Character-aware Subtitling through Improved Alignment and Diverse Priors

Adhiraj Deshmukh
CVIT, IIIT Hyderabad
adhiraj.deshmukh@research.iiit.ac.in

Makarand Tapaswi
CVIT, IIIT Hyderabad
makarand.tapaswi@iiit.ac.in

Vineet Gandhi
CVIT, IIIT Hyderabad
vgandhi@iiit.ac.in

***Abstract*—Subtitles are essential for accessibility and global dissemination of video content, yet conventional formats often omit critical cues such as speaker identities. Character-Aware Subtitling (CAS) aims to address this gap, however, existing approaches are limited by the absence of public checkpoints, weak pipeline coordination leading to detrimental performance, and insufficient failure analysis. This work presents a systematic study of subtitling pipelines, demonstrating that improved transcript–speech alignment substantially enhances downstream performance, achieving state-of-the-art results on the LLR dataset by 2.77% DER-point zero-shot, and by 4.78% DER-point after VAD finetuning. In addition, we propose design refinements, adopt open, out-of-the-box model alternatives, and introduce a scalable method for curating diverse character priors, thereby improving the robustness and replicability of CAS.**

I. Introduction

With the growth of streaming and short-form video content, subtitles have become increasingly critical for accessibility and global reach. However, conventional subtitles only provide timestamped dialogues, offering limited support for audio-impaired audiences and often falling short of SDH standards [1] by omitting essential cues such as speaker identification, sound effects, and background music. Incorporating these elements not only improves accessibility, but also enriches video datasets with annotations that facilitate the development of models for higher-level narrative understanding [2], [3].

Character-Aware Subtitling (CAS) task addresses part of this challenge by identifying *what* is spoken *when*, and by *whom* in a video (Fig. 1). Unlike conventional subtitling, CAS requires integrating speech, vision, and contextual cues. Although Active Speaker Recognition (ASR) [4], [5], Speaker Diarization [6], [7], and Character Recognition [8] have been well studied individually, only recently have a few works [9], [10] begun to address the full CAS task by integrating these subtasks through multi-stage pipelines and providing curated datasets. While promising, these approaches require further analysis of failure cases and component contributions. Moreover, reliance on models without public checkpoints (e.g., CLIP-PAD [11]) and limited details on character-prior curation hinder reproducibility and fair comparison.

We build on this line of work with three main contributions:

1) Through component-wise error analysis, we identify pipeline design choices that hinder performance.
2) We propose a simple Voice Activity Detection (VAD) filtering step that substantially improves transcript–speech alignment and downstream identification. Furthermore, fine-tuning it on a single episode from the target-domain exceeds prior results on the LLR dataset [9], without requiring post-hoc refinements from previous work [10].
3) We improve method replicability by employing open-source components and proposing a scalable method for character-prior curation from web-search results (e.g., Google Images [12]), corroborated by online databases (e.g., IMDb [13]) to filter noisy results.

II. Related Works

A. Automated Video Character Labeling

This task has been extensively studied for its utility in curating character-annotated video datasets, crucial for developing models with higher-level narrative understanding [3], [14]. Prior works used strong priors such as scripts, timed subtitles, or curated actor images [15], [16] to do so. In contrast, [17] proposes a scalable alternative using automatically retrieved web images as corroborative evidence for character identities. Following this, we use web search (e.g., Google Images [12]) to collect diverse character priors, and online databases (e.g., IMDb [13]) to validate identities and reduce noise.

B. Audio-only Speaker Diarization

This task involves segmenting an audio stream by speaker, answering *who* spoke *when*. It incorporates many components relevant to CAS, typically based on clustering [6] or end-to-end architectures [7]. These methods assume that speaker embeddings form separable clusters in the embedding space. However, this assumption becomes less reliable as number of speakers increases and acoustics conditions or domains vary, especially in TV shows with short dialogue segments, thereby reducing clustering purity.

C. Audio-Visual Person Identification

Limitations of audio-only diarization can be mitigated by leveraging cross-modal cues to improve cluster purity [17], [18]. Building on this, [9], [10] combine: (i) Visual Person Identification models that are effective for short segments, and (ii) Active Speaker Detection (ASD) models to transfer identities to remaining segments through audio.

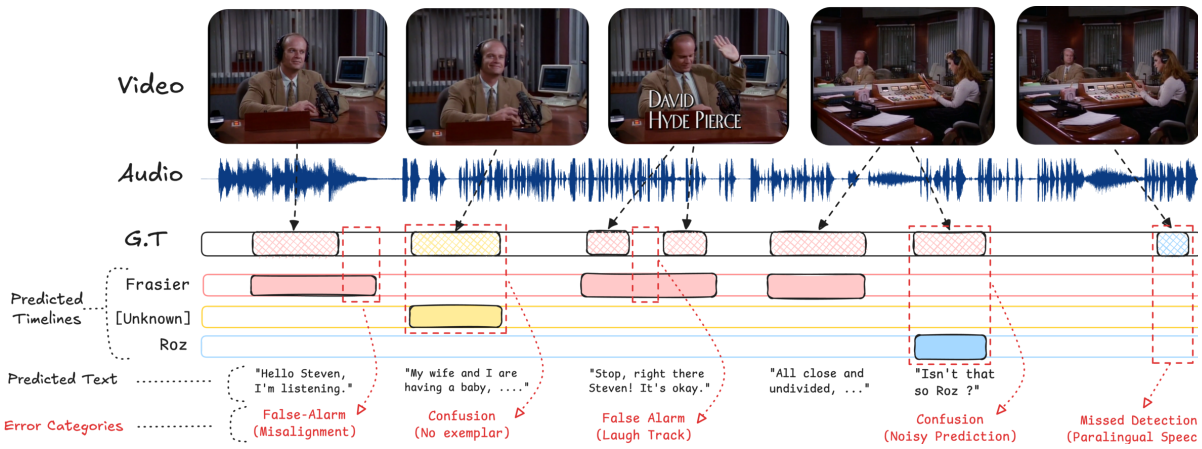

Fig. 1. Overview of the *Character-aware Subtitling* task: predicting *what*, *when*, and *whom* for conversational segments. Common errors in recent methods include transcript–speech misalignments, missing character priors, noisy acoustics (e.g., laugh tracks), and paralinguistic speech (e.g., emotive sounds). Errors are categorized into three types: (i) **Confusion**: predict incorrect speaker identity, (ii) **False Alarm**: predict speech over ground-truth non-speech, and (iii) **Missed Detection**: missed ground-truth speech.

D. Differences to Prior Work

LLR [9] often struggles with short utterances, which constitute majority of the LLR dataset ($\sim 71\,\%$ segments are $< 2\,\mathrm{s}$ [10]). Such segments often fall outside reliable speech-encoder distributions, yielding poor identification performance. CAS [10] addresses this with a refinement stage for low-confidence predictions, leveraging limited local scene-level acoustic variation to classify them using nearby labelled segments. Then remaining unassigned segments are resolved using local dialogue context via LLM-based reasoning. Since these refinements mainly target short segments, they yield only incremental downstream gains. In contrast, we systematically identify current system-level errors and show that improving transcript–speech alignment and fine-tuning VAD on target domain yields substantially larger improvements without relying on such post-hoc refinement strategies.

III. Method

We build upon the approaches of LLR [9] and CAS [10] for character-aware subtitling using video data and their cast list. As illustrated in Fig. 2, our multi-stage pipeline first uses ASR to transcribe the video into speech utterances. These utterances are then refined by the proposed VAD filtering stage, fine-tuned on a single episode per series, to improve temporal alignment with speech and filter ASR hallucinations. Next, our prior curation pipeline gathers diverse actor images from web search to curate visual priors for each character (Fig. 3). These visual priors enable high-confidence classification of audio segments for which active speakers are visible, yielding a gallery of audio exemplars. Finally, audio exemplars are refined with k-NN filtering and used for centroid-based classification of remaining segments, while segments exceeding a predefined distance threshold remain unlabelled.

A. Stage-1: Transcription

Following prior work, we use WhisperX [5] to generate time-stamped ASR segments. These are split into speaker-specific segments under the assumption that each transcribed sentence is spoken by a single speaker. A VAD model [6] is typically paired with ASR to isolate speech regions and mitigate hallucinations. However, ASR performance degrades in sitcom-like domains with rapid turn-taking, laugh tracks, and background noise, amplifying misalignments in the *VAD cut-and-merge* step from WhisperX pipeline. We address this by introducing an additional post-ASR VAD filtering step, that refines sentence-level segments by trimming or splitting them, thereby correcting any misalignments introduced by ASR . To bridge the domain gap, we fine-tune the VAD model on the validation set of the LLR dataset. These adaptations substantially improve the alignment between generated transcripts and ground-truth speech segments, thus enhancing downstream identification performance and enabling our approach to outperform prior work on the LLR dataset.

B. Stage-2: Building Exemplars

Audio exemplars are speech segments where the speaker is visible and clearly identifiable using visual cues.

1) ***Audio-Visual Speaker Detection****:* Following CAS [10], we use an Active Speaker Detection (ASD) model [19] to identify visible speaker segments, and obtain audio exemplars.

2) ***Visual Prior Curation****:* Prior work [2], [9] typically use actor profile images from online databases (e.g., IMDb [13]) as visual cues. While effective, these profiles lack diversity (e.g., mostly frontal views), miss appearance variations, and often omit minor characters. We instead retrieve diverse actor images via web search, using show-specific (in-context) and general (out-of-context) search queries by including or omitting the series name. This scalable approach benefits when shows have large casts with many minor characters (e.g., *Scrubs* 56-character cast). Database profiles still remain useful for corroborating and filtering noisy results via visual recognition with a fixed threshold. For characters without IMDb profiles, we use top-3 in-context results instead. For a fair comparison with prior work, we restrict the priors to 1–10

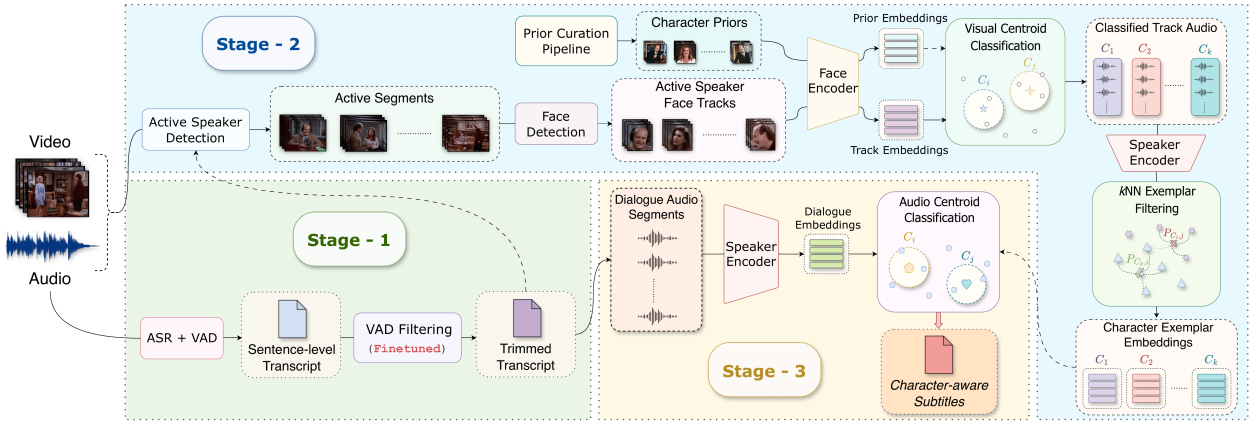

Fig. 2. Overview of the proposed *Character-aware Subtitling* pipeline. (i) **Stage-1** (Transcription): An ASR pipeline generates sentence-level transcripts, which are refined by fine-tuned VAD filtering step to improve alignment with speech and discard hallucinations. (ii) **Stage-2** (Exemplar Building): Visible speakers are detected from refined speech segments using ASD and visually classified with curated web priors (see Fig. 3). High-confidence visual matches are then used to construct audio exemplars, and refined via kNN filtering. (iii) **Stage-3** (Segment Classification): Remaining segments are classified by computing distances between speaker embeddings and audio exemplar centroids.

images per character, and find that using balanced in- and out-of-context web images as priors, yields the best performance.

*3) **Visual Classification**:* Active speaker segments are labelled by comparing averaged track embeddings to cluster centroids derived from visual priors. Segments above a threshold remain unassigned. Following [10], we use cropped face regions for ASD, yielding more exemplars than the global frame embeddings of [9]. Since [9], [10] rely on CLIP-PAD [11], without a publicly available checkpoint, we adopt ArcFace [20] as an off-the-shelf alternative.

*4) **Audio Exemplar Filtering**:* Prior methods [9], [10] further refine the audio exemplars using k-NN based filtering, retaining a segment only when their k=5 nearest exemplar audio embeddings share the same label. However, we find that using smaller k values, based on the validation set, prioritizes exemplar quantity and improves downstream classification performance with only marginal loss in exemplar precision

C. Stage-3: Classifying Segments

Audio exemplars from the previous stage are used to classify the remaining unassigned segments, whether or not the speaker is visible. Each segment is labelled based on their distance to the character-specific audio-exemplar centroids, similar to the visual classification step. Following prior work, we encode both segment and exemplar audio with ECAPA-TDNN [21]. Segments beyond a distance threshold to the nearest centroid are labelled `unknown`.

IV. Experiments & Results

A. Experimental Setup

***Implementation Details**:* Similar to prior work, we use the off-the-shelf WhisperX model [5] for ASR and the NLTK tokenizer for sentence-level segmentation. Voice embeddings are extracted with ECAPA-TDNN [21] pretrained on VoxCeleb [22], and visual embeddings from ArcFace [20] using the InsightFace library [23]. Unless specified otherwise, the remaining pipeline components follow CAS [10].

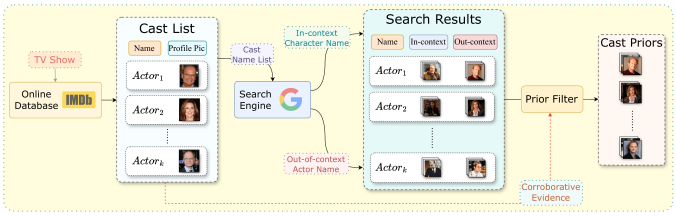

Fig. 3. Proposed *Prior Curation* pipeline, combining web search for diverse character images with online databases to filter noisy results.

***CAS Omissions**:* CAS [10] uses post-hoc refinement step that leverages local temporal context to revise low-confidence and unassigned predictions, yielding only modest gains. It also employs a Speech Enhancement (SE) model [24] for denoising input audio. However, in Section IV-B, we show that our Stage-2 modifications alone substantially improve over prior work, without such post-hoc steps. Moreover, SE degrades transcription quality and provides no downstream benefit. We therefore omit both components in our method.

***Dataset**:* We conduct our experiments on the LLR dataset [9], the only publicly available audio-visual TV/Movie dataset with diarization annotations. We exclude *Bazinga!* dataset [25] because it is audio-only. Following the evaluation protocol of prior work, we use the 6th episode of each series for validation and the remaining episodes for testing.

***Metrics**:* For downstream evaluation of character-aware subtitling, we report *Diarization Error Rate* (DER) and Character Recognition performance (*accuracy*, *precision*, *recall*) following prior work. We also include *Detection Error Rate* (DetER) and *Word Error Rate* (WER) to assess transcription alignment and textual accuracy. Exemplar performance is reported as *yield* and *accuracy* for *Main* and *Other* cast categories, grouped by frequency of occurrence, as in CAS [10]. A 0.25 s collar is applied wherever applicable.

TABLE I
ERROR BREAKDOWN (***Seg*** *- Avg. Error Segment Duration* (s) , ***Total*** *- Avg. Cumulative Error Duration* (s/ep))

Method	Confusion		False Alarm		Missed Detection		**Total Error**
	Seg	Total	Seg	Total	Seg	Total	
Base	0.49	64.36	0.31	129.11	0.08	**32.72**	225.45
Ours	0.43	42.51	0.11	**17.32**	0.13	57.35	**116.44**
w/o VAD FT	0.42	46.64	0.10	25.88	0.14	57.96	129.73
w/ Fixed k=5	0.48	51.20	0.11	17.32	0.13	57.35	125.13
w/ SE	0.44	**41.68**	0.11	30.42	0.13	59.38	130.69

Baselines & Proposed Method*:* We compare our approach against prior works LLR [9] and CAS [10], as well as a standard *Base* pipeline. *Base* approach corresponds to the 3-stage classification (Fig. 2), without the modifications to VAD or input pre-processing. Our final proposed method *Ours*, builds upon *Base* by integrating a 2nd VAD filtering step and fine-tuning it on target domain.

Ablations*:* To isolate the impact of our design choices, we evaluate 3 variants of proposed method: (i) *w/o VAD FT*: Removes the target-domain fine-tuning from the VAD step, evaluating impact of additional VAD filter alone. (ii) *w/ Fixed k=5*: Restricts our exemplar filtering to a strict $k = 5$, matching the standard used in prior works. (iii) *w/ SE*: Integrates a Speech Enhancement model on the input audio.

Hyper-parameters*:* We tune all pipeline hyper-parameters via grid search on the validation episode, selecting the best downstream configuration for each series.

B. Results & Discussion

1) ***Error Analysis****:* We analyze errors produced by these methods to better understand their failure modes. The small *avg. error segment duration* of $0.1\,\mathrm{s}$ to $0.15\,\mathrm{s}$ in Table I suggests that most False Alarms and Missed Detections arise from minor misalignments with ground-truth speech boundaries, motivating our VAD improvements. Stricter filtering (*w/ Fixed k*=5), reduces False Alarms substantially from 129.10 to 17.32 ($\sim 112\,\mathrm{s/ep}$), but at the cost of a modest increase in Missed Detections from 32.72 to 57.35 ($\sim 25\,\mathrm{s/ep}$).

Confusion Errors occur predominantly in short utterances ($\sim 0.5\,\mathrm{s}$), where state-of-the-art speaker encoders often fail to produce reliable embeddings. Although VAD filtering (*w/o VAD FT*) further reduces avg. Confusion duration by $0.06\,\mathrm{s}$, the Cumulative Errors also decrease by nearly $22\,\mathrm{s/ep}$, likely from corrected transcript–speech alignments. Overall, our approach (*Ours*) reduces cumulative errors by $\sim 110\,\mathrm{s/ep}$ relative to the *Base*.

2) ***Transcription Quality****:* Table II shows that adding a 2nd VAD filter (*w/o VAD FT*) reduces DetER by 12.65% on avg. for transcript alignment with ground-truth speech, with VAD finetuning (*Ours*) providing a further 1.5% reduction. Table III shows that Speech Enhancement (*w/ SE*) degrades transcription quality, likely by introducing artifacts and distribution shift, while providing no improvement in alignment.

TABLE II
TRANSCRIPT–SPEECH ALIGNMENT (*Detection Error Rate (**DetER** ↓)*)

Method	Frasier	Seinfeld	Scrubs	**Average**
Base	27.36	29.18	21.76	26.10
Ours	**11.82**	**11.89**	**12.14**	**11.95**
w/o VAD FT	14.29	13.91	12.15	13.45
w/ SE	14.63	14.83	13.30	14.25

TABLE III
TRANSCRIPTION ACCURACY (*Word Error Rate (**WER** ↓)*)

Method	Frasier	Seinfeld	Scrubs	**Average**
Whisper [4]	13.5	13.2	10.6	12.4
WhisperX [5]	**11.2**	**11.8**	**9.2**	**10.7**
WhisperX + SE [24]	14.7	15.2	13.4	14.4

TABLE IV
EXEMPLAR QUALITY (***Yield*** *(count)* , ***Accuracy*** *(%)*)

Method	Main Cast		Others		**Total**	
	Count	Acc.	Count	Acc.	Count	Acc.
LLR [9]	329	99.70	78	100	407	99.80
CAS [10]	1483	99.60	251	100	1743	99.65
Base	2035	96.56	404	95.79	2439	96.43
Ours	2357	96.73	470	96.17	2827	96.64
w/o VAD FT	2438	96.06	518	94.40	2956	95.77
w/ Fixed k=5	1780	99.72	272	97.79	2052	99.46
w/ SE	2296	96.08	456	95.39	2752	95.97

3) ***Exemplar Performance****:* Prior work emphasizes exemplar precision via strict filtering, we instead prioritize exemplar coverage by relaxing k. Table IV shows that improved speech–transcript alignment lets *Ours* yield 15.9% more exemplars compared to *Base*, while maintaining comparable precision. Reintroducing strict kNN filtering in ablation (*w/ Fixed k=5*) lowers exemplar yield and improves precision, bringing performance back in line with CAS [10].

4) ***Downstream Performance****:* Our findings validate the importance of exemplar quantity: priotizing high-precision exemplars through strict kNN filtering, as in prior work, leads to reduced downstream performance. Speech enhancement also yields no downstream gains, while substantially degrading transcription quality. Crucially, Table V shows that improved transcript alignment translates into significant downstream gains. Even without VAD fine-tuning (*w/o FT VAD*), it reduces DER by 15.33% points over *Base* and 2.77% over state-of-the-art. VAD fine-tuning (*Ours*) further reduces DER by 2.01%, gaining a total improvement of 4.78% over state-of-the-art.

These gains are also reflected in character-recognition performance (Table VI), with a 1.77% improvement in avg. accuracy. Notably, our approach surpasses prior work without any post-hoc refinements such as local audio embeddings or LLM-based reasoning, as in CAS [10].

5) ***Prior Curation****:* Table VII compares our web-search based prior curation against IMDb profile images. Across

TABLE V
DIARIZATION PERFORMANCE (*Diarization Error Rate (**DER** ↓)*)

Method	Frasier	Seinfeld	Scrubs	**Average**
SimpleDiar [26]	24.2	24.5	24.4	24.37
`pyannote` [6]	24.7	35.4	31.1	30.40
LLR [9]	26.4	28.0	26.7	27.03
CAS [10]	20.3	23.3	25.7	23.1
Base	35.57	37.07	34.35	35.66
Ours	**17.49**	**17.9**	**19.58**	**18.32**
w/o VAD FT	20.56	19.88	20.56	20.33
w/ Fixed k=5	18.36	18.11	22.39	19.62
w/ SE	20.66	21.04	20.93	20.88

TABLE VI
CHARACTER RECOGNITION PERFORMANCE (*Micro avg. **Accuracy**, **Precision**, and **Recall** across cast (%)*)

	Frasier			Seinfeld			Scrubs		
Method	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
LLR [9]	87.0	**91.6**	87.0	84.5	89.0	84.6	84.3	**89.2**	84.9
CAS [10]	88.9	89.8	88.9	85.8	87.1	86.0	84.4	84.8	85.1
Base	88.5	88.9	88.3	87.7	88.7	86.4	84.9	86.6	84.6
Ours	**90.2**	90.4	**90.0**	**88.6**	**89.9**	87.3	85.6	86.2	85.4
w/o VAD FT	88.6	90.5	88.4	88.2	89.7	86.7	84.6	86.1	84.4
w/ Fixed k=5	89.0	89.3	88.8	88.1	89.9	86.7	82.6	84.5	82.3
w/ SE	89.8	89.9	89.7	88.6	89.0	**87.4**	**85.9**	86.3	**85.7**

TABLE VII
PRIOR CURATION COMPARISON (*Diarization Error Rate (**DER** ↓)*)

	Frasier		Seinfeld		Scrubs		**Average**	
Method	IMDb	Web	IMDb	Web	IMDb	Web	IMDb	Web
Base	36.34	**35.57**	39.83	**37.07**	35.63	**34.35**	37.27	**35.66**
Ours	18.56	**17.49**	18.94	**17.90**	20.90	**19.58**	19.47	**18.32**

baselines, web-curated priors consistently reduce error rates, yielding average improvement of 1.3% DER points over IMDb priors. This highlights the role of high-quality priors in creating robust character exemplars.

V. CONCLUSION

In this work, we present an extensive analysis of design choices and failure cases in prior methods. Building on which, we propose improvements to transcript–speech alignment and exemplar construction. These enhancements achieve state-of-the-art performance on the LLR dataset without relying on the post hoc refinement strategies used in prior works. Future work can reduce reliance on labeled data for VAD adaptation and to explore architectural improvements.

REFERENCES

[1] A. Szarkowska, *Subtitling for the Deaf and the Hard of Hearing*. Springer International Publishing, 2020, pp. 249–268.

[2] T. Han, M. Bain, A. Nagrani *et al.*, "Autoad ii: The sequel-who, when, and what in movie audio description," in *International Conference on Computer Vision (ICCV)*, 2023.

[3] S. Bai, Y. Cai, R. Chen, *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[4] A. Radford, J. W. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, 2023.

[5] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.

[6] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Interspeech*, 2023.

[7] S. Horiguchi, Y. Fujita, S. Watanabe *et al.*, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," *arXiv preprint arXiv:2005.09921*, 2020.

[8] J. Huh, J. S. Chung, A. Nagrani *et al.*, "The vox celeb speaker recognition challenge: A retrospective," *ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[9] B. Korbar, J. Huh, and A. Zisserman, "Look, listen and recognise: Character-aware audio-visual subtitling," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[10] J. Huh and A. Zisserman, "Character-aware audio-visual subtitling in context," in *Asian Conference on Computer Vision (ACCV)*, 2024.

[11] B. Korbar and A. Zisserman, "Personalised clip or: how to find your vacation videos," in *British Machine Vision Conference (BMVC)*, 2022.

[12] "Google images," https://images.google.com.

[13] "International movie database," https://www.imdb.com.

[14] Y. Zhang, J. Wu, W. Li *et al.*, "Llava-video: Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[15] A. Nagrani and A. Zisserman, "From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script," *arXiv preprint arXiv:1801.10442*, 2018.

[16] M.-L. Haurilet, M. Tapaswi, Z. Al-Halah, and R. Stiefelhagen, "Naming tv characters by watching and analyzing dialogs," in *Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[17] A. Brown, E. Coto, and A. Zisserman, "Automated video labelling: Identifying faces by corroborative evidence," in *International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2021.

[18] A. Brown, V. Kalogeiton, and A. Zisserman, "Face, body, voice: Video person-clustering with multiple modalities," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224.

[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017.

[23] "Insightface: 2d and 3d face analysis project," https://github.com/deepinsight/insightface.

[24] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[25] P. Lerner, J. Bergoënd, C. Guinaudeau, *et al.*, "Bazinga! a dataset for multi-party dialogues structuring," in *Conference on Language Resources and Evaluation (LREC)*, 2022.

[26] J. Huh, "Simple Diarization," https://github.com/JaesungHuh/SimpleDiarization, 2024.