WHAT CAN WIFI SAY? LANGUAGE-ALIGNED REPRESENTATION LEARNING FOR WIFI SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding everyday activity could support smarter homes, personal assistants, and remote care. Camera-based monitoring provides rich observations but raises privacy concerns, while WiFi offers a privacy-aware alternative. We therefore ask: What can WiFi say in natural language? With language, WiFi could describe movement, retrieve related activities, and answer questions about what happened. We introduce SIGNALLANG, a unified framework that aligns WiFi channel state information (CSI) with motion-focused language. Its captions retain posture, body-part involvement, motion attributes, and temporal structure while excluding unsupported visual details. A transformer over subcarrier-time tokens preserves the link, frequency, and temporal structure of CSI, while contrastive, matching, generative, and auxiliary activity objectives learn a shared WiFi-text representation. Class-aware negative masking prevents recordings of the same activity from being treated as negatives. A single backbone supports caption generation, bidirectional retrieval, and WiFi question answering. On XRF55 activities represented during training, SIGNALLANG improves over an adapted WiFi2Cap baseline by 19.9 BLEU-4 points, 51.8 percentage points in WiFi-to-text retrieval R@1, and 25.3 percentage points in WiFi question answering accuracy. It also shows measurable transfer to held-out activity classes in retrieval and question answering.

1 INTRODUCTION

Understanding human activity is important for smart homes, personal assistants, elderly care, and health and wellness monitoring. Activity monitoring systems often rely on cameras for rich visual information, but continuous video can expose people's identities, appearance, household objects, and private living spaces. WiFi sensing offers a privacy-aware alternative by measuring how human movement changes wireless propagation without capturing images. If these measurements could be linked to language, personal assistants could answer questions about observed activity and daily routines, helping caregivers understand an older adult's mobility. This motivates our central question: *What can WiFi say in natural language?*

Human motion alters wireless propagation paths, producing measurable variations in WiFi channel state information (CSI) that have been used for activity recognition, gesture recognition, localization, and pose estimation (Wang et al., 2015; 2014; Yang et al., 2023; Zheng et al., 2019). Existing WiFi sensing research, however, has primarily focused on task-specific inference from a fixed activity vocabulary rather than explicitly aligning WiFi signals with natural language. A categorical label identifies an activity but does not explicitly describe its body-part movements, motion attributes, or temporal structure. In parallel, vision-language and audio-language models have shown that aligning perceptual inputs with natural language can support generation, retrieval, and reasoning through a shared semantic interface (Radford et al., 2021; Alayrac et al., 2022; Elizalde et al., 2023; Girdhar et al., 2023). Aligning WiFi with language could similarly support free-form motion descriptions, semantic retrieval of signals and text, and natural-language questions about observed activities.

Realizing this potential requires addressing several challenges specific to WiFi-language learning. First, existing public WiFi sensing datasets typically provide only coarse activity labels (Guo et al., 2019; Zheng et al., 2019; Yang et al., 2023; Lan et al., 2025). Although XRF55 (Wang et al., 2024) provides synchronized video that can be used to generate captions with vision-language models (Bai et al., 2025), unconstrained captions often include clothing color, background appearance,

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.

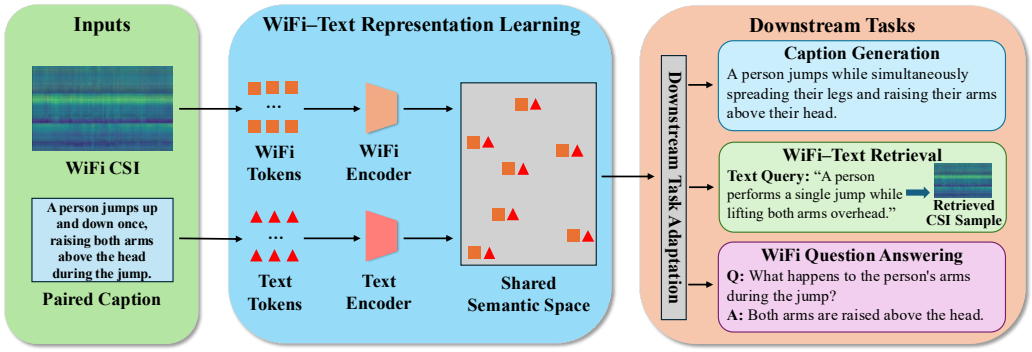

Figure 1: **Overview of SignalLang.** SignalLang learns a unified language-aligned representation from paired WiFi CSI and motion-focused text. The resulting representation supports caption generation, bidirectional WiFi–text retrieval, and WiFi question answering (WiQA) within a common framework. Representative outputs for the same WiFi observation are shown on the right.

camera viewpoint, and other visual details that are observable in video but unsupported by CSI. Training on such captions can encourage a model to reproduce plausible visual details from learned priors rather than ground its outputs in wireless measurements. Effective supervision should instead focus on information shared across the modalities, including posture, body-part involvement, direction, repetition, and temporal order. Second, WiFi-text alignment must account for the physical structure of CSI. CSI contains distinct wireless link, subcarrier, and temporal dimensions that should be preserved in its representation. Third, recordings of the same activity can share similar motion patterns and admit semantically compatible descriptions. Treating every unpaired WiFi-text pair as a contrastive negative can therefore introduce false negatives.

To address these challenges, we introduce SIGNALLANG, a unified framework for language-aligned WiFi representation learning built around three complementary designs, as illustrated in Figure 1. First, we develop motion-focused language supervision that retains motion information supported by CSI while excluding unsupported visual details. Second, we design a structured WiFi encoder that represents local subcarrier-time patches spanning all wireless links as transformer tokens, preserving complementary information across links, subcarriers, and time. Third, we incorporate auxiliary activity supervision and class-aware negative masking into joint alignment and generation training. Together, these mechanisms encourage activity-discriminative WiFi representations without treating unpaired recordings of the same activity as contrastive negatives. The resulting pretrained backbone supports caption generation, bidirectional WiFi-text retrieval, and WiFi question answering (WiQA).

We evaluate SIGNALLANG on XRF55 (Wang et al., 2024) and XRF V2 (Lan et al., 2025). Our main contributions and findings are as follows:

- We introduce SIGNALLANG, a unified framework that combines structure-preserving CSI encoding with activity-aware WiFi-text alignment. The pretrained backbone supports free-form caption generation, bidirectional WiFi-text retrieval, and WiFi question answering.
- We develop motion-focused language supervision and construct 28,759 captions and 103,240 WiQA questions on XRF55. Controlled analyses show reduced reference-unmatched visual-term usage under motion-focused supervision and higher caption-based WiQA accuracy with per-recording captions than with one fixed caption per class.
- On XRF55 activities represented during training, SIGNALLANG improves over an adapted WiFi2Cap baseline by 19.9 BLEU-4 points, 51.8 percentage points in WiFi-to-text retrieval R@1, and 25.3 percentage points in WiQA accuracy. It also shows measurable transfer to held-out activity classes in retrieval and WiQA, while retraining on XRF V2 demonstrates the framework's portability to a different WiFi acquisition setting.

Table 1: Comparison with wireless–language learning approaches. Modality columns indicate sensing inputs directly encoded during model training, including by frozen teacher encoders, but exclude modalities used solely for offline text annotation. Task columns indicate tasks reported in the original papers; ✗ means not reported. SignalLang uses video only to construct text annotations.

Work	Sensing Modalities			Downstream Tasks		
	WiFi	Video	mmWave	Retrieval	Captioning	QA
WiFi2Cap (Wei et al., 2026)	✓	✓	✗	✗	✓	✗
RadarLLM (Lai et al., 2026)	✗	✗	✓	✗	✓	✗
mmExpert (Yan et al., 2025)	✗	✗	✓	✗	✓	✓
mmMind (Zhang et al., 2026)	✗	✗	✓	✗	✓	✓
WirelessSenseLLM (Keya et al., 2026)	✓	✓	✗	✗	✓	✓
SignalLang (Ours)	✓	✗	✗	✓	✓	✓

2 Related Work

2.1 Language-Aligned Representation Learning

A central direction in multimodal learning is to align non-textual inputs with language. CLIP learns transferable visual representations through image–text contrastive learning (Radford et al., 2021). ALBEF aligns modalities before cross-modal fusion (Li et al., 2021). BLIP unifies image–text contrastive learning, matching, and language modeling for both understanding and generation (Li et al., 2022). CoCa combines contrastive and captioning objectives (Yu et al., 2022), while Flamingo conditions a language model on interleaved visual observations (Alayrac et al., 2022). Beyond vision, CLAP aligns audio with text (Elizalde et al., 2023), and ImageBind learns a joint multimodal embedding space (Girdhar et al., 2023). Complementary to cross-modal alignment, masked autoencoding learns representations through masked-input reconstruction (He et al., 2022). However, existing visual and audio formulations are not directly suited to CSI, which couples wireless-link, subcarrier, and temporal structure. Our method adapts BLIP-style multimodal alignment and masked pretraining to a WiFi Transformer operating on subcarrier–time patches across all nine links. Together with auxiliary activity supervision and class-aware negative masking, the framework learns a unified WiFi–language representation for both understanding and generation.

2.2 WiFi and RF-based Human Sensing

WiFi CSI has been widely studied for device-free human sensing, including activity recognition (Wang et al., 2014; 2015) and cross-domain gesture recognition (Zheng et al., 2019). SenseFi provides a benchmark and model library for deep-learning-based WiFi human sensing (Yang et al., 2023), and XRF55 offers a large synchronized multimodal RF–vision dataset covering 55 activities across multiple environments (Wang et al., 2024). More recent work explores broader sensing representations: X-Fi (Chen & Yang, 2025) learns modality-invariant representations across heterogeneous sensors, including WiFi CSI on XRF55, while RFMamba (Zhang et al., 2025b) introduces frequency-aware state-space modeling for RF-based human perception. These works remain focused on conventional sensing tasks rather than aligning wireless signals with natural language.

Recent work introduces language into wireless sensing through LLM-based zero-shot WiFi activity recognition (Zhang et al., 2025a), text-enhanced activity recognition and temporal action localization (Yang et al., 2025), and language-grounded multisensory human sensing and reasoning (Zhou & Yang, 2025). More directly related to language generation, WiFi2Cap transfers semantics from a synchronized vision–language teacher to CSI and uses a prefix-tuned language model for captioning (Wei et al., 2026), while WirelessSenseLLM maps unsegmented CSI to a language-aligned space for zero-shot motion description and reasoning (Keya et al., 2026). Compared with prior WiFi-language approaches, our method learns a unified WiFi–language representation for bidirectional retrieval, caption generation, and WiQA under a common seen/unseen activity protocol.

2.3 Time Series and Language Alignment

Language has been incorporated into time-series and sensor representation learning. Time-LLM, TEST, and ChatTS respectively study forecasting, text-aligned time-series representations, and language-based time-series understanding (Jin et al., 2024; Sun et al., 2024; Xie et al., 2025). A parallel line of work aligns human-centric sensors with language, including IMU2CLIP for inertial signals (Moon et al., 2023), SensorLLM for motion-sensor–text alignment (Li et al., 2025), and SensorLM for large-scale wearable sensor–language pretraining (Zhang et al., 2025c). Our setting differs in that WiFi CSI is an environmental measurement structured across wireless links, subcarriers, and time, and we target motion-description generation and bidirectional retrieval. As summarized in Table 1, SignalLang differs from prior WiFi-language approaches by supporting bidirectional retrieval, caption generation, and WiQA through a unified WiFi-language representation.

3 WiFi–Text Dataset Construction

Existing WiFi sensing datasets typically provide coarse activity labels or human pose annotations (Zheng et al., 2019; Wang et al., 2019; Guo et al., 2019), which offer limited supervision for fine-grained motion semantics. To avoid costly manual collection of such fine-grained motion descriptions, we use synchronized XRF55 videos (Wang et al., 2024) and a pretrained vision-language model (Bai et al., 2025) to generate motion-focused captions for each WiFi sample. The resulting WiFi–Text dataset supports caption generation, retrieval, and question answering.

3.1 Motion-Focused Caption Generation

Visual sensing observes objects through reflected light, whereas WiFi sensing captures how human motion perturbs wireless signal propagation, leading the two modalities to reveal different information. Visual attributes such as color, clothing, facial appearance, object identity, scene layout, and camera viewpoint are visible in video but not directly perceptible from WiFi CSI, making them potentially misleading as language supervision. Therefore, we design a motion-focused captioning scheme that constrains the VLM to use only motion-related information inferable from WiFi signals. The complete caption generation procedure is provided in Appendix B.1.

3.2 WiQA Generation

We construct a WiFi Question Answering (WiQA) dataset to evaluate fine-grained WiFi understanding. For each sample, we design four multiple-choice questions covering action recognition, body-part involvement, motion attributes, and temporal reasoning, each with one correct answer and three distractors. We manually audited the question–answer pairs associated with 500 randomly selected recordings for obvious question-only answer cues. Of these recordings, 482 contained no such cues, whereas 18 contained at least one question whose correct answer could be inferred from the question and answer options alone. For example, asking which body part is used for kicking may reveal "legs" without requiring observation of the motion. See Appendix B.3 for details.

4 Method

We propose SignalLang, a unified framework to learn from WiFi-Text pairs. SignalLang is trained in three stages: (i) WiFi encoder pretraining through masked reconstruction, (ii) WiFi-text alignment in a shared semantic space, and (iii) task-specific fine-tuning for caption generation, retrieval, and QA. Figure 2 shows the model architecture. We first describe the structure of the WiFi CSI input and its tokenization before introducing the pretraining and multimodal alignment objectives.

4.1 WiFi CSI Representation and Tokenization

Channel State Information. WiFi commonly uses orthogonal frequency-division multiplexing (OFDM) over multiple orthogonal subcarriers. For subcarrier s, the received signal is modeled as

$$Y_s = H_s X_s + n_s, \tag{1}$$

where X_s and Y_s are the transmitted and received signals, n_s is noise, and H_s is the channel state information (CSI) characterizing the wireless channel on subcarrier s. CSI is complex-valued and captures propagation-induced attenuation and phase shifts.

Indoor WiFi propagation is inherently multipath, with signals reaching the receiver through direct and reflected paths from surrounding objects and the human body. Human motion alters path lengths, attenuation, and relative phases, producing CSI variations across subcarriers, time, and TX-RX links; SignalLang preserves this structure when constructing transformer tokens.

XRF55 configuration and CSI representation. We use XRF55 (Wang et al., 2024), which contains synchronized WiFi CSI and RGB video for 55 indoor activities across four scenes. XRF55 collects CSI with ThinkPad X201 laptops equipped with Intel 5300 network interface cards, using one transmitter with a single active antenna and three receivers; the transmitter operates at 5.64 GHz and broadcasts 200 packets per second over 30 OFDM subcarriers. Each subcarrier therefore provides the channel response at a distinct frequency within the operating bandwidth. Each receiver is equipped with three receive antennas, resulting in 9 TX-RX links in total. Each activity sample spans 5 s, yielding 1,000 temporal measurements. The raw CSI tensor therefore has dimensions $1 \times 3 \times 3 \times 30 \times 1000$, corresponding to transmit antenna, receiver, receive antenna, subcarrier frequency, and time. We combine the three receivers and their three receive antennas into nine ordered wireless links and retain the CSI amplitude, producing the processed input $X_i \in \mathbb{R}^{9\times30\times1000}$.

We use a fixed split of 45 seen and 10 unseen activity classes. All model-training stages, including masked WiFi encoder pretraining, WiFi-text alignment, and task-specific fine-tuning, use only the seen activity classes; the unseen classes are reserved for evaluation. The complete seen and unseen activity lists are provided in Appendix B.2.

Motion-aware CSI tokenization. Given $\mathbf{X}_i \in \mathbb{R}^{9\times30\times1000}$, we construct tokens that preserve the wireless-link, subcarrier, and temporal dimensions. We retain the 9 wireless links as ordered input channels and partition the 30×1000 subcarrier-time plane into non-overlapping 10×100 patches. This yields 30 patches per sample; each spans all 9 links, 10 neighboring subcarriers, and 100 consecutive temporal measurements and is projected into a 256-dimensional transformer token. At 200 packets/s, each token summarizes 0.5 s of CSI evolution over a local group of neighboring frequencies, while the full 5 s activity is represented by 30 frequency-time tokens.

This motion-aware design captures complementary motion cues across time, frequency, and wireless links. Temporally, each token captures short-term channel evolution, while successive patches preserve how the motion unfolds over time. Across frequency, human motion affects subcarriers differently under frequency-selective multipath propagation, so grouping neighboring subcarriers captures local frequency-dependent motion patterns without collapsing subcarrier structure. Across TX-RX links, distinct multipath channels provide complementary spatial views of the same motion. The resulting tokens preserve local frequency–time motion structure while allowing the transformer to model dependencies across frequency groups, TX-RX links, and different stages of an activity.

4.2 SignalLang Architecture

WiFi Encoder Pretraining. We use masked reconstruction pretraining to initialize the WiFi encoder before WiFi-text alignment. Given a WiFi CSI sample $X_i \in \mathbb{R}^{A\times S\times T}$, where A, S, and T denote the numbers of wireless links, subcarriers, and time steps, respectively ($A = 9$ for XRF55 (Wang et al., 2024)), we divide it into non-overlapping patches and randomly mask 50% of them. The remaining visible patches are fed into the WiFi encoder, and a transformer decoder reconstructs the masked patches by minimizing mean squared error between the reconstructed and original masked patches. The pretrained WiFi encoder is subsequently used to initialize the encoder for WiFi-text alignment.

WiFi-Text Alignment. To align WiFi and text representations in a shared feature space, we jointly optimize four objectives during training. Given a training tuple $(\mathbf{X}_i, \mathbf{C}_i, a_i)$, where $\mathbf{C}_i$ denotes the caption associated with $\mathbf{X}_i$ and a_i denotes its activity label, we obtain the WiFi and text representations as

$$\mathbf{w}_i = \mathrm{Norm}(g_w(\mathrm{Pool}(E_w(\mathbf{X}_i)))), \qquad \mathbf{t}_i = \mathrm{Norm}(g_t(E_t(\mathbf{C}_i))), \tag{2}$$

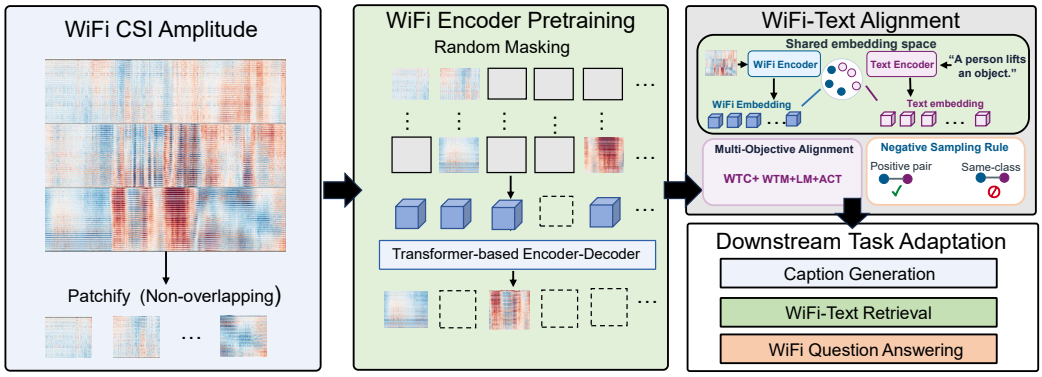

Figure 2: SIGNALLANG architecture. The WiFi encoder produces 30 subcarrier-time tokens. WTC aligns pooled WiFi and text representations with class-aware negative masking, while WTM and LM attend to the full WiFi token sequence. ACT provides auxiliary activity supervision. Same-class samples are excluded from WTC and WTM negatives.

where $g_w(\cdot)$ and $g_t(\cdot)$ are projection layers that map the WiFi and text representations into the same embedding space. $\mathrm{Norm}(\cdot)$ and $\mathrm{Pool}(\cdot)$ denote ℓ_2 normalization and the pooling operation, respectively. The four objective losses are shown below.

(1) WiFi-Text Contrastive Learning (WTC). Following Li et al. (2021), we use WTC to learn modality-invariant WiFi-text representations by computing bidirectional similarity distributions between WiFi and text embeddings, encouraging paired samples to be closer than unpaired samples in the shared embedding space. The WTC loss is defined as

$$\mathcal{L}_{\mathrm{WTC}} = \frac{1}{2}\mathbb{E}_i \left[\mathrm{H}\big(\mathbf{q}_i^{w\to t}, \mathbf{p}_i^{w\to t}\big) + \mathrm{H}\big(\mathbf{q}_i^{t\to w}, \mathbf{p}_i^{t\to w}\big) \right], \tag{3}$$

where $\mathrm{H}(\mathbf{q}, \mathbf{p})$ denotes the cross-entropy between the target distribution $\mathbf{q}$ and the predicted distribution $\mathbf{p}$. $\mathbf{p}_i^{w\to t}$ and $\mathbf{p}_i^{t\to w}$ denote the online WiFi-to-text and text-to-WiFi similarity distributions, respectively. $\mathbf{q}_i^{w\to t}$ and $\mathbf{q}_i^{t\to w}$ are the corresponding soft pseudo-target distributions generated by the momentum encoders. Different WiFi recordings of the same activity can share similar motion semantics, and their captions may remain compatible even when they are not explicitly paired. Treating all unpaired samples as negatives can therefore push semantically related WiFi-text pairs apart. To alleviate this issue, we introduce class-aware negative masking. For each anchor sample in a batch, unpaired samples with the same activity label are excluded from the negative set, while only the explicitly paired WiFi-text sample is retained as the positive. The remaining unpaired samples with different activity labels are treated as negatives.

(2) WiFi-Text Matching (WTM). To learn multimodal interactions between WiFi and text, we adopt the image-text matching objective of (Li et al., 2021) as WiFi-Text Matching (WTM). Specifically, given a WiFi-text pair $(\mathbf{X}_i, \mathbf{C}_j)$, caption tokens interact with the full WiFi token sequence through cross-attention, and the fused representation is used to predict whether the pair is matched or mismatched. WTM is optimized with a binary cross-entropy loss as follows:

$$\mathcal{L}_{\mathrm{WTM}} = \mathbb{E}_{(i,j)} \left[\mathrm{BCE}(m_{ij}, \hat{m}_{ij}) \right], \tag{4}$$

where $m_{ij} \in \{0, 1\}$ denotes the ground-truth matching label and $\hat{m}_{ij}$ denotes the predicted matching probability. We perform hard negative sampling based on WTC similarity scores (Li et al., 2021), while excluding unpaired samples with the same activity label from the negative sampling pool to avoid ambiguous negatives.

(3) WiFi-Conditioned Language Modeling (LM). To generate textual descriptions given a WiFi sample, we employ a WiFi-conditioned language modeling objective, where the language decoder autoregressively generates the paired caption by attending to the WiFi representation. This objective encourages the WiFi encoder to preserve motion-related information.

$$\mathcal{L}_{\mathrm{LM}} = -\sum_{\ell=1}^{L_i} \log P_\theta \left(c_{i,\ell} \mid c_{i,<\ell}, \mathbf{Z}_i^w \right), \tag{5}$$

where $c_{i,\ell}$ denotes the ℓ-th token of the caption, $c_{i,<\ell}$ denotes the preceding caption tokens, and $\mathbf{Z}_i^w$ represents the encoded WiFi token sequence. The loss is averaged over all non-padding tokens.

(4) Auxiliary Activity Classification (ACT). While WTC, WTM, and LM encourage cross-modal alignment and language generation, they do not explicitly constrain the WiFi representation to preserve activity-discriminative information. Following prior multimodal learning approaches that complement cross-modal objectives with modality-specific representation supervision (Zhu & Yang, 2020; Mu et al., 2022), we introduce an auxiliary WiFi-only activity classification objective:

$$\mathcal{L}_{\mathrm{ACT}} = \mathbb{E}_i \left[\mathrm{CE} \left(a_i, h_{\mathrm{act}} \left(\mathrm{Pool}(\mathbf{Z}_i^w) \right) \right) \right], \tag{6}$$

where a_i is the activity label and $h_{\mathrm{act}}(\cdot)$ is a linear classifier. Importantly, this objective operates only on the WiFi representation, without access to caption tokens or fused cross-modal features, encouraging the encoder to retain activity-discriminative motion information during alignment. After WiFi-text alignment, the activity classification head is discarded for downstream tasks. The complete objective is

$$\mathcal{L}_{\mathrm{Align}} = \mathcal{L}_{\mathrm{WTC}} + \mathcal{L}_{\mathrm{WTM}} + \mathcal{L}_{\mathrm{LM}} + \mathcal{L}_{\mathrm{ACT}}. \tag{7}$$

4.3 Downstream Task Adaptation

We adapt the aligned WiFi-text model to three downstream tasks: caption generation, WiFi-text retrieval, and WiFi question answering.

Caption Generation. We fine-tune the aligned WiFi encoder and pretrained language decoder end to end using the language-modeling objective in Eq. (5). The decoder autoregressively generates a natural-language description conditioned on the WiFi representation, with beam search at inference.

WiFi-Text Retrieval. Retrieval requires no additional fine-tuning. Following BLIP (Li et al., 2022), we first rank all candidates using the cosine similarity between the WiFi and text representations. The top-k candidates are then reranked using the matching probability p^{match} predicted by the WTM head. We evaluate both WiFi-to-text retrieval (Wi2TR) and text-to-WiFi retrieval (T2WiR).

WiFi Question Answering. We formulate WiQA as answer generation, following the generative formulation of VQA (Li et al., 2021). Given a WiFi sample and a question, the cross-modal text encoder attends to the WiFi tokens to produce a WiFi-grounded question representation $\mathbf{Q}_i$. An answer decoder initialized from the pretrained language decoder is then fine-tuned using the conditional language-modeling objective

$$\mathcal{L}_{\mathrm{WiQA}} = -\sum_{\ell=1}^{L_i^{\mathrm{ans}}} \log P_\theta \left(y_{i,\ell} \mid y_{i,<\ell}, \mathbf{Q}_i \right), \tag{8}$$

where $y_{i,\ell}$ denotes the ℓ-th token of the target answer. At inference, the decoder scores each candidate answer by its length-normalized log-likelihood and selects the highest-scoring candidate.

5 Evaluation

5.1 Experimental Setup

We evaluate SignalLang under the XRF55 configuration and activity split described in Section 4.1. We use a deterministic repetition-level split: for each subject-activity pair, repetitions 1 to 12 are assigned to training, repetitions 13 to 14 to validation, and repetitions 15 to 20 to testing. Subjects and scenes may therefore appear across partitions. Of the 55 activities, 45 are designated as seen, yielding 16,176 training, 2,697 validation, and 8,087 test samples. The remaining 10 activities are excluded from training and validation and used only for testing, contributing 1,799 unseen test samples. Caption generation and retrieval are therefore evaluated on 9,886 test samples in total, consisting of 8,087 seen and 1,799 unseen samples. For WiQA, the training, validation, and test sets contain 58,008, 9,688, and 35,544 questions, respectively, for a total of 103,240 questions.

Baselines. We compare SignalLang with WiFi2Cap*, an amplitude-only adaptation of WiFi2Cap (Wei et al., 2026), using the same CSI amplitude tensors and data splits. Because the processed XRF55 arrays used in our experiments contain amplitude only, we retain its amplitude

Table 2: Caption generation and activity-level WiFi–text retrieval on XRF55. Retrieval R@K counts a query as correct when at least one of the top-K candidates shares its activity label; recovering the exact paired recording is not required. B@4 denotes BLEU-4. LLM Cls.@1 denotes LLM-based activity classification accuracy, where the activity label is predicted from the generated caption only. Caption metrics use the multi-reference protocol in Appendix A.2. Wi2TR denotes WiFi-to-text retrieval, and T2WiR denotes text-to-WiFi retrieval. WiFi2Cap* is the amplitude-only XRF55 adaptation described in Section 5.1.

Activity Set	Method	Caption Generation					Retrieval					
		B@4	METEOR	ROUGE-L	CIDEr	LLM Cls.@1	Wi2TR@1	Wi2TR@5	Wi2TR@10	T2WiR@1	T2WiR@5	T2WiR@10
Seen	WiFi2Cap*	23.9	39.8	41.2	42.6	15.9	19.7	28.6	33.4	28.0	62.3	76.9
	SignalLang	**43.8**	**61.7**	**58.8**	**129.8**	**60.6**	**71.5**	**81.1**	**84.2**	**89.8**	**96.2**	**97.2**
Unseen	WiFi2Cap*	**13.4**	32.1	35.1	12.3	21.3	27.7	45.7	51.5	25.4	60.7	71.1
	SignalLang	11.5	**35.0**	**36.5**	**14.1**	**24.8**	**32.7**	**54.6**	**64.8**	**35.6**	**62.8**	**76.3**

encoder and remove the phase encoder and amplitude–phase fusion module. This comparison evaluates the two approaches under matched amplitude-only inputs rather than comparing SignalLang with the full amplitude–phase version of WiFi2Cap. We additionally adapt WiFi2Cap* for retrieval and WiQA evaluation.

Evaluation Metrics. We use task-specific metrics to evaluate SignalLang. For caption generation, we employ BLEU-4 (B@4), METEOR, ROUGE-L, and CIDEr to measure the similarity between generated and reference captions. We further report an LLM-based activity identification accuracy (LLM), where Qwen3.6-35B predicts the underlying activity using only the generated caption. For WiFi–text retrieval, we report Recall@K ($K \in \{1, 5, 10\}$) for both WiFi-to-text retrieval (Wi2TR) and text-to-WiFi retrieval (T2WiR). For WiFi question answering, we report multiple-choice accuracy for each of the four question categories.

5.2 WiFi Captioning

As shown in Table 2, SignalLang substantially outperforms WiFi2Cap* on all captioning metrics for seen activities, improving BLEU-4 and CIDEr by 19.9 and 87.2, respectively, and increasing LLM-based activity classification accuracy from 15.9% to 60.6%. On unseen activities, SignalLang achieves higher METEOR, ROUGE-L, CIDEr, and LLM accuracy, although BLEU-4 is lower and the improvements are more modest. Qualitative comparisons in Figure 3 and Table 5 of the appendix illustrate better recovery of body-part movements and motion sequences, with less reliance on visually specific assumptions. Figure 3 also highlights remaining errors in activity identity and motion direction, indicating that fine-grained motion interpretation remains challenging.

5.3 WiFi–Text Retrieval

We evaluate activity-level WiFi–text retrieval, where a retrieved item is relevant if it shares the query's activity label. This protocol measures cross-modal association at the activity level rather than recovery of the exact paired recording. Under this criterion, SignalLang improves seen WiFi-to-text and text-to-WiFi R@1 from 19.7% and 28.0% to 71.5% and 89.8%, respectively. On unseen activities, the corresponding scores improve from 27.7% and 25.4% to 32.7% and 35.6%. These results indicate improved activity-level cross-modal association on both activity sets, but do not establish instance-specific discrimination among recordings of the same activity.

5.4 WiFi Question Answering

As shown in Table 3, SignalLang consistently improves WiQA performance on seen activities, increasing the overall accuracy from 55.5% to 80.8%. In particular, action accuracy improves from 57.2% to 92.6%, while motion-attribute accuracy increases from 36.0% to 71.3%. On unseen activities, overall accuracy improves from 39.5% to 46.2%. It outperforms WiFi2Cap* on action, body-part, and motion-attribute questions, while remaining comparable on temporal reasoning (63.0% vs. 65.4%). These results show that SignalLang can answer questions about activity identity, body-part involvement, motion attributes, and temporal structure from WiFi CSI.

Table 3: WiQA accuracy (%) on seen and unseen activities. Overall denotes the average accuracy across all samples.

Activity Set	Method	Action	Body Part	Motion Attr.	Temporal	Overall
Seen	WiFi2Cap*	57.2	70.7	36.0	57.9	55.5
	SignalLang	92.6	83.5	71.3	76.1	80.8
Unseen	WiFi2Cap*	25.0	45.0	24.4	65.4	39.5
	SignalLang	31.7	59.9	30.1	63.0	46.2

Table 4: Ablation of the main SIGNALLANG components on seen and unseen activities. Retrieval reports activity-level R@1 for Wi2TR and T2WiR retrieval, using the same relevance criterion as Table 2. LLM Cls.@1 denotes LLM-based activity classification accuracy from generated captions.

Variant	Activity Set	Wi2TR R@1	T2WiR R@1	LLM Cls.@1
w/o class-aware negative masking	Seen	49.9	57.1	58.1
	Unseen	23.7	27.2	**26.9**
w/o auxiliary activity loss	Seen	40.0	61.4	45.7
	Unseen	29.1	29.8	22.5
w/o MAE initialization	Seen	64.9	85.2	57.3
	Unseen	28.8	32.6	25.9
SignalLang (full)	Seen	**71.5**	**89.8**	**60.6**
	Unseen	**32.7**	**35.6**	24.8

5.5 ABLATION STUDY AND GENERALIZATION ANALYSIS

Class-aware negatives and activity supervision. Table 4 shows that removing class-aware negative masking sharply reduces seen retrieval, with Wi2TR/T2WiR R@1 dropping from 71.5/89.8% to 49.9/57.1%, while LLM Cls.@1 changes only slightly. Removing auxiliary activity supervision reduces retrieval to 40.0/61.4% and LLM Cls.@1 to 45.7%. Under the activity-level retrieval criterion, class-aware negative masking improves cross-modal association, while auxiliary activity supervision also improves caption-based activity identification on seen activities. These results do not by themselves establish improved discrimination between instances of the same activity.

MAE initialization. MAE initialization improves seen LLM Cls.@1 from 57.3% to 60.6% and unseen Wi2TR/T2WiR R@1 from 28.8/32.6% to 32.7/35.6%, supporting masked reconstruction pretraining for WiFi-text alignment.

Motion-focused supervision. Appendix F reports lexicon-based hallucination proxy rates of 74.3/77.1% under unrestricted supervision and 3.4/4.5% under motion-focused supervision on seen/unseen activities. Training with per-recording captions also improves caption-based WiQA over training with one canonical caption per activity class, achieving 67.3/43.5% versus 58.3/35.3%. These comparisons show reduced reference-unmatched visual-term usage and improved caption-based question answering under the evaluated supervision settings.

Cross-dataset portability. Retraining SignalLang on XRF V2 with the same architecture and training procedure achieves 41.5 BLEU-4, 45.4/70.9% WiFi-to-text/text-to-WiFi R@1, and 78.3% WiQA accuracy. These results demonstrate framework-level portability after retraining, whereas poor zero-shot performance indicates limited transfer of the XRF55-trained checkpoint. Further analysis is provided in Appendix E.

6 CONCLUSION

We introduced SIGNALLANG, a unified framework for aligning WiFi CSI with motion-focused language. A single backbone supports caption generation, bidirectional retrieval, and question answering while discouraging visual details unsupported by the signal. Experiments on XRF55 demonstrate strong performance for activities represented during training and measurable transfer to held-out activity classes in retrieval and question answering. Retraining on XRF V2 further demonstrates the framework's portability to a different WiFi acquisition setting. These findings establish language alignment as a promising interface for WiFi sensing and motivate further work on faithful generation and robust generalization.

AI USE STATEMENT

Generative AI tools were used in several parts of this work. As described in the paper, we used large language and vision-language models to construct motion-focused captions and WiFi question-answering annotations from synchronized videos, and LLMs were used as part of the caption evaluation protocol. Generative AI tools were also used for proofreading the manuscript. All research decisions, experimental designs, implementations, analyses, and reported conclusions were reviewed and verified by the authors. AI-generated annotations were subject to the validation procedures described in the paper, including rule-based or LLM-based checks and human auditing where specified. The authors take full responsibility for the final content of the paper and for all claims and results reported herein.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *International Conference on Learning Representations*, volume 2025, pp. 97378–97397, 2025.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 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*, pp. 15180–15190, 2023.

Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo. Wiar: A public dataset for wifi-based activity recognition. *IEEE Access*, 7:154935–154945, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Mahmuda Keya, Sneh Pillai, Jiawei Yuan, Kai Zeng, and Long Jiao. WirelessSenseLLM: Zero-shot human activity understanding by bridging wireless signals and human language. In *IEEE International Conference on Sensing, Communication, and Networking*, 2026.

Zengyuan Lai, Jiarui Yang, Songpengcheng Xia, Lizhou Lin, Lan Sun, Renwen Wang, Jianran Liu, Qi Wu, and Ling Pei. Radarllm: Empowering large language models to understand human motion from millimeter-wave point cloud sequence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5791–5799, 2026.

Bo Lan, Pei Li, Jiaxi Yin, Yunpeng Song, Ge Wang, Han Ding, Jinsong Han, and Fei Wang. Xrf v2: A dataset for action summarization with wi-fi signals, and imus in phones, watches, earbuds, and glasses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–41, 2025.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12888–12900, 2022.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. SensorLLM: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 354–379, 2025.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13246–13253, 2023.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

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*, pp. 8748–8763, 2021.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. TEST: Text prototype aligned embedding to activate llm's ability for time series. In *International Conference on Learning Representations*, 2024.

Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-wifi: Fine-grained person perception using wifi. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5451–5460. IEEE, 2019.

Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. XRF55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1), 2024.

Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pp. 65–76, 2015.

Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, pp. 617–628, 2014.

Tzu-Ti Wei, Chu-Yu Huang, Yu-Chee Tseng, and Jen-Jee Chen. Wifi2cap: Semantic action captioning from wi-fi csi via limb-level semantic alignment. *arXiv preprint arXiv:2603.22690*, 2026.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. ChatTS: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 18, 2025.

Yifan Yan, Shuai Yang, Xiuzhen Guo, Xiangguang Wang, Wei Chow, Yuanchao Shu, and Shibo He. mmexpert: Integrating large language models for comprehensive mmwave data synthesis and understanding. In *Proceedings of the Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 1–10, 2025.

Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing. *Patterns*, 4 (3):100703, 2023.

Zhenkui Yang, Zeyi Huang, Ge Wang, Han Ding, Tony Xiao Han, and Fei Wang. Talk is not always cheap: Promoting wireless sensing models with text prompts. *arXiv preprint arXiv:2504.14621*, 2025.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Duo Zhang, Zhehui Yin, Zhiyun Yao, Haotong Qin, Xusheng Zhang, Hongliu Yang, Jianyu Sun, Junzhe Wang, Zizhou Fan, Michele Magno, et al. Teaching foundation models to read mmwave: Pose-guided kinematic representation for human behavior understanding. *arXiv preprint arXiv:2608.04127*, 2026.

Haopeng Zhang, Yili Ren, Haohan Yuan, Jingzhe Zhang, and Yitong Shen. Wi-chat: Large language model powered wi-fi sensing. *arXiv preprint arXiv:2502.12421*, 2025a.

Rui Zhang, Ruixu Geng, Yadong Li, Ruiyuan Song, Hanqin Gong, Dongheng Zhang, Yang Hu, and Yan Chen. Rfmamba: Frequency-aware state space model for rf-based human-centric perception. In *International Conference on Learning Representations*, volume 2025, pp. 11453–11469, 2025b.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025c.

Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 313–325, 2019.

Chuhao Zhou and Jianfei Yang. HoloLLM: Multisensory foundation model for language-grounded human sensing and reasoning. In *Advances in Neural Information Processing Systems*, 2025.

Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8743–8752. IEEE, 2020.

A Additional Experimental Details

A.1 Implementation Details

The WiFi encoder embeds non-overlapping 10×100 subcarrier–time patches from the nine-link CSI tensor into 256-dimensional tokens, producing 30 tokens per sample, and processes them with eight Transformer blocks. MAE pretraining masks 50% of the patches and reconstructs them with a lightweight decoder for 300 epochs using batch size 1,024 and 100 warmup epochs. Only the encoder is transferred to alignment.

The alignment stage jointly optimizes WTC, WTM, LM, and ACT with equal weights. Class-balanced batches contain 11 activity classes and 5 samples per class, for a total batch size of 55. The momentum coefficient is 0.99, the feature queue contains 2,750 entries, and the distillation weight $\alpha = 0.4$ is linearly warmed up during the first epoch. Same-class candidates are masked from the contrastive denominator, the momentum targets, and the WTM hard-negative pool. The pretrained BERT modules use a base learning rate of 1×10^{-5} with cosine decay. The WiFi-side projection layers and the activity head use 1×10^{-3}, and weight decay is 5×10^{-4}.

For captioning, the WiFi encoder and language decoder are initialized from the alignment checkpoint and fine-tuned with LM only. We use batch size 64, a decoder learning rate of 1×10^{-4}, an encoder

rate scaled by 0.1, and 30 epochs. Captions are decoded with beam size 3 and a length range of 8–40 tokens. WiQA is fine-tuned for 10 epochs with the same learning-rate schedule and evaluated by length-normalized log-likelihood over the candidate options. All experiments use PyTorch 2.6 on one NVIDIA A100 GPU with 80 GB memory.

For WiFi2Cap*, we retain the amplitude encoder and remove the phase encoder and gated amplitude–phase fusion because the processed XRF55 arrays used in our experiments contain amplitude only. The remaining teacher alignment and prefix-guided generation stages follow the published architecture and training protocol. Its training set contains only the 45 seen activities, and the same checkpoint is evaluated on both activity sets.

A.2 Evaluation Protocol

Retrieval. Every test caption and CSI sample is included in the gallery. A retrieved item is relevant when it belongs to the query activity, and we report Recall@K. Wi2TR uses a WiFi query and a caption gallery, whereas T2WiR uses a text query and a WiFi gallery. Chance R@1 is 2.2% for the 45 seen activities and 10.0% for the 10 unseen activities.

Captioning. Lexical quality is measured with BLEU-4, METEOR, ROUGE-L, and CIDEr. The multi-reference protocol uses up to 10 captions: the sample's paired caption and frequent captions from the same activity class. Because most captions are unique paraphrases, the multi-reference protocol avoids penalizing a valid motion description solely for using different wording. LLM Cls.@1 measures activity-identification accuracy from generated captions using the LLM-judge protocol described below.

LLM judge. The primary judge is Qwen3.6-35B. For each generated caption, the judge receives the caption and the candidate activity labels for the corresponding evaluation set. The candidate set contains 45 labels for seen evaluation and 10 labels for unseen evaluation. The judge receives no CSI, video frame, reference caption, or ground-truth label. We use constrained decoding at temperature 0 and require one candidate label as the complete response. A prediction is counted as correct when the selected label matches the sample's activity label. The prompt is:

> *You are evaluating a motion caption generated from a WiFi signal. Select the single activity label that is best described by the caption. Choose only from the candidate list and return exactly one label with no explanation. Candidate activities: [candidate list].*
> *Caption: [generated caption].*

WiQA. For each question, the decoder selects the option with the highest length-normalized log-likelihood. We report accuracy by question type and over the retained set.

B WiFi–Text Dataset Construction Details

B.1 Representative Caption Prompt

The following template summarizes the prompt used to generate motion-focused captions. The prompt is paired with 16 uniformly sampled frames and the activity label.

> **Task.** Given the video frames and activity label `[ACTIVITY]`, write exactly one English sentence of 12–22 words that describes only the human motion visible in the clip.
> **Content.** Use "a person" or "two people" as the subject. Describe posture, active body parts, motion direction, displacement, contact, repetition, and temporal order when they are visible. For an object-mediated activity, describe the body motion caused by the interaction rather than naming the object.
> **Exclusions.** Do not mention gender, age, identity, appearance, clothing, color, object identity, background, scene, camera viewpoint, sound, or any detail that WiFi sensing cannot infer.
> **Output.** Return only the caption, without a prefix, explanation, or list.

The caption generator uses temperature 0.7. The activity label is used to disambiguate the intended action during data construction; it is not provided to the captioning model at test time.

Generation and decoding. For each recording, Qwen3.6-35B receives 16 uniformly sampled video frames, the corresponding activity label, and the motion-focused instruction above. The activity label is used only to disambiguate the intended action during offline data construction and is not provided to the WiFi captioning model at inference time. Caption generation uses a temperature of 0.7.

Rule-based validation and regeneration. Each candidate caption is processed by a deterministic rule-based validator. The validator checks whether the caption (i) satisfies the required word-count range, (ii) follows the prescribed subject format, (iii) contains prohibited terms related to appearance or object identity, and (iv) triggers the object-leak detection rules. The validator returns either a pass decision or the detected violation reason.

When a caption fails validation, the violation reason is added to the regeneration instruction and returned to the same Qwen3.6-35B model. The model then generates a new sentence from scratch; the pipeline does not edit individual words or spans in the previous caption. The generate–validate–regenerate loop is capped at five rounds. A caption that still violates any rule after the fifth round is discarded from the final caption set.

The validator operates only on the generated text. It does not observe the video, use a second LLM, or perform open-ended semantic verification of whether the caption is visually correct. Its role is to enforce the predefined formatting and WiFi-perceivability constraints.

B.2 Activity Lists

The following lists use the activity-name mapping in our caption-generation and experiment configuration.

Seen activities (45). Carrying Weight; Mopping the Floor; Cutting; Wearing Hat; Using a Phone; Throw Something; Put Something on the Table; Put on Clothing; Picking; Smoking; Eating; Brushing Teeth; Blow Dry Hair; Brush Hair; Shake Hands; Hugging; Hand Something to Someone; Kick Someone; Hit Someone with Something; Choke Someone's Neck; Push Someone; Body Weight Squats; Tai Chi; Boxing; Weightlifting; Hula Hooping; Jump Rope; Jumping Jack; High Leg Lift; Waving; Clap Hands; Fall on the Floor; Running; Standing Up; Stretch Oneself; Pat on Shoulder; Playing Erhu; Playing Ukulele; Playing Drum; Shaking Head; Draw Circles; Draw a Cross; Pushing; Pulling; and Swipe Up.

Unseen activities (10). Drinking; Jumping; Sitting Down; Turning; Walking; Stomping; Nodding; Swipe Left; Swipe Right; and Swipe Down.

B.3 WiQA Construction Details

(i) Generation. For each recording, Qwen3.6-35B generates four multiple-choice questions from the synchronized video, covering action recognition, body-part involvement, motion attributes, and temporal reasoning. Each question contains one correct answer and three mutually exclusive distractors. The questions are designed to focus on motion information that can plausibly be reflected in WiFi CSI. Action questions test the performed activity, body part questions identify the primary moving body regions, motion-attribute questions characterize properties such as direction, symmetry, or displacement, and temporal questions test the ordering of motion phases.

(ii) Validation and Regeneration. We use a two-stage validation procedure. First, an LLM-based validator checks each generated question for formatting, answer-option consistency, and compliance with the required question structure. When a check fails, the identified error is returned to the generator, which regenerates the complete question. This process is repeated for up to five iterations. A recording is retained in the WiQA dataset only if it has one valid question in each of the four categories. If any category still lacks a valid question after the final regeneration attempt, the recording and its associated questions are excluded from WiQA. Separately, we conduct a human audit of 500 randomly selected recordings, covering all 2,000 associated question–answer pairs, to identify obvious question-only answer cues. Among these recordings, 482 (96.4%) contain no such cues, whereas 18 (3.6%) contain at least one question whose correct answer can be inferred from

the question and answer options alone. This audit assesses potential language-only shortcuts rather than overall annotation correctness. It is not used to filter the dataset, and evaluation uses the complete retained question set. The final WiQA question set contains 103,240 questions over 25,810 recordings, with one question from each of the four categories for every retained recording.

B.4 Controlled Caption-Design Ablation

The ablation without motion-focused caption design uses the same 16 frames, activity label, caption count, word range, VLM, and downstream training schedule. The only change is the instruction: the baseline asks for a general video description and removes the WiFi-perceivability constraints. We then retrain alignment and captioning from the same initialization and evaluate both activity sets with the metrics in Table 2. This design prevents caption count, model capacity, or optimization budget from confounding the effect of motion-focused supervision.

C Visual Priors Beyond WiFi Observability

Figure 3 presents representative examples of both successful predictions and remaining failure cases. In several seen activities, WiFi2Cap* introduces visually specific concepts that are not supported by the observed motion. For example, it associates Eating with "drumsticks" and a "fretboard," and interprets Using a Phone as strumming strings and pressing frets. Similar activity substitutions appear in other cases, where object- or scene-related priors dominate the generated description. In comparison, our motion-focused model generally places greater emphasis on body configuration, active body parts, motion trajectory, and temporal progression.

The figure also shows that motion-focused supervision does not eliminate all errors. For Using a Phone, the generated caption still refers to a handheld object and gaze direction. In the unseen examples, the model often preserves coarse motion structure but may confuse the underlying activity or motion direction. For example, the Sitting Down output describes a transition toward standing, while the Swipe Down and Drinking examples retain only partial motion cues. These cases indicate that motion-focused supervision reduces reliance on visually specific scene priors, but fine-grained action identity, direction, and temporal interpretation remain challenging when the CSI pattern is ambiguous or the activity is unseen.

Overall, Figure 3 suggests that restricting supervision to motion-related semantics changes the type of errors made by the captioning model. WiFi2Cap* more frequently completes ambiguous motion patterns with visually specific activity hypotheses, whereas our model more often remains centered on human motion. However, the remaining failures show that motion-focused supervision alone does not guarantee instance-level motion correctness.

D LLM-Judge Robustness

Our primary caption evaluation uses Qwen3.6-35B as an LLM-based activity classifier. We define LLM-based activity-identification accuracy as the percentage of generated captions whose described activity can be correctly identified by the LLM. For each generated caption, the LLM receives only the caption together with the candidate activity labels and is asked to select the single activity best described by the caption. The judge has no access to the WiFi signal, video, reference caption, or ground-truth activity label. The accuracy is then computed as the fraction of generated captions for which the selected activity matches the ground-truth activity. The candidate set contains 45 labels for seen-activity evaluation and 10 labels for unseen-activity evaluation.

To evaluate whether the results depend on the choice of LLM judge, we repeat the same evaluation with GPT-5.6-terra and Gemini-3-flash-preview. All three judges receive the identical generated captions, candidate activity sets, and evaluation prompt, and are constrained to return exactly one activity label. As shown in Table 6, the three judges produce similar activity-identification accuracy on seen activities, while the unseen results show somewhat greater variation. The overall pattern is consistent across judges: activity information is substantially easier to recover from generated captions for seen activities than for unseen activities.

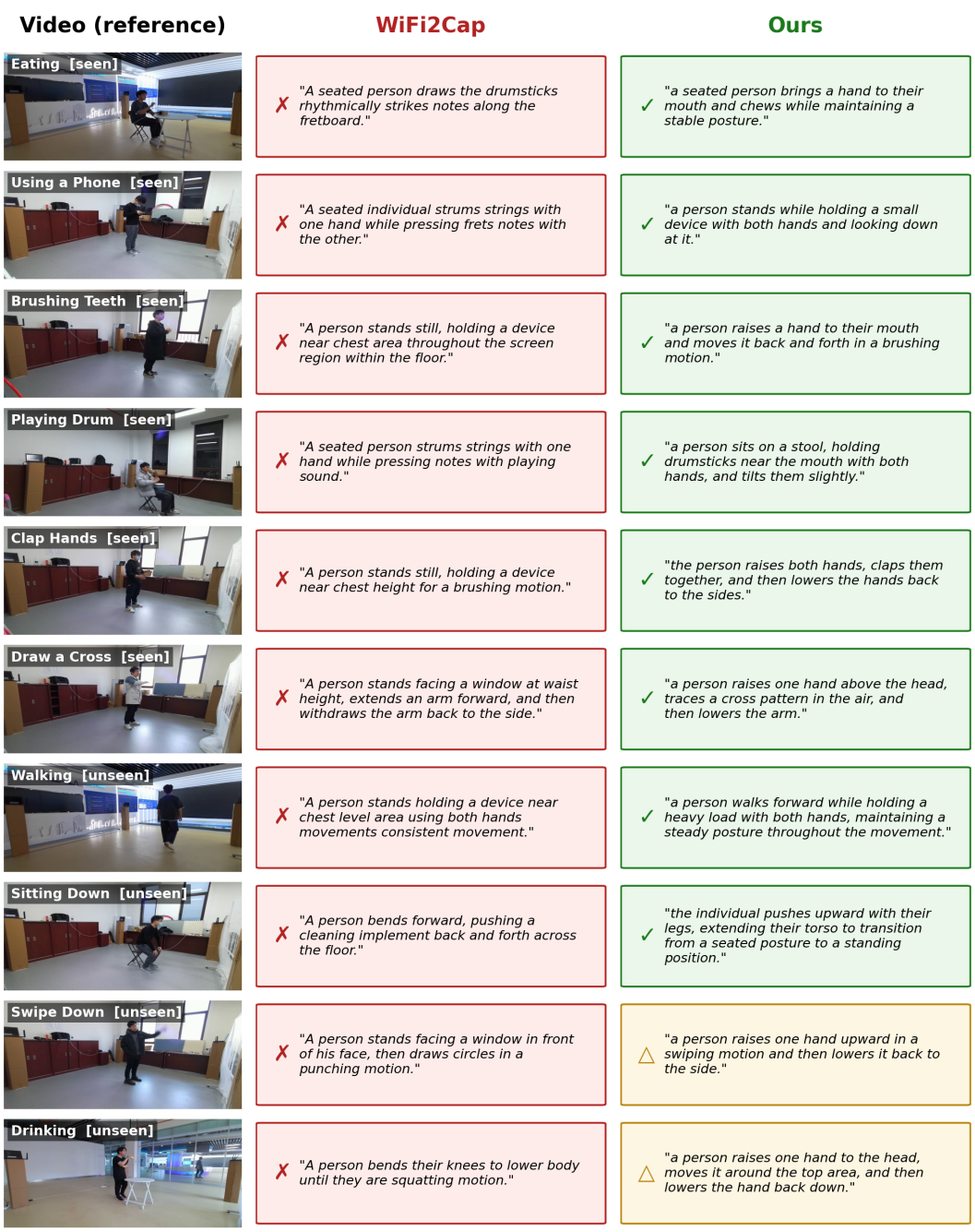

Figure 3: Representative captioning examples and failure cases. Video frames are shown only for reference and are not provided to either model at inference. WiFi2Cap* frequently introduces object- and scene-specific concepts when interpreting ambiguous CSI patterns. Our motion-focused model generally emphasizes body configuration and motion trajectory, although errors in activity identity, direction, and temporal progression remain. Unsupported visual assumptions are highlighted in red.

E Cross-Dataset Portability

We further evaluate SignalLang on XRF V2 to examine its performance under acquisition conditions different from those of XRF55. To distinguish the transferability of an XRF55 trained checkpoint from the portability of the overall framework, we compare three settings: Zero-shot, Fine-tuned,

Table 5: Representative caption outputs from WiFi2Cap* and SignalLang. In the third example, WiFi2Cap* describes an embrace, whereas SignalLang describes the neck-directed hand interaction specified by the reference caption.

Activity	Reference caption	WiFi2Cap*	SignalLang
Jumping Jack	A person jumps and spreads legs while raising arms upward, then jumps and returns arms and legs to the starting position.	A person bends their knees to lower body into a squatting position, then returns to a standing position.	A person jumps up and down while spreading their arms and legs apart in a rhythmic jumping jack motion.
Pulling	A person stands upright and extends one arm forward before drawing it back toward the torso in a controlled, repetitive pulling motion.	A person stands still, holding a device near the face slightly while keeping the limb extended wrist outward.	A person stands upright, extends an arm forward, and pulls it back toward the shoulder in a steady, controlled motion.
Choke Someone's Neck	Two individuals stand close while one extends both arms to encircle the other's neck, squeezing tightly in a choking gesture.	Two people stand facing each other then embrace by wrapping their arms around each other's backs.	Facing each other, one individual extends their arms to encircle the other person's neck with both hands in a grasping motion.

Table 6: LLM-based activity-identification accuracy (%) from generated captions across three independent LLM judges. Each judge receives only the generated caption and the candidate activity labels and selects the activity best described by the caption.

Activity Set	Qwen3.6-35B	GPT-5.6-terra	Gemini-3-flash-preview
Seen	60.6	63.1	65.2
Unseen	24.8	24.8	19.4

and Retrained, as reported in Table 7. Zero-shot directly applies the XRF55 trained model to XRF V2 without updating any model parameters. Fine-tuned initializes all transferable model parameters from the XRF55 alignment checkpoint, including the WiFi encoder, text encoder and decoder, cross-attention modules, and their corresponding momentum copies. The auxiliary activity classification head is newly initialized for the 34 XRF V2 classes, and the momentum queues are reset. We then perform 100 epochs of WiFi–text alignment on XRF V2, updating all online model components without freezing any layers. Retrained does not load the XRF55 alignment checkpoint. Instead, it initializes the WiFi encoder from MAE-pretrained weights and performs 450 epochs of WiFi–text alignment on XRF V2. Fine-tuned and Retrained use identical alignment objectives, all other hyperparameters, and the same subsequent caption fine-tuning procedure, differing only in initialization and alignment training duration. As shown in Table 7, direct zero-shot transfer performs poorly, with an LLM-based activity-identification accuracy of 2.3% against a 2.9% chance level and a BLEU-4 score of 0.0. The low METEOR, ROUGE-L, and CIDEr scores in Table 7 are consistent with generic lexical overlap rather than correct activity semantics. Qualitative inspection further shows that the model often generates fluent XRF55-style captions, including human–human interaction templates that are only weakly related to the XRF V2 input. This behavior suggests that the out-of-distribution CSI falls outside the region of the aligned WiFi–text manifold covered during XRF55 training, causing the decoder to rely primarily on its language prior.

We attribute this failure to a compound acquisition-domain gap. XRF V2 uses a different transceiver setup and antenna geometry and records CSI at 50 Hz, compared with 200 Hz in XRF55. Resampling every segment to 1,000 timesteps matches the required input size, but it cannot recover high-frequency motion components that were not captured at the original sampling rate. Moreover, mapping variable-duration segments to the same temporal length stretches or compresses their motion dynamics, changing the discrete-time periodicity observed by the encoder. Similarly, affine normalization matches global amplitude statistics but does not align device response, antenna-link structure, temporal spectra, or environment-specific multipath.

Table 7: Caption generation and retrieval on XRF V2 under the Zero-shot, Fine-tuned, and Retrained settings defined in Appendix E. LLM Cls.@1 denotes activity-identification accuracy from generated captions. Retrieval is reported as activity-level Recall@K for WiFi-to-text (Wi2TR) and text-to-WiFi (T2WiR) retrieval using WTC+WTM.

	Caption Generation					Retrieval	
Training setting	B@4	METEOR	ROUGE-L	CIDEr	LLM Cls.@1	Wi2TR R@1/R@5	T2WiR R@1/R@5
Zero-shot	0.0	10.8	9.9	0.3	2.3	2.2/2.4	8.5/15.2
Fine-tuned	41.1	**58.8**	63.0	311.1	35.3	43.1/43.2	60.6/87.2
Retrained	**41.5**	54.8	**63.7**	**326.6**	**36.2**	**45.4/45.5**	**70.9/93.1**

Table 8: Performance comparison under motion-focused and non-motion-focused caption supervision. BLEU-4 and CIDEr use the multi-reference protocol, with single-reference scores in parentheses. Lexical metrics are evaluated against references matching each supervision style and are not directly comparable across settings. LLM Cls.@1 denotes activity-identification accuracy from generated captions. Halluc. proxy is the percentage of generated captions containing at least one term from a fixed 45-word visual lexicon that is absent from the paired video-derived reference caption.

Supervision	Split	BLEU-4	METEOR	ROUGE-L	CIDEr	LLM Cls.@1	Halluc. proxy
No-motion-focus	Seen	44.1 (24.2)	55.9	54.4	79.1 (129.5)	**70.5**	74.3%
	Unseen	35.1 (20.5)	45.8	47.2	15.4 (66.4)	**28.0**	77.1%
Motion-focused	Seen	43.8 (16.8)	**61.7**	**58.8**	**129.8** (113.5)	60.6	**3.4%**
	Unseen	11.5 (4.2)	35.0	36.5	14.1 (21.8)	24.8	**4.5%**

The environment shift further compounds this mismatch. CSI jointly reflects human motion and the static propagation paths of the surrounding room. The XRF55 encoder is therefore optimized using motion patterns superimposed on the multipath backgrounds of the XRF55 environments, whereas XRF V2 is collected in a dining room, study room, and bedroom with different layouts and propagation conditions. Consequently, recordings of semantically similar activities may occupy substantially different regions of the learned signal space even after temporal resampling and global amplitude normalization.

The contrast with full retraining localizes this limitation. Using the same SIGNALLANG architecture and training recipe on XRF V2 yields 36.2% LLM-based activity-identification accuracy, 45.4% Wi2TR R@1, 70.9% T2WiR R@1, and 78.3% WiQA accuracy. Thus, the architecture, learning objectives, and question-generation and evaluation pipeline can be successfully applied to a new acquisition domain, while the signal representation learned from XRF55 does not transfer zero-shot. The current results therefore demonstrate framework-level portability rather than acquisition-invariant checkpoint generalization.

F The Effectiveness of Motion-Focused Caption Generation

Our caption supervision is deliberately *motion-focused*: the video captioner is constrained to describe posture changes, active body parts, movement direction, and temporal order, and is forbidden from mentioning appearance, objects' visual attributes, or the scene. A natural question is whether this restriction helps or hurts. We therefore build a no-motion-focus counterpart under identical conditions, using the same video captioner, the same length and label constraints, and the same training recipe, but with no restriction on what the LLM may describe. As a result, the generated captions freely mention clothing, objects, and the environment. Only the caption style differs between the two arms.

Table 8 reports the comparison, and the result is instructive precisely because it is not one-sided. On LLM-judge action recovery the unrestricted arm scores higher. The reason is structural rather than perceptual: unrestricted captions name activity-revealing objects and scenes ("spinning a hula hoop in a bright kitchen"), so a metric that only asks which activity is this can be satisfied by reciting class identity, which scene-inclusive language does most directly.

Table 9: Caption-based WiQA accuracy (%) without WiQA-specific fine-tuning. FullCap uses motion-focused captions generated separately for each recording during training, whereas SimpleCap uses one canonical caption shared by all recordings of each activity class. For each question, an LLM selects an answer using the model-generated caption as its only evidence about the observed activity. Overall denotes accuracy across all samples.

Activity Set	Caption type	Action	Body part	Motion attr.	Temporal	Overall
Seen	FullCap	**78.4**	**78.1**	**52.4**	**60.2**	**67.3**
	SimpleCap	72.6	71.7	46.4	42.7	58.3
Unseen	FullCap	**42.1**	**51.7**	24.8	**55.6**	**43.5**
	SimpleCap	32.7	44.6	24.5	39.4	35.3

We quantify reference-unmatched visual-term usage with a lexicon-based hallucination proxy. A generated caption is flagged if it contains at least one term from a fixed lexicon of 45 visual terms covering colors, clothing, objects, and room types that is absent from its paired video-derived reference caption. We report the percentage of flagged captions. This proxy is not a verified factual-error rate: reference captions may omit genuine visual details, and hallucinations outside the lexicon are not counted. The no-motion-focus model has a substantially higher proxy rate. Specifically, 74.3% and 77.1% of its captions on seen and unseen activities, respectively, contain at least one reference-unmatched lexicon term, and 46.4% contain two or more. In contrast, the corresponding rates for the motion-focused model are only 3.4% and 4.5%. Because the captioning model receives only WiFi at inference, such visually specific details are not directly observed and are likely influenced by correlations learned from the training captions. Figure 4 shows two representative cases that we verify against the source videos. In the first example, a mimed push is performed without interacting with a visible target, yet the no-motion output includes the phrase "extends his arm forward to push against the whiteboard." In the second example, the action is performed with an empty hand, while the no-motion output includes "uses a white object to strike the other man." These examples confirm that at least some of the visual terms captured by our metric correspond to factual hallucinations. In both cases, the motion-focused outputs instead include "extends one arm upward and pushes outward with the hand" and "swings an arm forward, striking the upper body of another person," respectively. Because the two supervision settings use references matching their respective caption styles, their lexical scores should not be interpreted as a direct measure of relative caption quality. Table 8 reports fewer reference-unmatched visual terms under motion-focused supervision, while the video-verified examples in Figure 4 illustrate cases where it avoids unsupported visual additions. However, the reduction in the proxy should not be interpreted as an equivalent reduction in verified factual errors. The lower LLM Cls.@1 under motion-focused supervision also indicates that restricting visually specific descriptions can make activity identity less directly recoverable from the generated captions.

Table 9 compares per-recording caption supervision with a single canonical caption per activity class. FullCap uses the motion-focused caption generated for each recording, whereas SimpleCap assigns the same canonical description to all recordings of an activity. FullCap improves caption-based WiQA accuracy on both activity sets. On seen activities, overall accuracy increases from 58.3% to 67.3%, and temporal accuracy from 42.7% to 60.2%. On unseen activities, overall accuracy increases from 35.3% to 43.5%, and temporal accuracy from 39.4% to 55.6%. The improvement on unseen motion-attribute questions remains small, from 24.5% to 24.8%. These results support the use of per-recording motion-focused captions rather than a single canonical description in this setting. However, the comparison does not separately control for linguistic diversity and the descriptive content of the captions. It therefore does not establish that the gains arise from recording-specific motion information or demonstrate discrimination among repetitions of the same activity.

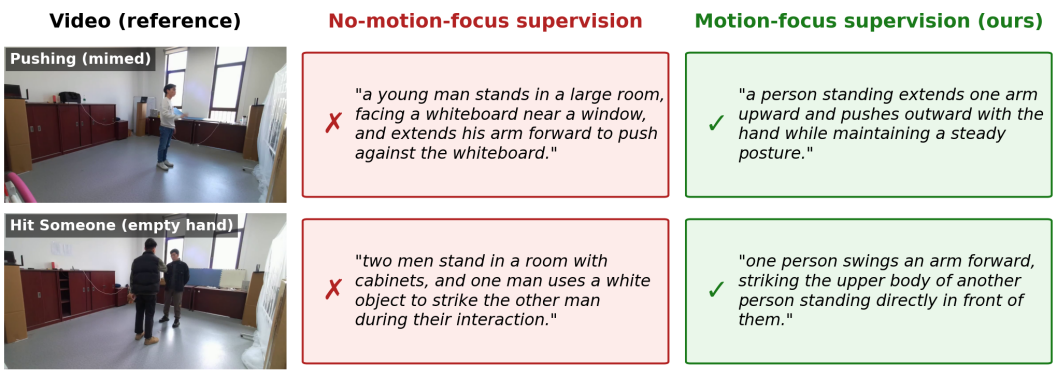

Figure 4: Hallucination under unrestricted supervision. When trained with scene-inclusive captions, the no-motion model completes its descriptions with objects suggested by scene priors but unsupported by the signal. These include an invented push target (top) and an invented striking implement (bottom). In contrast, the motion-focused outputs describe body movements without introducing these unsupported interactions.

G Limitations and Ethical Considerations

Our repetition-level split does not explicitly hold out subjects or scenes. Therefore, the reported results should not be interpreted as controlled cross-subject or cross-scene evaluation. The unseen-activity protocol evaluates transfer to held-out activity classes, but generalization across subjects and environments is not separately evaluated. Our evaluation also does not establish fine-grained discrimination among repetitions of the same activity or isolate the contribution of recording-specific motion information. Although the generated captions and questions undergo automated validation, only a subset of WiQA recordings is manually audited for question-only cues, and annotation errors may remain. Because WiFi sensing can monitor human activity without direct line of sight, practical deployments should require informed consent, data minimization, access controls, and safeguards against covert monitoring. This study uses existing research datasets and does not collect new participant data.