaplan et al. 2020). In this context, several studies have proposed extensive image-text pair datasets(Chen et al. 2015; Sharma et al. 2018; Schuhmann et al. 2022). Moreover, recent studies go beyond image-text models, proposing models that incorporate various modalities such as image-audio-text(Rubenstein et al. 2023; Zhang, Li, and Bing 2023).

Most large-scale multimodal datasets are generated based on web data, which captures the rich and diverse real world content. However, such web-crawled large datasets contain a significant amount of noise, which can hinder the model's optimal training(Li et al. 2022). Therefore, recent research has focused on enhancing the performance of multimodal models by filtering noisy large-scale web-crawled image-text data to improve quality(Kang et al. 2023; Nguyen et al. 2024; Fan et al. 2024). Additionally, there are studies that integrate the filtering process into the training process to enhance learning efficiency(Evans et al. 2024; Xu et al. 2023). While significant progress has been made in filtering image-text pair datasets(Sun et al. 2023), research on filtering multimodal datasets that include three or more modalities, such as images, text, and audio, remains relatively underexplored. Moreover, relying solely on the alignment between two modalities, such as images and text, is insufficient for fully assessing data quality. This limitation arises because evaluating alignment across multiple modalities cannot be based exclusively on the assessment of individual modality pairs. In datasets encompassing multiple modalities, it is essential to evaluate alignment across all modality combinations. For instance, even if images and text are well-aligned, other pairs, such as audio and text, may exhibit poor alignment.

As shown in Figure 1, a good sample—where every modality is well-aligned—displays high alignment scores across all modality pairs. In contrast, failed samples demonstrate strong alignment between two specific modalities but poor alignment with others. This illustrates that in datasets encompassing diverse modalities, assessing data quality solely based on image-text alignment is insufficient.

In this study, we propose a filtering method to filtering multimodal datasets aiming to identify the optimal subset. By considering the alignment levels across multiple modalities, we filter out low-quality data to improve the efficiency of multimodal alignment learning. Inspired by CLIP-Score(Hessel et al. 2021), we quantify the alignment levels between modalities using cosine similarity across multiple modalities that exist within a shared embedding space. Then we apply filtering based on the distribution of these alignment scores. Through several downstream tasks, we empirically demonstrate that even a smaller, filtered dataset can achieve effective multimodal alignment.

Method

We suggest a method for filtering datasets that include three or more modalities. Therefore, we mainly consider datasets containing three or more modalities and refer to these as *K-modality* datasets for convenience. The notation is defined as follows: $M = \{m_i\}_{i=1}^{K}$, where K is the total number of modalities, and m_i represents the i-th modality. In other words, a single data instance M is assumed to consist of K distinct modalities.

Filtering Through Alignment

We need encoders E_i that map each modality m_i into a shared embedding space to calculate the alignment of a *K-modality* dataset. To achieve this, we use LanguageBind(Zhu et al. 2023), a model pretrained on the large-scale multimodal dataset VIDAL-10m, to perform various multimodal semantic alignments centered on language. Since most multimodal alignment training is conducted through self-supervised manner such as contrastive learning(Chen et al. 2020b), we utilize cosine similarity as the alignment score. Specifically, the alignment score is calculated as follows:

$$\text{Align\_Score}_{i,j} = w * \max\left(\cos(m_i, m_j), 0\right) \tag{1}$$

where cos denotes the cosine similarity score. Following the approach of CLIPScore(Hessel et al. 2021), negative values were removed, and w was set to 2.5. Using modality-specific encoders, we compute the alignment between any two distinct modalities among the K modalities. However, evaluating data quality based solely on the alignment of a single modality pair is insufficient, as some pairs may align while others do not. Therefore, we calculate alignment scores for all possible modality pairs and assess the quality of a specific data point M by considering the mean and variance of these scores. For a single data point M, the mean alignment score across different modality pairs is calculated as follows:

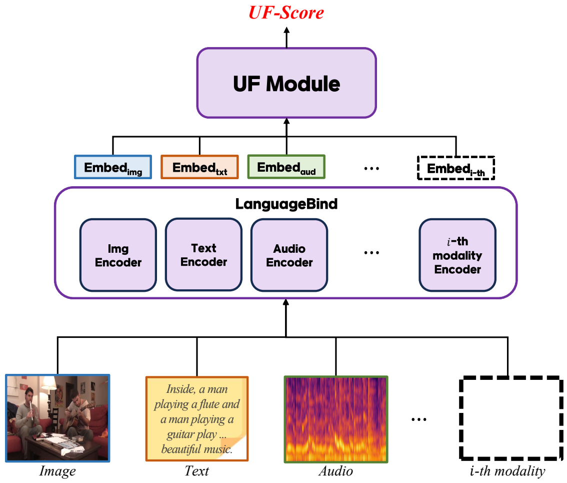

Figure 2: **Proposed Scoring Algorithm.** The UF-Score evaluates a multimodal data instance by considering the alignment score across all modalities.

$$\mu = \mathbb{E}[\text{Align\_Score}_{i,j}] = \frac{1}{N} \sum_{1 \le i < j \le K} \text{Align\_Score}_{i,j} \tag{2}$$

In a similar way, we can get the variance of the alignment scores.

$$\begin{aligned} \sigma^2 &= \text{Var}(\text{Align\_Score}_{i,j}) \\ &= \frac{1}{N} \sum_{i<j} \left(\text{Align\_Score}_{i,j} - \mu\right)^2 \end{aligned} \tag{3}$$

Here, since we calculate the alignment scores for each pair of distinct two modalities among the K modalities, N is defined as $\frac{K(K-1)}{2}$. We should consider the mean alignment score because a low mean indicates an overall low level of alignment. This suggests that multiple modalities exhibit weak alignment, potentially indicating the presence of noise in the data. In that context, we also should take into account the variance of the alignment scores. A high variance implies that some modality pairs are well-aligned while others are not, which can hinder the learning of overall alignment in a *K-modality* dataset.

Therefore, we propose the Unified Filtering Score for multimodal datasets (UF-Score), which takes both the mean and variance of alignment scores across different modality pairs into account. The UF-Score is calculated as follows:

$$\text{UF-Score} = \mu + \alpha \times \sigma^2 \tag{4}$$

Here, α is a hyperparameter that adjusts the weighted sum of the mean and variance. This allows for the simultaneous consideration of the central tendency and variability of the data, enabling a comprehensive evaluation of both the mean and the variance. Generally, our objective is to identify data with a large mean and a small variance. Consequently, α is set to a value less than zero. In the next section, we experimentally demonstrate how the UF-Score can be used to identify an optimal subset that outperforms the entire dataset.

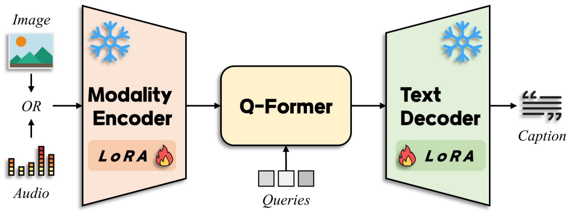

Figure 3: **Model Architecture.** An image or audio input is processed through a modality encoder with LoRA tuning, followed by a Q-Former, and then passed to a text decoder (also using LoRA) to generate captions.

Experimental Setup

Implementation Details

To evaluate the proposed UF-Score methodology in this study, we conduct a multimodal task, performing both Image Captioning (Vinyals et al. 2016) and Audio Captioning (Drossos, Adavanne, and Virtanen 2017) tasks. Inspired by the structure of BLIP (Li et al. 2023), we designed our model, with the overall architecture illustrated in figure 3. Our model utilizes a vision transformer (Dosovitskiy 2020) as the image encoder and an audio transformer (Dinkel et al. 2024) as the audio encoder, while the text decoder utilizes LLaMA-2 7B (Touvron et al. 2023). Q-former (Li et al. 2023) is integrated between the encoder and decoder, and both encoder and decoder are trained using LoRA (Hu et al. 2021) tuning.

For model training across both image and audio captioning tasks, we employed the AdamW (Loshchilov 2017) optimizer with β parameters set to $(0.9, 0.999)$. The learning rate is set at 5×10^{-6}, weight decay at 1×10^{-6}, with a batch size of 96, over 15 epochs, of which the initial 2 epochs use learning rate warm-up. To augment the dataset, we applied RandomResizedCrop (He et al. 2016) to images and used four audio augmentation techniques (Ko et al. 2015) —AddWhiteNoise, Shifting, Stretching, and Flipping—for audio data.

Datasets

We evaluate the proposed UF-Score based filtering methodology using two datasets. The first is the small-scale VALOR-32K (Chen et al. 2023) dataset, consisting of 32K videos paired with corresponding text descriptions. For this study, we apply the proposed methodology exclusively to the training set, which contains 25K samples. Additionally, we extract relevant snapshots (images) and audio from the videos to utilize in multi-modal downstream tasks. The second dataset is the large-scale VggSound (Chen et al. 2020a) dataset, comprising 200K visual-audio samples. Since VgSound does not include text descriptions aligned with the visual-audio data, we performed additional LLM-based (GPT-4o) (Shahriar et al. 2024) captioning to construct the dataset. Similar to VALOR, the proposed methodology is applied only to the training set for the experiments.

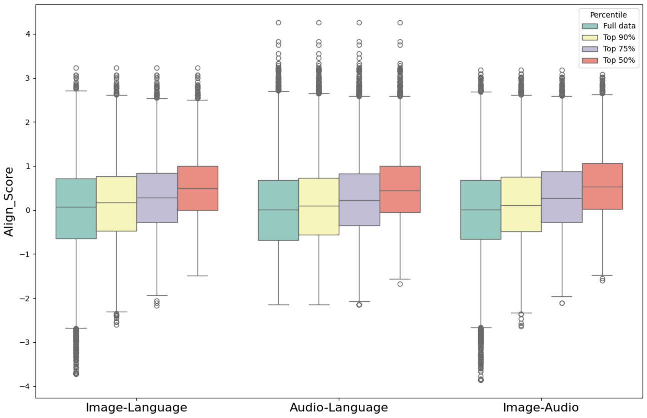

Figure 4: **Align Score Distributions Across Full and Top Percentile Groups.** We compare the distribution of alignment scores between two modalities under varying filtering ratios on the VALOR-32K. As the filtering ratio increases, the alignment scores tend to improve overall.

Metrics

We utilize a variety of evaluation metrics, including BLEU (BL) (Papineni et al. 2002), ROUGE-L (RG-L) (Lin 2004), METEOR (ME) (Denkowski and Lavie 2014), CIDEr (CD) (Vedantam, Lawrence Zitnick, and Parikh 2015), SPICE (SP) (Anderson et al. 2016), SPIDEr (SD) (Liu et al. 2017), SPIDEr-FL (SD-F) (Labbe, Pellegrini, and Pinquier 2022), Sentence-BERT (SB) (Reimers 2019), and FENSE (FS) (Zhou et al. 2022). These metrics are commonly used in natural language processing to assess language generation by measuring the similarity between generated and reference texts. BLEU and ROUGE-L are based on n-gram precision, while METEOR adjusts scores based on lexical similarity. CIDEr and SPICE evaluate semantic similarity, with SPIDEr and SPIDEr-FL combining CIDEr and SPICE to provide more refined assessments. Finally, Sentence-BERT and FENSE use embedding-based approaches to measure semantic similarity between sentences.

Results

Analysis of UF-Score Filtering

When we filter the data based on the UF-Score, we aim to examine how the alignment scores between specific modalities change with varying filtering ratios. The results are presented in Figure 4. We observed that as more data is filtered from the entire dataset, the average Align_Score between each modality increases. Additionally, outliers, depicted as points below the boxes, are effectively removed. This indicates that our UF-Score is suitable as an integrated metric for assessing the alignment among multiple modalities.

Multimodal Downstream Tasks on Filtered Data

In this section, we analyze the results of performing multimodal downstream tasks using the data filtered with the proposed UF-Score methodology.

In VALOR. We conducted image and audio captioning tasks on the VALOR-32K (Chen et al. 2023) dataset, with

	Image Captioning									Audio Captioning								
	BL	RG-L	ME	CD	SP	SD	SD-F	SB	FS	BL	RG-L	ME	CD	SP	SD	SD-F	SB	FS
Full data	27.2	22.3	9.0	8.1	7.3	7.7	7.7	13.1	13.1	**31.5**	25.7	11.7	18.9	12.0	15.4	13.4	36.0	29.5
Top 90%	**28.3**	**24.5**	**10.5**	**7.9**	**7.7**	**7.8**	**7.8**	**16.9**	**16.9**	30.4	24.9	11.1	15.8	10.7	13.2	11.4	33.3	27.2
Top 75%	26.3	23.0	9.6	6.6	7.3	7.0	6.0	16.1	13.5	31.2	**25.9**	**11.9**	**20.5**	**13.1**	**16.8**	**14.7**	**37.5**	**30.9**
Top 50%	26.7	23.2	9.5	6.6	7.4	7.0	6.1	16.7	14.4	30.2	24.9	10.8	15.1	10.2	12.7	11.0	32.0	26.3

Table 1: Performance of image/audio captioning task at different filtering ratios in VALOR-32k

	Full data	Top 80%	Top 50%
BL-1	33.0	**34.3**	33.8
BL-2	18.3	**19.4**	18.6
BL-3	10.9	**11.9**	11.1
BL-4	7.0	**7.9**	7.3
RG-L	29.1	**29.9**	29.4
ME	12.4	**14.0**	13.1
CD	47.6	**57.4**	51.6
SP	14.2	**15.6**	14.9
SD	30.9	**36.6**	33.4
SD-F	27.6	**31.0**	28.9
SB	61.8	**63.6**	62.1
FS	**55.3**	53.5	53.0

Table 2: Performance of audio captioning task at different filtering ratios in VggSound. The top 50% of the data demonstrates overall superior performance compared to the full data environment.

the results presented in table 1. The filtering ratios were set to the top 90%, 75%, and 50% for the experiments. Since VALOR is a high-quality dataset with human-labeled annotations, we set lower filtering ratios. Although results vary by task, we observe performance improvements when filtering is applied. By applying the proposed filtering methodology, samples with relatively weaker alignment between modalities are excluded from training, allowing the model to learn from higher-quality data. This process effectively removes data with alignment noise between modalities, resulting in a optimal subset compared to the full dataset.

In VggSound. Additionally, we conducted an audio captioning task on the VggSound (Chen et al. 2020a) dataset to validate the effectiveness of the proposed filtering methodology on a larger-scale dataset. The results are presented in Table 2. We experimented with filtering ratios of 80% and 50%. Since VggSound is a dataset with text descriptions generated based on an LLM, we assumed a lot of noise in modality alignment. Therefore, unlike the VALOR experimental setting, we applied higher filtering ratios. The experimental results indicate that the proposed methodology is also effective on the VggSound dataset. Specifically, using only the top 80% of the data for training yielded the best results. Notably, the results using only the top 50% of the data were superior to those obtained with the full dataset, which is remarkable. This suggests that the subset filtered using UF-Score is strongly aligned in a multimodal sense, positively impacting the training process.

Limitations

Computational cost. Calculating the proposed UF-Score requires generating embeddings for all samples, which consumes significant computational resources. As the dataset size increases, these computational costs can grow exponentially, potentially limiting the scalability of the methodology for very large datasets(Goyal et al. 2024).

Filtering accuracy. A simple scoring-based filtering approach may not effectively address false positives or false negatives due to biases in the pretrained data of the scoring model (Mahmoud et al. 2024). Therefore, it is necessary to consider more sophisticated filtering methods that are specifically designed to mitigate these biases and accommodate the specific conditions of our dataset.

Conclusion and Next Step

Conclusion This study highlights the limitations of conventional unified data filtering methodologies and proposes UF-Score, an approach applicable to various modality combinations. The UF-Score-based filtering method effectively eliminates low-quality data by filtering out samples with relatively weak alignment levels, thereby generating high-quality subsets better suited for model training. Experimental results on the VALOR (Chen et al. 2023) and VggSound (Chen et al. 2020a) datasets demonstrate that using the filtered data subsets can lead to performance improvements in multimodal downstream tasks. This suggests that UF-Score contributes to enhancing dataset quality and improving training efficiency. Consequently, UF-Score is a promising methodology for facilitating performance improvements in multimodal learning by creating refined, noise-reduced data subsets.

Next Step To address the limitations of this study, future research will focus on improving the computational efficiency of UF-Score. This may involve exploring approximate embedding techniques or selective sampling methods to enhance processing speed and reduce computational costs. Additionally, we plan to analyze filtering failure cases to develop scoring methods that enable more accurate filtering. This includes investigating adaptive scoring techniques and sophisticated scoring mechanisms that incorporate condition-based strategies. By doing so, we aim to enhance the robustness of UF-Score and strengthen its scalability to ensure stable performance across diverse large-scale multimodal datasets.

Acknowledgements

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00253782) and partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01341, Artificial Intelligence Graduate School Program, Chung-Ang University)

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020a. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, S.; He, X.; Guo, L.; Zhu, X.; Wang, W.; Tang, J.; and Liu, J. 2023. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Dinkel, H.; Wang, Y.; Yan, Z.; Zhang, J.; and Wang, Y. 2024. CED: Consistent ensemble distillation for audio tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 291–295. IEEE.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Drossos, K.; Adavanne, S.; and Virtanen, T. 2017. Automated audio captioning with recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 374–378. IEEE.

Evans, T.; Parthasarathy, N.; Merzic, H.; and Henaff, O. J. 2024. Data curation via joint example selection further accelerates multimodal learning. *arXiv preprint arXiv:2406.17711*.

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36.

Goyal, S.; Maini, P.; Lipton, Z. C.; Raghunathan, A.; and Kolter, J. Z. 2024. Scaling Laws for Data Filtering–Data Curation cannot be Compute Agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22702–22711.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. For RandomResizedCrop augmentation in image processing.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kang, W.; Mun, J.; Lee, S.; and Roh, B. 2023. Noise-aware learning from web-crawled image-text data for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2942–2952.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Ko, T.; Peddinti, V.; Povey, D.; Seltzer, M. L.; and Khudanpur, S. 2015. Audio augmentation for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3586–3590. For AddWhiteNoise, Shifting, Stretching, and Flipping in audio augmentation.

Labbe, E.; Pellegrini, T.; and Pinquier, J. 2022. SPIDEr-FL: An extension of SPIDEr for evaluating fluency and linguistic diversity.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, 873–881.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mahmoud, A.; Elhoushi, M.; Abbas, A.; Yang, Y.; Ardalani, N.; Leather, H.; and Morcos, A. S. 2024. Sieve: Multimodal dataset pruning using image captioning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22423–22432.

Nguyen, T.; Gadre, S. Y.; Ilharco, G.; Oh, S.; and Schmidt, L. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reimers, N. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Rubenstein, P. K.; Asawaroengchai, C.; Nguyen, D. D.; Bapna, A.; Borsos, Z.; Quitry, F. d. C.; Chen, P.; Badawy, D. E.; Han, W.; Kharitonov, E.; et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shahriar, S.; et al. 2024. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *arXiv preprint arXiv:2407.09519*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Touvron, H.; Martin, L.; Stone, K. R.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4): 652–663.

Xu, H.; Xie, S.; Huang, P.-Y.; Yu, L.; Howes, R.; Ghosh, G.; Zettlemoyer, L.; and Feichtenhofer, C. 2023. Cit: Curation in training for effective vision-language data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15180–15189.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhou, Z.; Zhang, Z.; Xu, X.; Xie, Z.; Wu, M.; and Zhu, K. Q. 2022. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 981–985. IEEE.

Zhu, B.; Lin, B.; Ning, M.; Yan, Y.; Cui, J.; Wang, H.; Pang, Y.; Jiang, W.; Zhang, J.; Li, Z.; et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.