Toward Understanding Subjective Spatial Interpretation: An Analysis of Action Deviation from Place Affordances

Yoshiki Watanabe
Nagoya University
Nagoya, Aichi, Japan
yoshiki@ucl.nuee.nagoya-u.ac.jp

Kaiya Shimura
Nagoya University
Nagoya, Aichi, Japan
kaiya@ucl.nuee.nagoya-u.ac.jp

Hideki Deguchi
Nagoya University
Nagoya, Aichi, Japan
deguchi@ucl.nuee.nagoya-u.ac.jp

Kenta Urano
Nagoya University
Nagoya, Aichi, Japan
urano@nagoya-u.ac.jp

Takuro Yonezawa
Nagoya University
Nagoya, Aichi, Japan
takuro@nagoya-u.ac.jp

Nobuo Kawaguchi
Nagoya University
Nagoya, Aichi, Japan
kawaguti@nagoya-u.ac.jp

Abstract

Mutual understanding is grounded in sharing internal states such as beliefs, evaluations, and feelings, which cannot be observed directly but can be inferred from externally observable behavior. Reading such interpretations from observation enables computational sharing that supports deeper mutual understanding. As a first step, we focus on how much an action deviates from what is typical of the place, since such deviations tend to reveal intentions and circumstances. We call the collective tendency of actions associated with a place its place affordance, and define semantic deviation as the degree to which an action departs from it. Detecting semantic deviation requires two capabilities: describing the place affordance from a background image, and extracting the deviating action component from an observation image. We investigate whether multimodal embedding models can provide foundational signals for both, using ImageBind. For the first, results suggest ImageBind can discriminate typical from atypical actions to some extent zero-shot. For the second, we propose a difference vector indicator subtracting the background embedding from the observation embedding. Experiments on synthetic images with controlled typicality suggest that, under atypical conditions, the deviating component surfaces as a rise in relative rank for most action classes.

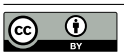
CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; **HCI theory, concepts and models**.

Keywords

place affordance; multimodal embedding models; semantic deviation; subjective interpretation; shared reality

ACM Reference Format:
Yoshiki Watanabe, Kaiya Shimura, Hideki Deguchi, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2026. Toward Understanding Subjective Spatial Interpretation: An Analysis of Action Deviation from Place Affordances. In *Companion of the 2026 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '26), October 11–15, 2026, Shanghai, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3798063.3839183



This work is licensed under a Creative Commons Attribution 4.0 International License.
UbiComp Companion '26, Shanghai, China
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2533-3/2026/10
https://doi.org/10.1145/3798063.3839183

1 Introduction

Mutual understanding between people is grounded in sharing with others the commonalities of internal states about the world such as beliefs, evaluations, and feelings, as characterized in the Shared Reality framework [7, 11]. We advocate an information and communication paradigm called the Internet of Realities (IoR) [1, 26]. IoR seeks to computationally represent not only physically observable facts but also individuals' subjective interpretations of the world. Its goal is to support deeper mutual understanding between people. Realizing this goal requires reading each person's subjective interpretation from what is externally observable and turning it into a form that can be represented and shared computationally.

Subjective internal states cannot be directly observed from the outside. But behavior is externally observable, and since people act with intentions of their own, behavior partly reflects the subjectivity of the actor. Thus, what we seek to understand is not the behavior itself but the internal states behind it. In particular, behaviors that are not usually seen in a given place tend to reveal that person's intentions and circumstances. In this paper, we refer to the collective tendency of typical actions that many people associate with a given place as the place affordance [8, 17], and we define semantic deviation as the degree to which a person's action diverges from this place affordance. This study computationally addresses this semantic deviation as a first step toward reading subjective interpretations from the outside, targeting deviation from population-level typicality rather than individual subjective interpretation.

The relation between a place and the actions performed in it is not a question limited to the understanding of subjectivity; it appears commonly across multiple fields such as action recognition and video anomaly detection. Choi et al. [4] pointed out the difficulty of dependence on scene context in action recognition, and Chen et al. [3] noted the difficulty of judgments in which the same behavior is normal or abnormal depending on the place, illustrated by the contrast between running on a beach and running on a street. However, to the best of our knowledge, no framework has been proposed that systematically and quantitatively evaluates this semantic deviation.

In this study, we propose a framework that determines semantic deviation by comparing the place affordance obtained from the

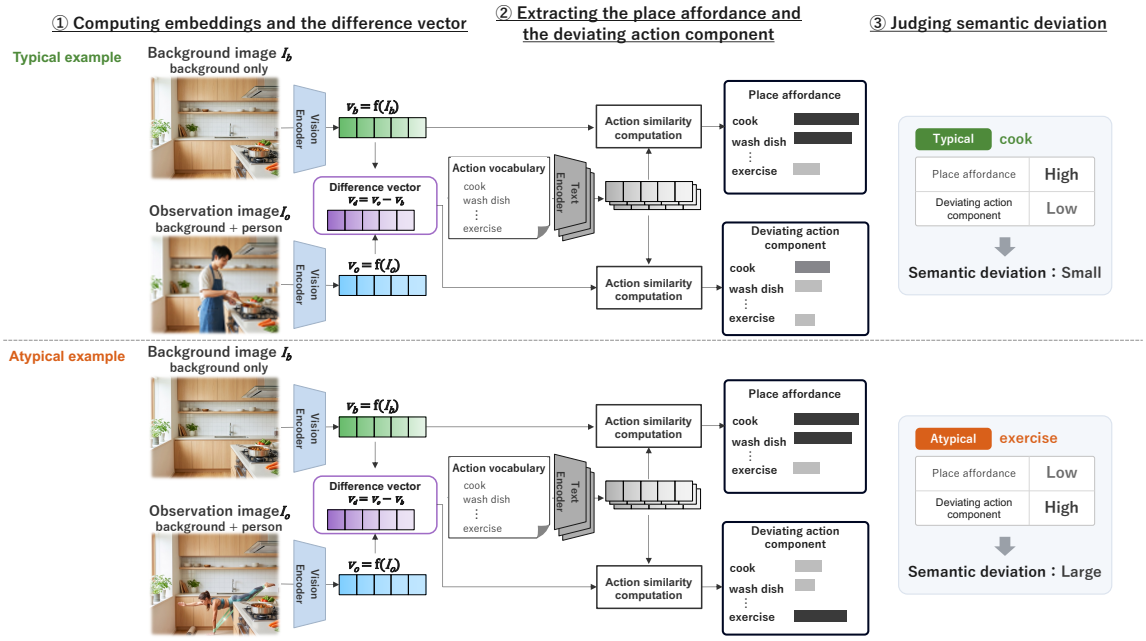


Figure 1: Conceptual diagram of semantic deviation determination: the place affordance from the background image is compared with the action component from the embedding difference between the observation and background images.

background image with the action component obtained from the embedding difference between the observation image and the background image (Figure 1). This framework requires two capabilities. The first is the capability to describe the place affordance from a background image. The second is the capability to extract the action component that deviates from the place affordance through the embedding difference between the observation image and the background image. Using ImageBind [9], a multimodal embedding model that maps heterogeneous modalities into a common latent space, as the evaluation target, this study poses the following questions corresponding to these two capabilities.

Question 1 Can a multimodal embedding model describe place affordances from a background image in a zero-shot manner?

Question 2 Can a difference vector indicator based on the embedding difference between the observation image and the background image extract, under atypical conditions, the action component that deviates from the place affordance?

To verify Question 1, Experiment 1 evaluates whether ImageBind can describe place affordances from background images, using SUN397 [23], a scene recognition dataset consisting of 397 scene categories, and taking as reference the annotations of SUN Action [21] in which the actions typically performed at each place are aggregated by hand. To verify Question 2, Experiment 2 uses as a difference vector indicator the difference $\mathbf{v}_d = \mathbf{v}_o - \mathbf{v}_b$ between the embedding $\mathbf{v}_o$ of the observation image and the embedding $\mathbf{v}_b$ of the background image. To generate observation images, we extract image regions from the MPII Human Pose Dataset [2]. Each region contains a person performing an action and the objects involved in that action. These regions are composited onto typical and atypical backgrounds from SUN397. We then obtain the rank of the performed action among many candidate actions. This rank is compared with the rank obtained when only the background image is used.

Our main contributions are as follows.

(1) We formulate semantic deviation detection as a problem consisting of two basic capabilities, namely "describing place affordances" and "extracting the deviating action component," and propose an evaluation framework for each.
(2) We quantitatively evaluate the place affordance description capability of the multimodal embedding model ImageBind across 397 scene categories, using the human annotations of the SUN Action dataset as the reference. The results suggest that ImageBind can discriminate typical from atypical actions to a certain extent, consistently across four prompt formats.
(3) We propose a difference vector indicator based on the embedding difference between the observation image and the background image. Experiments with synthetic images over 24 action classes of the MPII Human Pose Dataset suggest that, under atypical conditions, the deviating action component surfaces as a rise in relative rank for the majority of the action classes.

2 Related Work

2.1 Computational Description of Spatial Affordances

Gibson defined an affordance as the action possibility that the environment offers to an animal [8]. Gibson discussed affordances as offered not only by objects but also by surfaces, places, and the environment as a whole. In ecological psychology, Rietveld and Kiverstein reformulated affordances as "the relations between

aspects of a material environment and abilities available in a form of life," and proposed a "rich landscape of affordances" in which a whole space, rather than individual objects, evokes a repertoire of purposive actions [17]. Our notion of place affordance builds on this line of work.

In computer vision, the correspondence between scene images and human actions has been studied. Vu et al. built SUN Action by crowdsourcing typical human actions for each of the 397 scene categories of SUN397, formalizing the task of predicting typical actions from scene images [21]. Chuang et al. proposed the ADE-Affordance dataset, which annotates relations between actions and objects (classified as physical obstacle, functional mismatch, social inappropriateness, social prohibition, danger, etc.), together with a graph neural network model that infers, for each object, the appropriateness of an action conditioned on the surrounding scene [5]. More recently, OpenScene [15] realizes zero-shot affordance reasoning over 3D scenes by jointly embedding CLIP features in 3D, and AffordanceCLIP [6] uses CLIP alone to segment, within an image of an object, the parts that afford a given action.

These works describe what a place or an object offers, but do not ask how an actually observed action relates to what is typical of the place. Our work differs in that it describes the place affordance as a score distribution over actions for an arbitrary place image without additional training, and adds a mechanism for extracting the deviating action component via the embedding difference between the observation and the background.

2.2 Multimodal Embedding Models and Their Systematic Biases

CLIP, proposed by Radford et al., embeds images and text in a shared latent space by contrastive learning on a large corpus of image and text pairs collected from the web [16]. CLIP can identify classes not explicitly labeled at pretraining time via cosine similarity to text prompts in a zero-shot manner, and has been applied to a wide range of visual tasks.

The CLIP framework has been extended to modalities beyond image and language. AudioCLIP [10] integrates audio; X-CLIP [14], VideoCLIP [24], and ActionCLIP [22] extend CLIP to video and action recognition, all in a way compatible with the CLIP embedding space. ImageBind [9] pushes this direction further by integrating six modalities (image, text, audio, depth, thermal, and IMU) into a single embedding space.

In principle these models can evaluate the association between any spatial category and any action description without additional training. However, systematic biases in their embedding spaces have been pointed out. Yuksekgonul et al. showed that models of the CLIP family have limited capacity for representing object relations and attribute bindings [27], and Thrush et al. quantified the limits of compositional reasoning in vision and language models with the Winoground benchmark [20]. These results suggest that VLM representations are strong at "detecting individual concepts" and weak at "distinguishing relational structure between concepts." On ImageBind, we evaluate how well such a VLM representation captures the relation between place and action.

We use ImageBind in our implementation because it computes the cosine similarity between images and action descriptions in a zero-shot manner and is expected to have an embedding space whose structure aligns with action concepts. How accurately the model describes place affordances in its embedding space depends on its pretraining data and is not obvious. We verify this point in Experiment 1.

2.3 Semantic Consistency between Spatial Context and Action

Whether a given action is semantically appropriate in a given space has been discussed repeatedly in action recognition and video anomaly detection, but no systematic framework computationally describes the relation itself. In action recognition, the tendency of models to rely excessively on spatial cues, known as scene bias, has long been recognized as a problem. Choi et al. demonstrated that action recognition models rely on spatial information (e.g., "in a basketball court → playing basketball") and proposed a debiasing method based on an adversarial loss [4]. Choi et al. treat the semantic correlation between place and action as "error to be removed," whereas we treat this correlation itself as the basis of normative affordance and cancel the typical component from the background to extract the deviating action component.

In video anomaly detection, supervised, weakly supervised, and unsupervised methods on datasets such as UCF-Crime [19], ShanghaiTech [12], and CUHK Avenue [13], and recent zero-shot methods leveraging large language models (LLMs) [25, 28], all rely on training data specific to the domain or on LLMs' general knowledge as the criterion of judgment. They are not designed to incorporate the visual features of the space itself as the basis of the norm. DecoAD [3] decouples the relation between scene and action on a knowledge graph and shares our motivation, but it requires building a knowledge graph and weakly supervised learning.

These prior works show the importance of treating the relation between space and action, but none of them provides a framework for computationally describing the relation itself. We address this gap by verifying whether a multimodal embedding model can describe place affordances, and by proposing the difference vector indicator that cancels the typical component from the background and extracts the deviating action component from the observation image.

3 The Difference Vector Indicator

We propose a vector operation in the embedding space of a multimodal embedding model to computationally describe how much a person's action in a given space deviates from the place (Figure 1). The method is formulated as an abstract design built on an encoder model $f(\cdot)$. In this paper we evaluate the instantiation that uses ImageBind [9].

3.1 Formulation

Let $f(\cdot)$ be an encoder model that maps an image or text to a d-dimensional embedding space:

$$f : \mathcal{X} \to \mathbb{R}^d. \tag{1}$$

For an observation image I_o that contains a person and an action, and for an image I_b of the background region of the same image,

we define the embeddings

$$\mathbf{v}_o = f(I_o) \in \mathbb{R}^d, \tag{2}$$

$$\mathbf{v}_b = f(I_b) \in \mathbb{R}^d. \tag{3}$$

Here $\mathbf{v}_b$ is the *place affordance vector* describing the place, and $\mathbf{v}_o$ is the vector describing the whole observation including the person's action.

Next, we define the *difference vector* $\mathbf{v}_d$ that captures the semantic deviation:

$$\mathbf{v}_d = \mathbf{v}_o - \mathbf{v}_b \in \mathbb{R}^d. \tag{4}$$

$\mathbf{v}_d$ is the residual obtained by removing the typical component from the observation. When the background is typical for the action, both $\mathbf{v}_o$ and $\mathbf{v}_b$ have a strong component along that action direction, so the action component originating from the person and object pair is small relative to the typical background component, and the component along the action direction in $\mathbf{v}_d$ appears weakly. When the background is atypical, $\mathbf{v}_b$ carries almost no signal along the action direction, so the action component originating from the person and object pair appears strongly in $\mathbf{v}_d$. This component corresponds to the action component that deviates from the place affordance. We do not assume linear compositionality of the latent space. The subtraction is a heuristic diagnostic whose added value we quantify empirically in Section 5.

Several scalar reductions of the magnitude of semantic deviation are possible, such as the norm $\|\mathbf{v}_d\|$ or the inner product with a specific action text embedding. The difference vector indicator we propose is a framework level design that leaves the concrete scoring to the implementation. In this paper we adopt scoring by cosine similarity with the action text embedding $f(t_A)$:

$$s_d(A) = \cos\big(\mathbf{v}_d,\ f(t_A)\big). \tag{5}$$

This scoring is used throughout the paper: Experiment 1 (Section 4) applies the same cosine scoring to $\mathbf{v}_b$ to describe the place affordance, and Experiment 2 (Section 5) evaluates $s_d(A)$ itself.

The indicator is agnostic to the model and applies in principle to any multimodal embedding model that ties together heterogeneous modalities, and it adapts immediately to an arbitrary place without additional training.

3.2 Validation Plan

Detecting semantic deviation requires two capabilities. The first is to describe the place affordance from a background image. This corresponds to the validity of the formulation (Eq. 3) that $\mathbf{v}_b$ semantically describes the place affordance. The second is to extract, from the embedding difference between the observation and the background, the action component that deviates from the place affordance. This corresponds to whether the difference operation (Eq. 4) cancels the typical component aligned with the place affordance and surfaces the deviating component. Questions 1 and 2 in Section 1 correspond to these two points. We validate Question 1 in Experiment 1 (Section 4) and Question 2 in Experiment 2 (Section 5), with a metric chosen for the nature of each question: ROC AUC for how well the cosine similarity separates typical from atypical pairs, and the relative rank of the action performed in the observation image among the candidate actions for the discriminability within an image.

4 Experiment 1: Zero-Shot Place Affordance Description

This section addresses Question 1: can a multimodal embedding model describe place affordances in a zero-shot manner? We evaluate ImageBind [9] on its ability to distinguish typical from atypical (place, action) pairs, and we compare four prompt formats.

4.1 Dataset and Reference

We use SUN397 [23] as the image dataset and SUN Action [21] as the human reference for place affordances. SUN Action aggregates participants' freeform responses about the actions they imagine taking in each SUN397 scene category, which corresponds to our notion of place affordance. We reconstruct the annotations from the raw responses released by the SUN Action authors, following the original aggregation procedure.

SUN Action uses separate action vocabularies for outdoor and indoor scenes to exclude clearly mismatched pairs such as washing dishes at the beach. Specifically, 38 outdoor actions and 23 indoor actions receive vote counts, with 9 actions shared between the two sets (Table 1). Following this design, we select the action set according to each scene's indoor/outdoor attribute. The evaluation covers 94,111 images in total.

4.2 Metric and Prompts

For each (scene image I, action a) pair, we use the cosine similarity between the ImageBind image embedding $f(I)$ and the action prompt embedding $f(t_a)$ as the typicality score:

$$S(I, a) = \cos\big(f(I),\ f(t_a)\big). \tag{6}$$

Following the SUN Action paper, we derive a binary ground truth label from vote counts so that the evaluation focuses on unambiguous cases:

- **Human Typical** (positive): votes ≥ 20.
- **Human Atypical** (negative): votes $= 0$.

Pairs with intermediate vote counts are excluded, since their typicality admits multiple interpretations.

We report the ROC AUC of the cosine similarity for this binary task; higher values indicate better separation of typical from atypical pairs.

Because ImageBind's output depends on the wording of the text prompt, we compare four prompt formats that differ in how explicitly they describe the action (Table 2). The placeholder A is replaced with the action verb phrase (e.g., `climb`).

4.3 Results and Discussion

Figure 2 shows the ROC curves. All four prompt formats exceed an AUC of 0.7 (`word`: 0.733, `photo`: 0.759, `photo_people`: 0.785, `people`: 0.822), with the best result obtained by `people`, the simple prompt that focuses on the person. Prompt choice produces an AUC gap of about 0.089.

These results suggest that ImageBind can describe place affordances to a certain extent in a zero-shot manner. Associations between places and actions, accumulated during pretraining at web scale, appear to be reflected in ImageBind's embedding space and do not depend strongly on a particular prompt format. This gives a partial affirmative to Question 1. However, detecting semantic

Table 1: Composition of the action vocabulary in SUN Action.

Outdoor scenes only (29 actions)	Indoor scenes only (14 actions)	Shared by both (9 actions)
cheer, climb, cross bridge, dig, drive, explore, farm, fly, garden, go fishing, harvest, have a picnic, hike, live, open door, park, play, play baseball, play golf, ride, ride train, run, sail, sightsee, skate, ski, smell flower, take a walk, take picture	bowl, cook, dance, do laundry, exercise, gamble, listen, play game, play pool, read, study, wait, wash dish, watch	eat, learn, pray, relax, shop, sit, sleep, swim, work

Table 2: The four prompt formats compared. A denotes the verb phrase of the target action.

Prompt	Format
`word`	A
`photo`	a photo of A
`photo_people`	a photo of a place where people A
`people`	people A

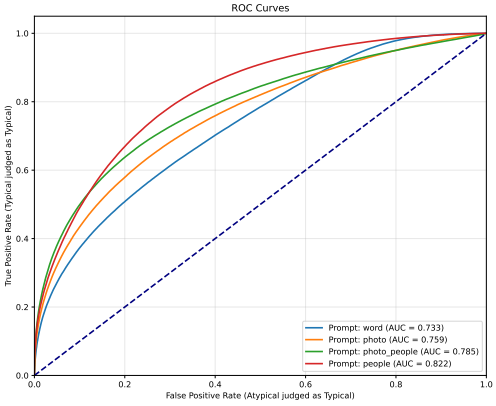


Figure 2: ROC curves for discriminating typical from atypical pairs with the four prompt formats.

deviation also requires extracting from an observation image the action component that departs from the place affordance. The next section addresses Question 2 by evaluating the proposed difference vector indicator.

5 Experiment 2: Extracting Deviating Action Components

This section addresses Question 2: can the difference vector indicator, defined as the embedding difference between an observation image and its background, extract the action component in the observation image that deviates from the place affordance? The case of interest for semantic deviation is the *atypical* combination of action and background. For each action, we obtain the rank of the action performed in the observation image among many candidate actions. We compare this rank under the difference vector indicator and under the score computed from the background alone, and examine what the difference operation contributes.

Table 3: Mapping between SUN Action and representative MPII actions.

SUN Action	Representative MPII action	Images
bowl	bowling	25
climb	rock climbing	102
cook	cooking or food preparation	226
dance[†]	ballet, modern, or jazz	115
dig	digging, spading garden	44
do laundry	laundry (home activities)	55
eat	eating, sitting	53
exercise	yoga	554
farm	farming, rice, planting	76
garden	planting, potting seedlings	159
go fishing	fishing, general	275
harvest	picking fruits/vegetables	33
hike	backpacking	67
play baseball	softball, general	79
play golf	golf	104
play pool	billiards	26
ride	horseback riding	85
run	running	98
sail	sailing	21
shop	food shopping, standing or walking	22
skate	skateboarding	207
ski	skiing, downhill	210
swim	swimming, general	62
take a walk	walking, general	42
wash dish	wash dishes	37
Total (25 candidates)		2,777

[†]Excluded from the evaluation because no atypical background can be constructed.

5.1 Synthetic Observation Images

Validating Question 2 requires observation images in which the typicality of the action and background pair is controlled. We generate such images by compositing person and object regions extracted from the MPII Human Pose Dataset [2] onto SUN397 [23] scenes. A person and object region is an image area containing a person performing an action together with the objects involved in that action.

We extract MPII actions that map one to one to SUN Action actions and, after manual screening, retain the 25 actions with at least 20 image samples. Table 3 lists the mapping and image counts. The action dance is excluded because no SUN397 category has zero votes for it, so no atypical background can be constructed. The final evaluation covers the remaining 24 actions.

For each action we randomly sample 20 source images and generate two composites per source as follows:

(1) **Typical background.** Randomly pick a SUN397 category in which action A receives ≥ 20 votes in SUN Action, then randomly pick one image from that category.

(2) **Atypical background.** Randomly pick a SUN397 category in which action A receives 0 votes, then randomly pick one image from that category.
(3) **Compositing.** Using the extracted person and object mask, overlay the person and object region onto both the typical and the atypical background, yielding two observation images.

Person and object regions are extracted with the method of [18]. The background image I_b is the SUN397 image used as is for compositing. With 24 actions × 20 samples we obtain 480 person and object regions, each composited onto a typical and an atypical background, for a total of 2×480 = 960 samples (480 typical pairs and 480 atypical pairs). Figure 3 shows example person and object regions and the resulting composites.

5.2 Metric and Prompts

For an (image I, action a) pair we compare three ways of computing the typicality score. Let $\mathbf{v}_o$ be the observation embedding, $\mathbf{v}_b$ the background embedding, and $\mathbf{t}_a$ the action prompt embedding.

- Background Only: Cosine similarity between the background embedding and the action prompt: $S_{\text{BG}}(I, a) = \cos(\mathbf{v}_b, \mathbf{t}_a)$.
- Observation Only: Cosine similarity between the observation embedding and the action prompt: $S_{\text{Obs}}(I, a) = \cos(\mathbf{v}_o, \mathbf{t}_a)$.
- Difference: Cosine similarity between the difference vector $\mathbf{v}_d = \mathbf{v}_o - \mathbf{v}_b$ and the action prompt: $S_{\text{Diff}}(I, a) = \cos(\mathbf{v}_d, \mathbf{t}_a)$.

Difference is the proposed indicator, and Background Only represents what the place alone suggests. Observation Only represents what the observation image alone suggests without any background suppression. It lets us test whether the difference operation adds signal beyond simply scoring the observation image directly. We compare the three to examine the effect of the difference operation.

Let the *target action* of an observation image denote the action performed by the person in the image and used to generate the composite. What we want to measure is the discriminability within the image: how well each score separates the target action from the other candidates inside the same image. The rank of the target action among the candidates captures this separation in an intuitively interpretable form, and it is unaffected by the fact that a cosine similarity that counts as high for one background can be ordinary for another. We therefore rank all candidate actions within each image and report the normalized relative rank of the target action. For an observation image I and target action a_{target}, with action set $\mathcal{A}$ drawn per image from the SUN Action vocabulary as in Experiment 1 (23 indoor actions when the background scene is indoor, 38 outdoor actions when it is outdoor, from the union of 52 actions), we compute

$$\text{RelRank}(a_{\text{target}}) = \frac{\left|\{a_i \in \mathcal{A} : S(I, a_i) < S(I, a_{\text{target}})\}\right|}{|\mathcal{A}| - 1}. \quad (7)$$

RelRank = 1.0 means the score function S ranks a_{target} at the top of the action set, and 0.0 at the bottom. Results are aggregated over the 24 actions in Table 3.

The prediction differs between the two conditions. On atypical pairs, the place does not suggest the target action, so the relative rank under Background Only should be low to begin with. If the difference operation extracts the deviating action component, the rank under Difference should exceed it by a large margin. On typical pairs the action component is cancelled together with the typical component of the background, so the rank under Difference may instead drop. To capture this, we compute two rank gain quantities per action: Δ_{BG}, the mean relative rank under Difference minus that under Background Only, and Δ_{Obs}, the same quantity against Observation Only. The following section reports the number of actions with a positive gain and the mean gain over all actions, for both conditions. The former captures the breadth of actions for which Difference outranks the corresponding baseline, and the latter captures the average size of that shift across actions. As in Experiment 1, we evaluate the four prompt formats.

Table 4: Mean relative rank of the target action under Background Only (BG), Observation Only (Obs), and Difference (Diff), by condition. Bold marks the highest score in each row.

	Atypical			Typical		
Prompt	BG	Obs	Diff	BG	Obs	Diff
`word`	0.415	0.606	**0.732**	0.803	**0.886**	0.659
`photo`	0.362	0.609	**0.753**	0.842	**0.911**	0.676
`photo_people`	0.324	0.640	**0.814**	0.870	**0.934**	0.746
`people`	0.328	0.632	**0.826**	0.814	**0.919**	0.781

5.3 Results and Discussion

Table 4 summarizes the aggregate results. On the atypical pairs, Observation Only already gives the target action a moderate rank (mean 0.61 to 0.66 across the four prompts), reflecting that the person and involved objects visible in the observation image carry action information on their own. Difference adds signal on top. It exceeds Background Only by +0.317 to +0.498 (positive for 18 to 23 of 24 actions) and Observation Only by +0.126 to +0.193 (positive for 14 to 23 actions). The largest gains occur under the `people` prompt, consistent with Experiment 1.

Figure 4 shows the per action breakdown on the atypical pairs under the `people` prompt. Actions whose Background Only rank is very low (left) receive the largest gain from Difference (e.g., `cook`, `play baseball`, `play golf`). Actions whose Background Only rank is already high (right, e.g., `take a walk`, `garden`) receive little or negative gain, but this is not a failure of the difference operation itself. ImageBind gives the target action a nontrivial similarity even to backgrounds humans labeled "atypical," so the premise of atypicality does not hold in the embedding space for these actions, leaving little room for the difference operation.

On typical pairs the sign reverses. Difference falls below both baselines (Δ_{BG} from −0.17 to −0.03, Δ_{Obs} from −0.24 to −0.14), and only 1 to 5 of 24 actions gain rank over Observation Only. When the background is itself aligned with the action, subtracting the background embedding also removes the action component, pushing the rank down. This is consistent with the formulation in Section 3. The difference operation suppresses the component that the place explains and surfaces the component that deviates from it.



Figure 3: Example person and object regions and their composites on typical and atypical backgrounds.

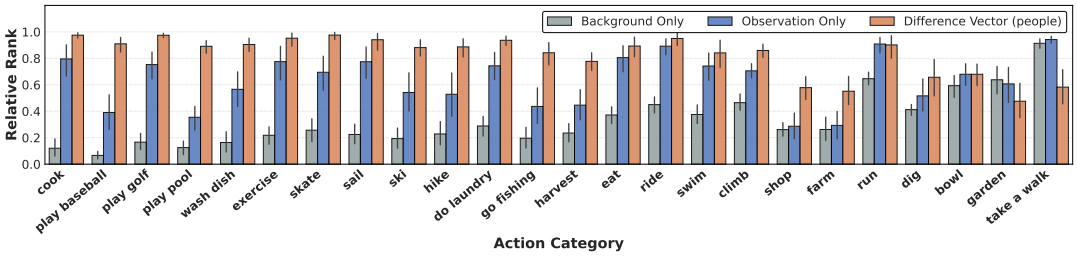


Figure 4: Relative rank per action on atypical pairs under Background Only, Observation Only, and Difference (`people` prompt).

The added value of the difference vector indicator over scoring the observation image directly is therefore conditional. It materializes when the background is atypical for the target action, and disappears or reverses when the background is typical.

6 Discussion and Future Work

6.1 Integrated Interpretation

The two experiments answer both questions affirmatively: ImageBind describes place affordances from a background image in a zero-shot manner (Question 1), and the difference vector indicator lifts the target action's rank in the majority of atypical cases (Question 2). The effect is not perfect, but it shows preliminary effectiveness as a foundational signal for semantic deviation detection.

6.2 Limitations

Our evaluation covers a limited setting, and the evidence should be regarded as preliminary. The following points should be kept in mind when interpreting the results.

- **Single model.** We evaluated only ImageBind. Whether the observed behavior is a general property of the pretraining regime of the CLIP family or specific to ImageBind remains open. Replication on other multimodal embedding models is left to future work.
- **Synthetic observation images.** Experiment 2 used composites of person and object regions and backgrounds. Composites have imperfect illumination, perspective, and visual integration with the background, so part of the signal picked up by the difference operation may stem from compositing artifacts rather than from the action itself; disentangling the two requires evaluation on natural images where an action genuinely deviates from its surrounding scene. In real deployment the background image would be obtained via background modeling from video rather than supplied as ground truth.
- **Limited action coverage.** The 24 evaluated actions are restricted to those with a tight one to one mapping between MPII and SUN Action. Generalization to a broader range of everyday actions is not established. The applicability of the indicator depends both on the action class and on ImageBind's place affordance description capability; for action groups where the latter is weak (e.g., `take a walk`, `garden`) the difference operation also helps less.

6.3 Future Directions

ImageBind's place affordance description capability (Experiment 1) and the preliminary effectiveness of the difference vector indicator (Experiment 2) suggest that strengthening the representation of the acting agent in multimodal embedding models is a central research direction. Natural extensions of our framework include:

- integrating embedding models specialized for action recognition or video understanding to strengthen the extraction of the action component, and the world knowledge of large language models (LLMs) to complement the place affordance description, and
- a hierarchical formulation that applies the difference vector indicator to features at the scene, pose, and object levels separately and combines them.

This problem setting also connects directly to the foundational agenda of the Internet of Realities (IoR) [1, 26]. Computationally representing an individual's subjective interpretation of the world requires inferring an actor's intent and situation from observable

behavior, and semantic deviation from place affordance is a promising observable signal in that direction. Our study is a first step toward computationally describing semantic deviation; the natural next step is to extend the framework to infer the actor's intent and situation from the detected deviation. The findings reported here offer concrete design guidance for future work on computationally modeling the semantic relation between place and action.

7 Conclusion

Semantic deviation detection captures how much a person's behavior deviates from what is typical of a place. This study framed the task as a foundational problem of the Internet of Realities (IoR), which seeks to computationally handle each individual's subjective interpretation of the world, and as a challenge shared with action recognition and video anomaly detection. Toward this goal, we evaluated the multimodal embedding model ImageBind on the two foundational capabilities required: describing place affordances, and extracting the action component in an observation image that deviates from those affordances.

The two experiments answered both questions affirmatively: ImageBind described place affordances from a background image in a zero-shot manner (Question 1), and the proposed difference vector indicator lifted the relative rank of the target action in the majority of atypical cases (Question 2). We position these results as foundational signals for semantic deviation detection and as preliminary evidence that multimodal embedding models can be applied to the problem.

Acknowledgments

This research was partially supported by JST CREST (JPMJCR22M4) and JST RISTEX (JPMJRS23K3).

References

[1] 2025. Vision and Challenges of Internet of Realities. https://internet-of-realities.org/vision/. Accessed: 2026-03-23.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Chenglizhao Chen, Xinyu Liu, Mengke Song, Luming Li, Shaojiang Yuan, Xu Yu, and Shanchen Pang. 2025. Unveiling context-related anomalies: Knowledge graph empowered decoupling of scene and action for human-related video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).

[4] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 2019. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems* 32 (2019).

[5] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. 2017. Learning to Act Properly: Predicting and Explaining Affordances from Images. *arXiv preprint arXiv:1712.07576* (2017).

[6] Claudia Cuttano, Gabriele Rosi, Gabriele Trivigno, and Giuseppe Averta. 2024. What Does CLIP Know About Peeling a Banana?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2238–2247.

[7] Gerald Echterhoff, E Tory Higgins, and John M Levine. 2009. Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on psychological science* 4, 5 (2009), 496–521.

[8] James J. Gibson. 1977. The Theory of Affordances. In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, Robert E. Shaw and John Bransford (Eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 67–82.

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space to Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15180–15190.

[10] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 976–980.

[11] Curtis D. Hardin and E. Tory Higgins. 1996. Shared reality: How social verification makes the subjective objective. In *Handbook of motivation and cognition, Vol. 3: The interpersonal context*. The Guilford Press, New York, NY, 28–84.

[12] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future Frame Prediction for Anomaly Detection - A New Baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6536–6545. doi:10.1109/CVPR.2018.00684

[13] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*. 2720–2727.

[14] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*. Springer, 1–18.

[15] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. 2023. OpenScene: 3D Scene Understanding with Open Vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[17] Erik Rietveld and Julian Kiverstein. 2014. A Rich Landscape of Affordances. *Ecological Psychology* 26, 4 (2014), 325–352. doi:10.1080/10407413.2014.958035

[18] Kaiya Shimura, Kazuma Kano, Tahera Hossain, Shin Katayama, Kenta Urano, Shun Taguchi, Hideki Deguchi, Hiroyuki Sakai, Takuro Yonezawa, and Nobuo Kawaguchi. 2025. Real-time Activity-Aware Video Streaming While Preserving Privacy. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 236–240.

[19] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6479–6488.

[20] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5238–5248.

[21] T.H. VU, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. 2014. Predicting Actions from Static Scenes. In *ECCV*.

[22] Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472* (2021).

[23] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3485–3492. doi:10.1109/CVPR.2010.5539970

[24] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. 6787–6800.

[25] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. 2024. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*. Springer, 304–322.

[26] Takuro Yonezawa. 2025. Towards Internet of Realities - Vision and Challenges. In *2025 IEEE International Symposium on Emerging Metaverse (ISEMV)*. 157–157. doi:10.1109/ISEMV67326.2025.00035

[27] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936* (2022).

[28] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18527–18536.