FRAME: A Fully Automated Pipeline for Scalable Multimodal Body Language Datasets

Anonymous authors
Paper under double-blind review

Abstract

Large-scale multimodal datasets are essential for training representations that capture non-verbal communication, yet most are limited by the cost of manual labeling and remain too small to support large-scale models. We introduce FRAME (Fully-automated Retrieval and Annotation of Motion and Expression), a fully automated pipeline for constructing large-scale multimodal datasets from natural dialogue videos, and use it to release POSTURE (Pose, Speech, and Talking-face Utterances for Representation Evaluation), a dataset of 156,020 movie clips with synchronized subtitles, facial expression frames, and body pose sequences. From segmentation to pose and face extraction, every stage of the pipeline runs without human input, unlike most existing multimodal datasets, which rely on manual annotation and consequently reach only a fraction of this scale. We show that this scale translates directly into better representations. Averaged across a zoo of 15 loss functions and pose encoder architectures, models pretrained on POSTURE achieve 6.8× the cross-dataset transfer R@1 of models pretrained on any other single comparable dataset (CAER, MELD, CMU-MOSEI, DFEW) at full size. This advantage does not come at the cost of quality: at matched dataset size, the advantage falls to near-parity (0.98×), confirming that scale, not source-specific quality, drives POSTURE's advantage. Because construction requires no manual annotation, the same pipeline can be given more source footage to produce proportionally larger datasets at low marginal cost, making POSTURE inherently scalable. We release the dataset, the full extraction pipeline, and evaluation code to support future work on scalable multimodal representation learning. Code for the dataset pipeline can be found through this link: here

1 Introduction

Body language is a critical component of human communication, providing information that spoken or written language alone cannot capture (Knapp et al., 2014). Learning representations that capture this signal across text, facial expression, and body pose takes data at a scale and modality diversity that manual annotation struggles to provide. Existing multimodal datasets are consistently limited by the cost of labeling, and consequently remain too small or too narrow in modality coverage to support this kind of research at scale (Section 2, Table 1).

We address this by constructing FRAME (Fully-automated Retrieval and Annotation of Motion and Expression), a fully automated pipeline in which every stage, from segmentation to pose and face extraction, requires no human input. The resulting dataset POSTURE (Pose, Speech, and Talking-face Utterances for Representation Evaluation) contains 156,020 video clips with synchronized subtitles, facial expression frames, and pose sequences, beyond the scale of manual annotation. As we show in this work, this data pipeline not only produces large-scale, but also high-quality data.

We show that pretrained contrastive encoders on POSTURE evaluated on cross-dataset transfer against three existing multimodal datasets (CAER, MELD, CMU-MOSEI) achieve

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

6.8× the transfer retrieval performance of models pretrained on any single other dataset. When size-matched to each of the other datasets, models trained on our dataset performed comparably, showing that while scale caused a substantive improvement, the data quality was comparable; movie-disjoint and compute-matched controls further confirm that this advantage is not an artifact of shared film identity or training length.

Our contributions are as follows: First, we introduce a fully automated pipeline, FRAME, for constructing large-scale multimodal datasets from natural dialogue video, and release POSTURE, the 156,020-clip dataset it produces, with synchronized text, face, and pose annotations. Second, we show that the scale of this dataset gives massive cross-dataset transfer advantage on a variety of architectures via multiple metrics. Third, despite this scale advantage, we show through scale-matching that the quality of our automated dataset is comparable to other human-annotated data.

2 Related Work

In this section, we review two areas related to our experiments: the datasets available for multimodal and emotion recognition research, and how prior work has trained multimodal systems for this task, from feature fusion to contrastive pretraining that aligns modalities into a shared representation space.

2.1 Multimodal Datasets

Existing multimodal and affective computing datasets are limited along at least one of three variables: temporal structure, modality coverage, or scale (Table 1). EMOTIC pairs facial expression with scene context, but at the level of single images rather than temporal clips (Kosti et al., 2019). CAER extends context-aware emotion recognition to video, but without a pose modality or a distributed transcript (Lee et al., 2019). MELD and CMU-MOSEI instead center text alongside audio and face (Poria et al., 2019; Bagher Zadeh et al., 2018), but neither includes pose, and both are an order of magnitude smaller than POSTURE. DFEW likewise centers face and audio at video scale, again without text or pose (Jiang et al., 2020). BoLD takes the opposite approach, providing pose annotations without any text or face modality (Luo et al., 2020). FERV39k and MAFW are, respectively, the largest and most richly annotated video datasets restricted to the face (and, for MAFW, audio) (Wang et al., 2022; Liu et al., 2022). None of these datasets jointly provide text, face, and pose at clip-level temporal resolution and large scale, and each relies on manual annotation for its labels. POSTURE addresses all three limitations at once, combining all three modalities at 156,020 clips through a construction pipeline that requires no manual labeling at any stage.

2.2 Contrastive Multimodal Pretraining

Contrastive pretraining has become a dominant method for aligning different modalities into a shared representation space, beginning with image-text alignment in CLIP (Radford et al., 2021) and extending to broader modality sets in frameworks such as ImageBind (Girdhar et al., 2023) and DeSPITE, which places skeleton, point-cloud, and inertial data alongside text within a single contrastive space (Kreutz et al., 2025). Within the motion domain, MotionCLIP aligns body motion sequences with text in a CLIP-derived space and reconstructs motion from the aligned latent representation (Tevet et al., 2022), while MoCLIP fine-tunes CLIP's text encoder with a dedicated motion head to improve text-to-motion generation (Maldonado et al., 2025).

Beyond contrastive alignment, prior work in multimodal emotion recognition has largely relied on feature fusion, combining face, body, and text through concatenation or attention-based pooling (Fang et al., 2022). These approaches optimize directly for classification accuracy, training the fusion mechanism and the encoders jointly toward the target labels. Our work instead centers on the quality of the representation itself, evaluated across multiple metrics beyond classification.

Unlike MotionCLIP and MoCLIP, which target generation quality on curated motion capture data, our work targets the representational quality of a frozen embedding across text, face, and pose on natural dialogue, evaluated independently of any downstream generation task.

3 Methodology

This section describes the FRAME pipeline and how we evaluate the dataset it produces. We first detail FRAME's source, segmentation, and modality extraction, which together produce the POSTURE dataset (Sections 3.1–3.2). We then describe how we compare POSTURE against existing multimodal datasets: the comparison datasets (Section 3.3), model training across pretraining recipes (Section 3.4), and cross-dataset retrieval evaluation (Section 3.5).

3.1 Dataset Source and Segmentation

For the dataset evaluated in this work, FRAME is run on raw movie clips sourced from the MovieClips YouTube channel, chosen for their pairing of full-body visibility with variable-length speech dialogue. This contrasts with existing public multimodal datasets, which typically feature seated speakers with limited body visibility, and are often far smaller in scale (Table 1). YouTube additionally provides timestamped captions, enabling large-scale segmentation without manual annotation. We downloaded 12,719 source videos from the channel, each with subtitles and corresponding timestamps, which were subsequently segmented into shorter clips.

We segmented these videos into 156,020 clips (mean duration ∼8s, range 0.5–10s), merging captions within 0.6s (natural pause) of each other and padding each clip by 0.15s to avoid clipping edge words.

3.2 Modality Extraction

While segmenting the clips of the dataset, FRAME extracts the subtitles while simultaneously processing each segmented clip through Google's MediaPipe pose model at 6 frames per second to obtain the pose modality. The output of the pose extraction is 33 landmark joints, each with coordinates and a confidence score. For the face, since our encoder uses MAE-DFER with a projection head, we processed the face images to match its input format by collecting 16 frames per clip, with each frame being a 224×224 pixel image of the face.

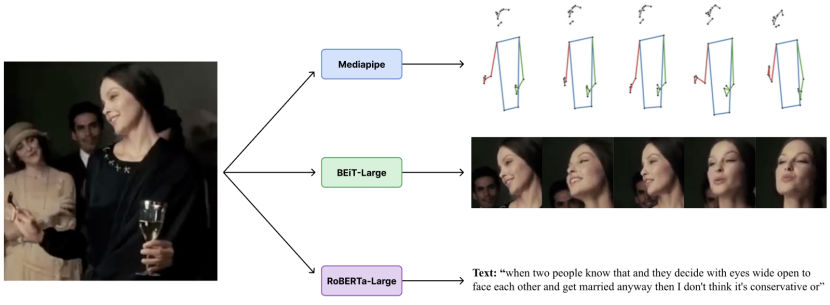


Figure 1: Modality extraction from a single clip

3.3 Baseline Datasets

To test the impact of POSTURE and whether it creates meaningful benefits compared to current standard multimodal body language datasets, we compare POSTURE's performance in contrastive alignment to other similar datasets.

We tested against CAER, MELD, CMU-MOSEI, and DFEW. CAER and MELD are sources from TV shows. Specifically, CAER is taken from 79 different TV shows and MELD is exclusively taken from the TV show *Friends*. CMU-MOSEI is sourced from YouTube videos

Table 1: Comparison of POSTURE with existing multimodal datasets.

Dataset	# Clips	Modalities	Source Domain	Temporal?	Annotation
POSTURE	156,020	Text + Face + Pose	Movies	Yes (clip)	Automated
EMOTIC (Kosti et al., 2019)	23,571 imgs	Face + Scene context	Web / COCO / ADE20K	No (single image)	Manual (AMT)
CAER (Lee et al., 2019)	13,201	Face + Context (video)	79 TV shows	Yes (clip)	Manual (crowd)
MELD (Poria et al., 2019)	13,708 utt.	Text + Audio + Face	*Friends* (1 TV show)	Yes (utterance)	Manual (crowd)
CMU-MOSEI (Bagher Zadeh et al., 2018)	23,453	Text + Audio + Face	YouTube monologues	Yes (segment)	Manual (crowd)
DFEW (Jiang et al., 2020)	16,372	Face + Audio	In-the-wild video	Yes (clip)	Manual (crowd)
BoLD (Luo et al., 2020)	9,876	Pose only	150 films (AVA)	Yes (clip)	Manual (crowd)
FERV39k (Wang et al., 2022)	38,935	Face only	22 scenes, 4 scenarios	Yes (clip)	Manual (30 annotators)
MAFW (Liu et al., 2022)	10,045	Face + Audio	In-the-wild video	Yes (clip)	Manual (crowd)

and DFEW is sourced from movies. We required datasets compared against POSTURE to contain text that semantically describes the video, whether text descriptions or what the person is saying; visible faces; and visible and mobile pose/human body.

The datasets differ in where their text comes from. MELD and CMU-MOSEI use human transcripts, and POSTURE uses the videos' subtitle track. CAER and DFEW contain no text, so we transcribed their audio with Whisper-small.

CMU-MOSEI consists of clips with a single speaker talking to the camera, usually sitting down and with most of the body not visible. DFEW's clips were selected by its creators specifically for clear and visible faces. POSTURE and CMU-MOSEI averaged around 7–8 seconds per clip, while MELD, CAER, and DFEW all averaged 3–3.5 seconds per clip.

POSTURE is the largest corpus by an order of magnitude. It contains 156,020 clips, compared with 13,175 for CAER, 13,708 for MELD, 22,856 for CMU-MOSEI, and 11,697 for DFEW.

Other popular and relevant body language datasets were excluded from experiments for a few reasons. EMOTIC was excluded because it contains still images. IEMOCAP was excluded because it is recorded in a laboratory with ten actors. We selected datasets that are publicly available and cover a wide variety of film footage. BoLD (Luo et al., 2020), which provides body language annotations for movie clips, would be a natural candidate for comparison, but it was unavailable when we ran these experiments. We plan to include it in future releases.

We ran all external datasets through parts of the FRAME pipeline and made sure that each extracted modality was processed the same way.

We used the official train/test splits for dataset evaluation whenever possible, as it is what other experiments utilize and it prevents favoring our dataset. CAER, MELD and CMU-MOSEI provide train, validation and test splits. We trained using each dataset's train split and combined the test and validation splits into a single held-out pool from which, as for every dataset, we draw a 3,500 clip test set. DFEW provides five cross-validation folds, and we use fold 1. POSTURE's held-out set is 4,000 clips.

3.4 Model Training

Each of the models we tested use the same basic architecture with some modifications to the training procedure enumerated below. For each, the text and face encoders are frozen, pretrained models: For text, we use RoBERTa-large, and we use MAE-DFER for faces. We freeze the pretrained models because it isolates the effects of the training data to small projection heads and the pose encoder, so differences cannot be attributed to one dataset fine-tuning an encoder more. Each modality's output is reduced into one vector using attention pooling; That vector passes through a projection head made up of a linear layer, GELU, and a second linear layer. The output of the projection head is a 512-dimensional vector, which is then L2-normalized. The default objective is pairwise InfoNCE, applied separately to each of the three modality pairs and summed, with a fixed temperature of 0.07. Optimization uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-4} and a cosine schedule; batch size 128, 30 epochs, and gradient clipping at 1.0. Every configuration was trained with three seeds: 6767, 1234 and 2024. Training ran in 32-bit floating point on a single Apple M4 Max Mac Studio.

We repeated this experiment with 15 pretraining recipes, chosen as representative published methods from each major family of multimodal contrastive learning. Recipes are not combined: each changes one component of the default setup (InfoNCE, transformer pose encoder) and keeps everything else, including the frozen text and face encoders, fixed.

The recipes cover six types:

- **Pairwise objectives:** InfoNCE, and SigLIP's sigmoid loss, which scores each pair independently rather than against the whole batch.
- **Anchored objectives:** two modalities are aligned only to a third, the anchor: text (LanguageBind) or face (ImageBind).
- **Three-way objectives:** all three embeddings are scored jointly: Symile, GRAM (Gram-matrix volume) and TRIANGLE (triangle area).
- **Fusion architectures:** PoseEmbroider's core fuses modalities into one joint representation; its aligner is the simpler pairwise variant.
- **Alternative pose encoders:** frozen pretrained MotionBERT, a MotionPatches-style ViT over body-part patches, a TMR-style encoder, a JEGAL-style gesture encoder, and the graph-based ST-GCN.
- **Alternative face encoder:** CLIP ViT-B/32 in place of MAE-DFER.

We run our experiments in two parts, once on each full training dataset and then once where each dataset is downsampled to the size of the minimum-sized dataset. In the full size test, every dataset is used as collected, so POSTURE uses all 152,020 clips. This measures the total impact of each dataset, including its size. In the same size test, every training set is cut down to 9,222 clips, the size of CAER's training set which is the smallest. This way, no dataset needs to be enlarged or repeated. CAER's two tests are therefore the same run. Each equal size subset is drawn using a uniform random sample, using the same seed as that run. Everything else is held constant.

3.5 Model Evaluation

For evaluating contrastive encoder alignment, we measured: cross-modal retrieval recall at 1, 5 and 10 (R@1, R@5, R@10), mean rank, mean reciprocal rank, and RankMe (Garrido et al., 2023), which measures how many embedding dimensions a model uses.

To evaluate cross-modal retrieval we take every clip in the held-out set and embed it using the encoders. We then calculate and rank the cosine similarity between the first modality and every single clip in the other modality. R@1 is the fraction of queries whose own clip ranks first. We also record R@5, R@10, mean rank, mean reciprocal rank, and RankMe. The transfer score of a model is the averaged R@1 score over all held-out sets that it was not trained on. We leave out testing a model on the same dataset it was trained on because it measures fit to a dataset rather than generalization, and it avoids the problem of neighboring scenes and actors leaking from the training set. We also report a random-initialization floor, evaluated without training. With a held-out set size of about 3,500, its R@1 score sits at around 0.0003, about the level of guessing. To make results from these experiments more concise and represent POSTURE's advantage, we calculate a ratio between POSTURE's transfer score and the transfer score of the models trained on the other datasets. To average these ratios, we use a geometric mean.

Each recipe is scored on the specific task that it optimizes for. The pairwise, anchored, and fusion recipes were scored based on retrieval between modality pairs. The three-way objectives are evaluated using their own retrieval tasks and scoring functions, since their losses never align individual modality pairs and pairwise retrieval would therefore understate them. Symile is scored on retrieving the pose clip given text and face together. GRAM and TRIANGLE are scored on retrieving the face–pose pair given text.

From each held-out set, we randomly selected 3,500 clips, picked with seed 6767, and used that same set for every model, seed, and recipe, since R@1 depends on both which clips are in the set and how many, with more clips making retrieval harder.

DFEW is an exception to this. Its fold-1 test split only contains 2,341 clips, so we used all of them. Since a smaller set boosts R@1, we do not directly compare R@1 scores tested on 2,341 clips with R@1 scores tested on 3,500 clips. Rather, when DFEW is the test set, we only compare models against each other that are tested on the same set. We can also indirectly compare performance across datasets by using a ratio between other datasets' performance and POSTURE's performance, as an easier test set boosts both equally.

In every dataset, the held-out clips share people or scenes with the training set: MELD's test scenes use the same six actors as its training set, CAER's test set shares the same TV shows, and in POSTURE's original split, every one of the 1,510 films in the held-out set also appeared in training, often through neighboring clips from the same scene. This inflates same-corpus retrieval, as the model can recognize familiar faces or scenes, but it does not affect transfer scores, since a model trained on POSTURE has not seen any data from the other datasets; our conclusions rely on transfer scores, so they are unaffected. To quantify this bias directly, we also release a movie-disjoint split, holding out 58 whole movies (4,037 clips) with no film appearing in both training and test.

3.6 Filtering

We explore two follow up questions: does retrieval keep improving as POSTURE grows and would fewer, less noisy clips outperform more noisy ones? We trained a model on default InfoNCE and another on ImageBind style, the highest performing option on our measure of pairwise retrieval.

We take random subsets of the training split at 20, 40, 60, and 80%, in addition to the full-size and 9,222-clip runs already reported. We then take the same-size subsets again, but this time keeping only the clips with the highest pose confidence rather than sampling randomly. Pose confidence is measured as the mean MediaPipe visibility confidence over a clip's sampled frames, restricted to landmarks 0–24 (legs are excluded, since many films are shot from the waist up), with missed frames counted as zero confidence.

All other settings match the main study, with three seeds. Models are evaluated on the same four held-out sets used elsewhere, which are left unfiltered. POSTURE's held-out set is split at its median confidence (0.577) into two halves of 1,750 clips each; we compare these two halves directly, since they are the only pair of splits at that test size. For each model type, size, and seed, we compute the ratio between the most-confident subset and the randomly selected subset, summarizing these ratios with a geometric mean, and testing them with the Wilcoxon signed-rank and sign tests.

3.7 Compute Control

During training, because every model trains for 30 epochs, a larger dataset also receives more optimization steps. At full size, POSTURE takes 35,610 steps while models trained on the comparison datasets range from 2,160 to 3,810 steps. POSTURE's full size advantage could therefore be the result of longer training time rather than more data. To test this, we retrained CAER, MELD, CMU-MOSEI and DFEW at full size for about as many steps as POSTURE receives, raising their epochs to 495, 457, 280 and 488 respectively (35,560-35,646 steps). The cosine learning rate schedule spans the whole longer run, and all other settings match the main study. We used default InfoNCE with two seeds (6767 and 1234) and computed transfer scores on the same test sets. We then compare POSTURE's advantage over these compute-matched models with its advantage over the standard 30 epoch models in subsection 4.4.

4 Results

4.1 Experimental Setup

We compare POSTURE against three existing multimodal emotion datasets: CAER (Lee et al., 2019), MELD (Poria et al., 2019), and CMU-MOSEI (Bagher Zadeh et al., 2018). For each dataset, we train a contrastive encoder using a pairwise InfoNCE (van den Oord et al.,

Table 2: Cross-dataset transfer matrix, full dataset size. Rows: training dataset (size in clips). Columns: evaluation dataset. Diagonal entries are in-domain R@1; the rightmost column is the transfer score (mean R@1 across all non-diagonal columns).

Trained on	**POSTURE**	**CAER**	**MELD**	**CMU-MOSEI**	**DFEW**	**Transfer**
POSTURE (152,020)	0.1183	0.0411	0.0292	0.0133	0.0580	0.0279 ± 0.0014
CAER (9,222)	0.0060	0.0873	0.0080	0.0037	0.0114	0.0059 ± 0.0002
MELD (9,989)	0.0036	0.0123	0.1323	0.0032	0.0094	0.0064 ± 0.0004
CMU-MOSEI (16,326)	0.0044	0.0055	0.0028	0.0345	0.0070	0.0042 ± 0.0001
DFEW (9,356)	0.0098	0.0186	0.0093	0.0072	0.0498	0.0112 ± 0.0003

Table 3: Cross-dataset transfer matrix, equal dataset size (9,222 clips each). Layout as in Table 2.

Trained on	**POSTURE**	**CAER**	**MELD**	**CMU-MOSEI**	**DFEW**	**Transfer**
POSTURE	0.0251	0.0088	0.0051	0.0042	0.0135	0.0060 ± 0.0003
CAER	0.0060	0.0873	0.0080	0.0037	0.0114	0.0059 ± 0.0002
MELD	0.0037	0.0106	0.1246	0.0027	0.0081	0.0057 ± 0.0004
CMU-MOSEI	0.0034	0.0041	0.0024	0.0282	0.0064	0.0033 ± 0.0001
DFEW	0.0094	0.0171	0.0093	0.0070	0.0471	0.0107 ± 0.0002

2018) over the modalities that dataset provides, then evaluate the resulting embeddings on every dataset in a cross-modal retrieval task, reporting Recall@1 (R@1). We define a model's *transfer score* as its mean R@1 across all datasets other than the one it was trained on, excluding the self-tested result (the diagonal). We define *advantage* as the ratio of POSTURE's transfer score to a comparison dataset's transfer score, computed under matched conditions and averaged across random seeds.

To isolate the effect of dataset scale from other differences between datasets (source, modality composition, annotation method), we repeat this comparison twice: once with each dataset trained at its natural full size, and once with every dataset downsampled to a common size, which is that of the smallest dataset, CAER, at a size of 9,222 clips. To test whether a single choice of contrastive objective and encoder architecture could bias this comparison, we further repeat it across a "recipe zoo" of 15 loss-function and encoder-architecture combinations. These results are averaged across all tests unless otherwise noted.

4.2 Full-Size vs. Equal-Size Transfer

Table 2 reports the complete transfer matrix for the default pairwise contrastive recipe at full dataset size. POSTURE achieves a transfer score of 0.0279 ± 0.0014, exceeding every comparison dataset by a wide margin (CAER: 0.0059 ± 0.0002; MELD: 0.0064 ± 0.0004; CMU-MOSEI: 0.0042 ± 0.0001), despite POSTURE's in-domain R@1 (0.1183) being only modestly higher than CAER's (0.0873) or MELD's (0.1323).

Table 3 repeats this comparison with every dataset downsampled to 9,222 clips. At matched size, POSTURE's advantage largely disappears: its transfer score (0.0060 ± 0.0003) is statistically indistinguishable from CAER's (0.0059 ± 0.0002) and MELD's (0.0057 ± 0.0004), and only modestly above CMU-MOSEI's (0.0033 ± 0.0001).

This pattern, a large POSTURE advantage at full size that collapses under a matched-size control, holds beyond this single recipe. Averaged across the 12 recipes that learned (of the 15 in the zoo), POSTURE achieves an advantage of 6.8× over the best comparison dataset at full size, versus 0.98× at equal size. This confirms that POSTURE's transfer advantage is driven primarily by dataset scale rather than by our particular choice of contrastive objective or encoder architecture.

4.3 Robustness Across Pretraining Recipes

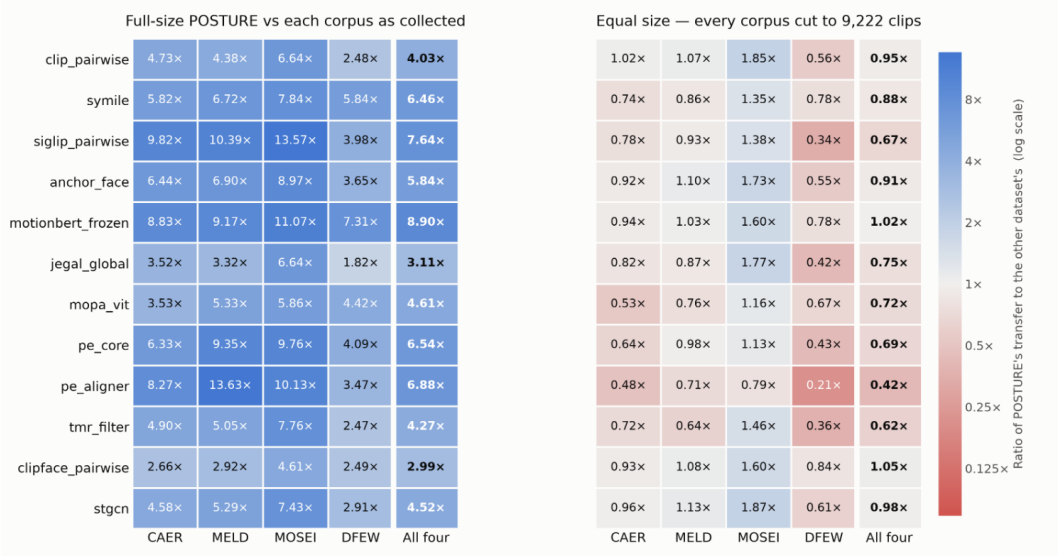


Figure 2: POSTURE's transfer score divided by each comparison dataset's, for the 12 recipes that learned, with every dataset at full size (left) and cut to 9,222 training clips (right). Blue: POSTURE ahead; red: POSTURE behind. Transfer is mean R@1 on the 3,500-clip test sets a model was not trained on, averaged over three seeds. "All four" divides by the mean of the four comparison datasets.

Figure 2 breaks this result down by recipe and comparison dataset. At full size, every cell is blue, which means POSTURE transfers better than every comparison dataset under every recipe, by 1.8× to 13.6×. At equal size, POSTURE's transfer is on par with existing datasets on average, despite being collected without manual curation. POSTURE stays ahead of CMU-MOSEI in all but one recipe, but it falls behind CAER in 11 of 12 recipes and behind DFEW in all 12. Because DFEW is also drawn from films, this suggests that POSTURE's advantage comes from its scale rather than from film footage itself.

4.4 Eliminating Confounding Variables

The transfer advantage established in Section 4.2 could in principle be explained by factors other than dataset scale itself. Namely, shared visual identity between training and evaluation clips drawn from the same films, or simply longer training time afforded by a larger dataset. We address each in turn.

Because POSTURE clips are drawn from a shared pool of source films, a model could achieve inflated transfer performance by exploiting recurring actors, sets, or other visual cues in movies rather than by learning genuine multimodal structure. To rule this out, we construct a movie-disjoint split, in which no film contributing clips to the training set also contributes clips to the evaluation set. Under this split, POSTURE's self-transfer R@1 decreases by 9%, while its transfer score changes by +1% , indicating that POSTURE's cross-dataset transfer advantage is not attributable to identity or set leakage between training and evaluation clips drawn from the same film.

Since POSTURE's larger size also results in a longer training schedule under a fixed number of epochs, we additionally control for total training compute directly, training models on POSTURE and on each comparison dataset for a matched number of gradient steps rather than a matched number of epochs. Under this control, POSTURE's transfer advantage changes by +176%, showing that the advantage established in Section 4.2 is not simply a byproduct of longer training time afforded by a larger dataset.

Taken together, these two controls indicate that POSTURE's transfer advantage reflects the scale and diversity of its training signal, rather than identity leakage across clips from the same film or a training-time confound.

POSTURE's pose tracking is noisier than that of the comparison datasets, so we tested whether fewer, better tracked clips would outperform noisier data. Transfer rises steadily with the number of random POSTURE clips, 4.6× (InfoNCE) and 7.0× (ImageBind-style) from 9,222 to 152,020, with no sign of saturation. Keeping only the clips with the highest pose confidence helps when applied moderately. At 20–80% of the training split, filtered subsets transfer 1.11–1.32× better than random subsets of the same size, and the top 80% outperform the full dataset. Filtering down to 9,222 clips hurts instead (0.75× and 0.85×), because the remaining clips come from far fewer films. Scale therefore remains the main driver, with light quality filtering adding a further gain.

5 Discussion and Conclusion

Because POSTURE requires no manual annotation, it can grow simply by pointing the pipeline at more source footage, at a cost that does not scale with dataset size the way annotation budgets do. At matched size, models pretrained on POSTURE perform comparably to models pretrained on existing, manually annotated datasets, indicating that removing manual annotation from the construction process does not come at the cost of the resulting data's usefulness for representation learning. The advantage POSTURE shows at full size, then, reflects what this quality-preserving automation makes possible: scale that a manually annotated dataset could not reach at the same cost. Two additional controls rule out the most obvious alternative explanations for this advantage: retraining on a movie-disjoint split leaves it essentially unchanged, showing it is not an artifact of shared actors, sets, or scenes between training and evaluation clips, and matching total training compute rather than epoch count shows it is not simply a byproduct of longer training afforded by a larger dataset. Together, these results indicate that FRAME's contribution is not scale for its own sake, but scale achieved without the quality cost automation is usually assumed to carry.

This has a practical implication beyond POSTURE itself. Because construction requires no manual annotation, the same pipeline can be pointed at additional footage to produce proportionally larger datasets at low marginal cost, in contrast to manually annotated datasets, whose sizes are fundamentally capped by annotation budget. Our results suggest that this scalability is not merely a convenience but the primary mechanism by which an automatically constructed dataset can outperform smaller, manually curated alternatives on cross-dataset transfer.

In this paper, we introduced FRAME, a fully automated pipeline for constructing large-scale multimodal datasets from natural dialogue video, and released POSTURE, a 156,020-clip dataset it produced, with synchronized text, face, and pose annotations. We showed that models pretrained on POSTURE achieve a substantial cross-dataset transfer advantage over models pretrained on existing comparison datasets, that this advantage is driven by scale rather than intrinsic per-clip quality, and that it holds under movie-disjoint and compute-matched controls that rule out shared film footage and training-time as alternative explanations.

We see this work as a starting point rather than a complete account of what a dataset like POSTURE can support. Future work may extend the pipeline to additional source footage to test how the transfer advantage scales beyond its current size, apply it to domains beyond movie dialogue, or use POSTURE to study encoder architectures and training objectives for multimodal pose-based representation learning, questions this paper's scope did not allow us to address directly. We release the dataset, the full construction pipeline, and our evaluation code to support this and other future work on scalable multimodal representation learning.

Ethics Statement

POSTURE is constructed from movie clips sourced from a public YouTube channel (MovieClips), and consequently includes the faces, bodies, and voices of real professional actors. These individuals consented to appear in the films from which the clips are drawn, but not specifically to having their expressions extracted, labeled, and distributed as a machine learning dataset. We acknowledge that this is an ethical tension that the footage's public availability does not resolve.

To mitigate this tension, we release processed modalities (pose skeletons, face crops, subtitles), rather than redistributing raw video frames. This is consistent with the copyright status of the footage and limits our redistribution to derived, non-reproducible representations (pose landmarks, embeddings, and labels) rather than the original creative work itself.

We note that this approach does not ensure full anonymity. Because clip identifiers can be traced back to their source film, and because the individuals depicted are usually publicly recognizable actors, re-identification is possible in a way that would not be true of a dataset built from genuinely anonymous subjects. This is an inherent property of using recognizable film footage rather than a limitation we can fully engineer around.

This project did not undergo formal Institutional Review Board (IRB) review. We did not collect new data through direct interaction with human subjects as POSTURE is constructed entirely from pre-existing, publicly available media. We have not sought a formal exemption determination from an IRB, and we note this here rather than assume one applies.

AI Use Statement

We used AI in three main ways during the preparation of this manuscript: reviewing and revising the wording of various sections, retrieving relevant sources and assisting with literature review, and supporting research ideation. While code for experiments was generated by AI, experimental design was not, along with analysis or interpretations of results. Furthermore, it did not generate any of the paper's technical content. All experimental design, implementation, analysis, and conclusions are the authors' own.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1208.

Zhongyu Fang, Aoyun He, Qihui Yu, Baopeng Gao, Weiping Ding, T. Zhang, and Lei Ma. Faf: A novel multimodal emotion recognition approach integrating face, body and text. *arXiv preprint arXiv:2211.15425*, 2022.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. DFEW: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020. doi: 10.1145/3394171.3413620.

Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. *Nonverbal Communication in Human Interaction*. Wadsworth Cengage Learning, 8th edition, 2014.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(11):2755–2766, 2019. doi: 10.1109/TPAMI.2019.2916866.

Thomas Kreutz, Max Mühlhäuser, and Alejandro Sanchez Guinea. Despite: Exploring contrastive deep skeleton-pointcloud-imu-text embeddings for advanced point cloud human activity understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14633–14643, 2025.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10143–10152, 2019.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. MAFW: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022. doi: 10.1145/3503161.3548190.

Yu Luo, Jianbo Ye, Reginald B. Adams, Jia Li, Michelle G. Newman, and James Z. Wang. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision*, 128(1):1–25, 2020. doi: 10.1007/s11263-019-01215-y.

Gabriel Maldonado, Armin Danesh Pazho, Ghazal Alinezhad Noghre, Vinit Katariya, and Hamed Tabkhi. Moclip: Motion-aware fine-tuning and distillation of clip for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision (ECCV)*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20922–20931, 2022.