Uncertainty-Aware Symbolic State Monitoring with Vision-Language Models and Gaussian Naive Bayes

Lioba Schürmann[1], Kento Kawaharazuka[2], Freek Stulp[1], Samuel Bustamante[1,2]

***Abstract*—Autonomous robots require generalizable reasoning abilities to interact intelligently with unknown environments, including the ability to estimate the symbolic state of their surrounding objects. Vision-language models (VLMs) have shown promising advances towards novel scene understanding due to their strong generalization capabilities. Contrastive VLMs encode the semantics of vision embeddings by learning to align them with textual embeddings. These models, however, lack an uncertainty quantification of the resulting alignment, leading to unreliable results when monitoring an environment's symbolic state. To address this limitation, we propose a method that integrates vision-language embeddings into a Bayesian framework for classification. Our approach constructs a representational subspace of the embeddings that captures the symbolic states to be monitored and learns Gaussian distributions on a small set of scene-agnostic support data. We model the symbolic state estimation as a classification problem with a Naive Bayes classifier, which provides information on the classification confidence and enables the integration of informative priors into the estimation process. We demonstrate the effectiveness of our method through two proof-of-concept experiments utilizing real-world data from kitchen environments.**

I. Introduction

In pursuit of full autonomy, robotic systems must understand the symbolic state of their environment and its objects. The symbolic state is the semantic interpretation of what the robot perceives to reason about the environment's affordances and constraints for task execution. A cooking robot that manipulates a pot must, for instance, monitor whether the appliance is *closed or open*, or whether the *water inside is already boiling*, a task illustrated in Fig. 1, to decide how to proceed in the cooking process.

In recent years, vision-language models (VLMs) have facilitated this goal by grounding symbolic states to visual input to provide robots with open-world understanding of novel scenes. VLMs based on contrastive encoders, such as CLIP [1] and SigLIP [2], for instance, assign meaning to vision features by learning a shared embedding space of image and text embeddings from large-scale data where semantically similar entities are closer aligned. The learned encoders can be applied as a frozen backbone to estimate a scene's symbolic state (e.g. in [3]) by relying on the cosine similarity between the visual embeddings and the text embeddings that describe each state (such as *"a picture of an open pot"* or *"a picture of a closed pot"*) to obtain the label of new image observations.

[1]German Aerospace Center (DLR), Robotics and Mechatronics Center (RMC), Münchener Str. 20, 82234 Weßling, Germany.[2]Jouhou System Kougaku Laboratory (JSK), Department of Mechano-Informatics, Graduate School of Information Science and Technology, The University of Tokyo.

While this approach removes the requirement to fine-tune models on a large data set for downstream tasks, it has three limitations: First, the cosine similarity is prone to assigning high confidence scores to misclassifications [4] and cannot provide reliable confidence estimates on the downstream class. Second, classification typically requires single text descriptions per class to infer alignment with the image embedding, thus making the success of the method dependent on the adequate phrasing of the text label. And third, it limits the class labels to a textual modality only, ignoring potentially useful supervision data in other modalities such as images or audio. These problems are confounded by the so-called *modality gap*, i.e. a phenomenon observed in CLIP-like models that shows that text and vision embeddings of VLMs are embedded in distinct spaces of their shared representation space [5].

To overcome these issues, we propose a method that facilitates the uncertainty-aware estimation of an environment's symbolic state from sensor data and weakly supervised class labels. Since we do not have any information about the scene we are observing, we utilize a small data set of support data that represents each symbolic state to train a Gaussian Naive Bayes classifier. We demonstrate how we can use a very small data set of 5 to 10 data points per class to weakly supervise the state estimation model.

We describe our proposed method in detail in Section II. In Section III, we provide two proof-of-concept experiments to validate our method: one on the DROID data set [6] and one on the water boiling task in the PR2 cooking data set from previous work [3]. In both examples, we show how our method enables a robot to detect symbolic state changes and estimate its confidence on scenes it has never seen before. Our proposed method provides more robust estimations than directly using cosine similarity with text labels, and adds a quantification of its certainty. We conclude by discussing future research directions in Section IV and related work in Section V.

II. Method

In order to make meaningful decisions in a new environment, autonomous robots must have the ability to interpret the symbolic state of their environment without prior exposure. We propose a method that computes low-dimensional feature vectors from the robot's sensor readings and classifies these into discrete classes. Our method relies on a small set of scene-agnostic support data. It consists of three steps:

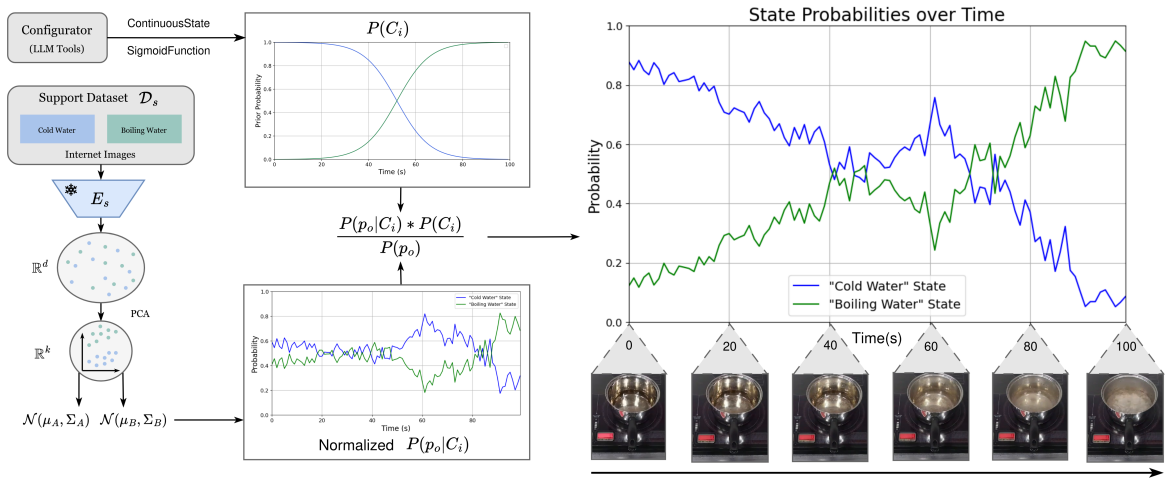


Fig. 1: We propose a Gaussian Naive Bayes classifier for estimating the symbolic state of novel scenes. Our proposed method can integrate an informed prior $P(C_i)$ (middle top) to compute the probability over time for each state $P(C_i|p_o)$ (right).

(1) We encode the small data set of support data into the shared vision-language embedding space of a contrastive encoder, such as SigLIP. The support data consists of either (a) descriptive text samples or (b) images of similar situations, not including the current scene or objects. In the shared embedding space, semantically similar entities are closer aligned. We can apply Principal Component Analysis (PCA) to find the latent sub-space that clusters the support data embeddings by the symbolic states we want to monitor, due to the variance in meaning of the embeddings.

(2) We fit a Gaussian distribution per symbolic state class on the support data. In Fig. 1, this represents $P(p_o|C_i)$, i.e. the probability of a new observation to belong to class C_i. We then train a Gaussian Naive Bayes classifier that computes the posterior $P(C_i|p_o)$, i.e. the probability of class C_i given the current observation, as our state estimation model. This enables us to estimate a discrete class label for the new observation by selecting the class that maximizes this probability and additionally to model the class uncertainty and the data typicality $P(p_o)$ explicitly due to the Bayesian formulation, which can support robot decision-making in downstream tasks.

(3) Finally, our method supports the integration of situational inductive biases via a class prior $P(C_i)$. As shown in the example in Fig. 1, we can explicitly model the time when a pot of water is expected to boil and use this to update the posterior. We furthermore propose a configurator that utilizes a large language model (LLM) to autonomously generate both the class priors and the support data set.

Fig. 2 shows an overview of our proposed method which is described in more detail in the next subsections.

A. Creating low-dimensional support data

Let $\mathcal{D}_s$ be a set of support data that represents the possible symbolic states of the environment. For simplicity, we assume that the environment can only exist in two binary states A and B. When using text labels as support data, the set would consist of synonyms and antonyms describing these states, such as *"an open pot"* and *"a closed pot"*. In the case of images as support data, $\mathcal{D}_s$ would contain image examples of the state we want to monitor, such as images of a pot with cold water and boiling water. Importantly, we do not need any support data specific to the observed scene and can apply the proposed method to any novel scene zero-shot. Let $\mathcal{D}_o$ be a set of sensor observations of the scene, such as a camera stream. The support data and sensor observations are embedded into a shared representation space using the modality-specific encoders E_s and E_o of a contrastive VLM, such as SigLIP. The resulting embeddings $f_s = E_s(s),\ s \in \mathcal{D}_s$ and $f_o = E_o(o),\ o \in \mathcal{D}_o$ are 768-dimensional feature vectors. Since we are only concerned with classifying the support data - and any new image observation - into symbolic state A or B, we want to reduce the high-dimensional vectors to a representational sub-space that extracts the dimension representing the semantics of the symbolic states. We apply PCA to find the k principal components that group the support data by its semantics. As shown in Section III, this clustering arises naturally within the data without requiring labeled data, as variance results from the contrasting meaning of the embeddings. We can project the support data embeddings $f_s \in \mathbb{R}^d$ to the fitted sub-space to receive $p_s = W^\top (f_s - \mu) \in \mathbb{R}^k$ where $k \ll d$. The same transformation can be applied to the sensor observations to get the projected embeddings $p_o \in \mathbb{R}^k$.

B. Learning a probabilistic model with Gaussian Naive Bayes

For classifying the symbolic state of a novel scene observation, we employ the Naive Bayes algorithm. We assume that the symbolic states can be modeled by two Gaussian distributions in the sub-space $\mathbb{R}^k$ and fit two Gaussian models $\mathcal{N}(\boldsymbol{\mu_A}, \boldsymbol{\Sigma_A})$ and $\mathcal{N}(\boldsymbol{\mu_B}, \boldsymbol{\Sigma_B})$ to the reduced support data embeddings $\{p_{s,A}\} \in \mathcal{D}_{s,A}$ and $\{p_{s,B}\} \in \mathcal{D}_{s,B}$ with $\mathcal{D}_s = \mathcal{D}_{s,A} \dot\cup \mathcal{D}_{s,B}$ respectively.

When receiving a novel scene observation p_o, we compute the likelihood that the observation belongs to the i-th symbolic state (i.e., $P(p_o|C_i),\ i \in [A, B]$) using the probability density

function of the Gaussian distribution $\mathcal{N}_i$. Subsequently we derive the posterior of each class $P(C_i|p_o)$ using Bayes rule:

$$P(C_i|p_o) = \frac{P(p_o|C_i) * P(C_i)}{P(p_o)} \quad (1)$$

where $P(C_i)$ is a prior that models the probability of the class and $P(p_o)$, which represents the typicality of the data point, is the normalization factor computed by the law of total probability $P(p_o) = P(p_o|C_A)P(C_A) + P(p_o|C_B)P(C_B)$.

In certain situations, it is sufficient to assume a constant prior $P(C_A) = P(C_B) = 0.5$ as there are no prior information regarding the symbolic state the environment is likely in.

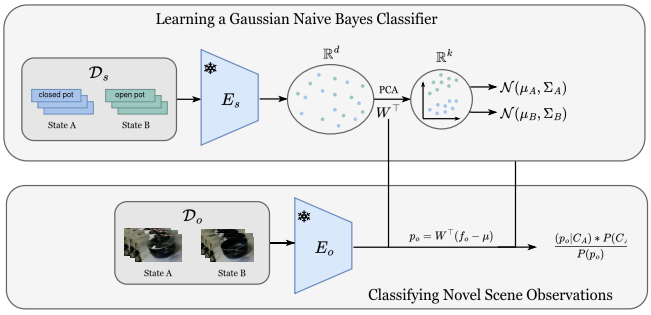


Fig. 2: Method overview. **Top.** For learning a symbolic state classifier, we choose a small data set $\mathcal{D}_s$ of support data. We embed the support data with a frozen VLM to a high-dimensional embedding space and learn a reduced embedding space $\mathbb{R}^k$ on our data points. Finally, we can train two Gaussian distributions on the reduced embeddings. **Bottom.** When receiving novel scene observations $\mathcal{D}_o$, we reduce the embeddings to the same space $\mathbb{R}^k$ and compute the posterior of the learned Gaussian distributions for state classification.

C. Extending Naive Bayes with an informative prior

A new challenge arises when looking at symbolic states that change continuously over time. These states - and especially their transitional phases - cannot be fully modeled by support data. In many of these cases, we can however make assumptions on the expected state change over time. For example, looking at the boiling water scenario, we can expect the water to be cold at the beginning and roughly know the expected time it takes for the water to start boiling [1]. We utilize this information by replacing the constant prior with an informative time-dependent prior when investigating a sequence of scene observations as shown in Fig. 1. The influence of the prior can be adapted with a weighing factor α.

Generating metadata with an LLM configurator: Instead of hand-crafting informative priors, we can leverage the common sense reasoning of LLM models to generate them autonomously. We call the module that provides this information a *configurator*. As a proof-of-concept, the configurator selects one prior out of a set of prior templates in our examples, using the LLM tool call functionality[2]. It first chooses whether the expected state change is continuous or discrete, and in the continuous case, selects the expected profile of the state change *linear vs. sigmoid vs. sinusoidal*. We hypothesize that the configurator can also provide support data and provide an example in Section III-B.

[1]We note that this depends on the specific context, e.g. if the stove is induction or electricity based.

[2]We use a locally-running Mistral Small 3 model.

III. Experiments

We test our proposed method by conducting two proof-of-concept experiments: one on the DROID data set [6] and one on the cooking data presented by Kawaharazuka et al. in previous work [3], to classify discrete and continuous symbolic states respectively.

A. Experiment 1: Boiling or not boiling water

One typical use case of symbolic states that change continuously over time are cooking scenarios, where it is important to determine the current stage of the food preparation process. To test our proposed method for this application, we investigate how well it classifies the *"boiling water"* scenario that was presented in [3], looking at a video stream of scene observations. In that work, state estimation required training an optimization algorithm on previous episodes with the same scene and objects.

Experiment: As support data, we use Internet images that depict the two states we are interested in - cold water and boiling water in a pot - and choose 10 images for each state. We note that the choice of images was arbitrary, and that they do not include the observed scene or objects. We first encode the support images with SigLIP-2 [7], fit PCA with $k = 2$ principal components to the embeddings and finally train a Naive Bayes classifier on the small data set (10 images per class). The configurator decides that the observed *"boiling water"* state is expected to change continuously over time, and chooses a sigmoid function ranging from 0 to 1 as the optimal informative prior to model the state change. The sigmoid $\sigma(x) = \frac{1}{1+e^{-a(x-x_0)}}$ uses two hyper-parameters (predefined by a human): a, defining the steepness of the curve, and x_0 as the point of the curve's largest steepness. The prior is weighted by $\alpha = 0.2$ to compute the posterior.

During inference, we employ an object detector to extract a cropped image of the *"pot"* from each video frame. We use the LLMDet open-vocabulary object detector for our experiments [8]. The cropped images are encoded by the SigLIP-2 model and transformed to a reduced embedding space using the fitted PCA model with $k = 2$.

Results: Fig. 1 shows the probability of the two states at different points in time in the video. The figure also illustrates the effect of the informative prior (top left). Using a constant prior ($P(C_A) = P(C_B) = 0.5$) results in the depicted likelihood function (bottom left), normalized such that $P(p_o|C_A) + P(p_o|C_B) = 1$. In contrast, the posterior (right) incorporates the informative prior to increase or decrease the model's certainty at different time steps.

Comparison with cosine similarity: In comparison, Fig. 3 visualizes how the cosine similarity between vision embeddings and different text embeddings of synonyms (blue) and antonyms (green) change over time. The similarity score is highly dependent on the choice of the text descriptions and does not provide a clear classification of the state.

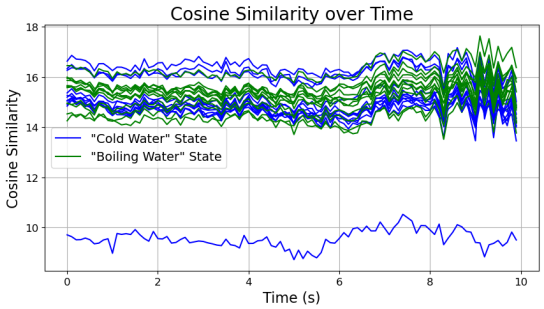


Fig. 3: Cosine similarity between vision and text embeddings of multiple synonyms (blue) and antonyms (green) over time.

B. Experiment 2: Open or closed pot

We select an episode from the DROID data set in which a robotic arm opens a pot in a kitchen environment and assess if our proposed method can correctly classify the pot as either *"closed"* or *"open"* in each video frame.

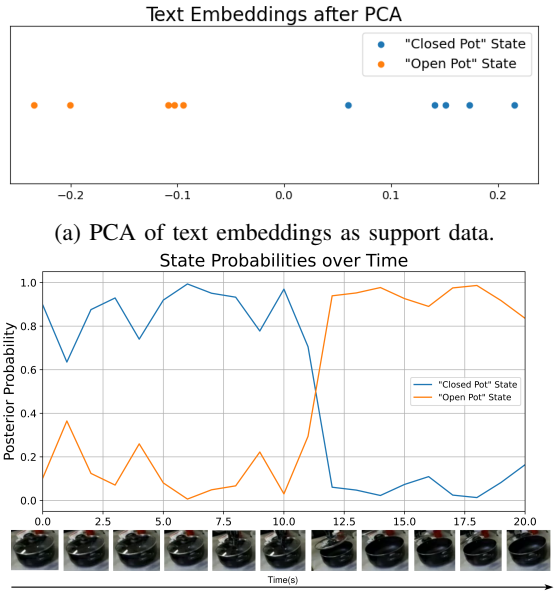
Experiment: The configurator predicts the observed state to be discrete and best modeled with a constant prior of 0.5 for each class, i.e. there are no prior assumptions that can be made about whether the observed pot in the novel scene is more likely to be *"closed"* or *"open"*. Instead of hand-crafting the support data set, our configurator autonomously generates suitable text descriptions for classification, choosing the following 5 synonyms and 5 antonyms of the symbolic state: *[[**S:** "a closed pot", "a sealed pot", "a shut pot", "a locked pot", "an fastened pot"], [**A:** "open pot", "uncovered pot", "unsealed pot", "unlocked pot", "loose pot"]]*. The generated text descriptions are encoded with SigLIP-2 and we apply PCA to the text embeddings to compute the $k = 1$ principal component of the data. As shown in figure 4a, the reduced embedding space clusters the text descriptions by their meaning. We finally train a Naive Bayes classifier on the small data set of reduced text embeddings. The resulting classifier is agnostic of any image observations or scene information.

During inference, we generate cropped images of the pot for each video frame and compute the posterior after generating reduced embeddings with SigLIP-2 and the fitted PCA model.

Results: We determine the probability of each image observation to represent a *"closed pot"* or an *"open pot"*. The resulting probabilities and their change over time are visualized in figure 4b.

IV. Discussion and Future Directions

In this work, we present a novel method for symbolic state estimation with VLMs in unknown scenes. Our method relies on a small set of scene agnostic support data with 5 to 10 data points per class and learns a Naive Bayes classifier. The Bayesian formulation provides a certainty quantification for the state classification and supports the seamless integration of informative priors into the estimation process. We demonstrate how a LLM-based configurator can be used to autonomously select informative priors and support data for the state estimation.



(a) PCA of text embeddings as support data.

(b) Symbolic state probabilities over time.

Fig. 4: Experimental results for the *"open/closed pot"* example.

In future work, we want to benchmark our proposed classifier and configurator on a large-scale data set and evaluate its robustness with more conclusive metrics, such as AUC-PR. We additionally want to extend the configurator by more complex tools. Useful features include the generation of support data sets of images, e.g. by searching the Internet for suitable images, and the parameterization of the prior functions. Furthermore, we are interested in exploring support data from other modalities. Since our method is modality-agnostic and can be used with any contrastive embedding model that was trained to project two or more different modalities onto the same embedding space, we want to investigate models such as AudioCLIP [9] or ImageBind [10] as alternative modalities.

V. Related Work

Traditional state estimation methods typically rely on hand-crafted heuristics or require fine-tuning state classifiers on each novel scene [11]. In contrast, autoregressive VLMs and LLMs have been widely applied to task planning and state estimation in unknown scenes where they have demonstrated strong generalization capabilities [12]–[14]. These works, however, focus on identifying task failures and do not consider the uncertainty of failure state predictions. One common limitation of VLMs is their overconfidence, leading to misclassifications and false predictions. To overcome this limitation in contrastive VLMs, previous work has extended these models by training probabilistic adapters and embeddings [15]–[17] or by integrating the retrieved embeddings into a Bayesian framework during test-time for uncertainty estimation [4], [18]. Our method extends these approaches to learn distributions on support data samples for classifying symbolic states with uncertainty estimation, supporting cross-modal training samples and the integration of informative priors.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[2] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, " Sigmoid Loss for Language Image Pre-Training ," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2023, pp. 11 941–11 952. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01100

[3] K. Kawaharazuka, N. Kanazawa, Y. Obinata, K. Okada, and M. Inaba, "Continuous object state recognition for cooking robots using pre-trained vision-language models and black-box optimization," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, p. 4059–4066, May 2024. [Online]. Available: http://dx.doi.org/10.1109/LRA.2024.3375257

[4] Z. Lin, M. Haghighat, W. Browne, and D. Miller, "Intra-class probabilistic embeddings for uncertainty estimation in vision-language models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2026, pp. 2327–2337. [Online]. Available: https://arxiv.org/abs/2511.22019

[5] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: understanding the modality gap in multi-modal contrastive representation learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022. [Online]. Available: https://arxiv.org/abs/2203.02053

[6] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, and et al., "DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. [Online]. Available: https://www.roboticsproceedings.org/rss20/p120.pdf

[7] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," 2025. [Online]. Available: https://arxiv.org/abs/2502.14786

[8] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, X. Xie, and W.-S. Zheng, "Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models," 2025. [Online]. Available: https://arxiv.org/abs/2501.18954

[9] A. Guzhov, F. Raue, J. Hees, and A. R. Dengel, "Audioclip: Extending clip to image, text and audio," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235624127

[10] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. [Online]. Available: https://arxiv.org/abs/2305.05665

[11] T. Migimatsu, W. Lian, J. Bohg, and S. Schaal, "Symbolic state estimation with predicates for contact-rich manipulation tasks," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE Press, 2022, p. 1702–1709. [Online]. Available: https://doi.org/10.1109/ICRA46639.2022.9811675

[12] J. Duan, W. Yuan, W. Pumacay, Y. R. Wang, K. Ehsani, D. Fox, and R. Krishna, "Manipulate-anything: Automating real-world robots using vision-language models," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 5326–5350. [Online]. Available: https://proceedings.mlr.press/v270/duan25a.html

[13] A. Mei, G.-N. Zhu, H. Zhang, and Z. Gan, "Replanvlm: Replanning robotic tasks with visual language models," *IEEE Robotics and Automation Letters*, vol. 9, pp. 10 201–10 208, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271571227

[14] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, T. Jackson, N. Brown, L. Luu, S. Levine, K. Hausman, and brian ichter, "Inner monologue: Embodied reasoning through planning with language models," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=3R3Pz5i0tye

[15] U. Upadhyay, S. Karthik, M. Mancini, and Z. Akata, "Probvlm: Probabilistic adapter for frozen vison-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 1899–1910. [Online]. Available: https://arxiv.org/abs/2307.00398

[16] S. Chun, W. Kim, S. Park, and S. Yun, "Probabilistic language-image pre-training," in *International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: https://arxiv.org/abs/2410.18857

[17] P. Morales-Álvarez, S. Christodoulidis, M. Vakalopoulou, P. Piantanida, and J. Dolz, "Bayesadapter: Enhanced uncertainty estimation in clip few-shot adaptation," *Int. J. Comput. Vision*, vol. 134, no. 2, Jan. 2026. [Online]. Available: https://doi.org/10.1007/s11263-025-02630-0

[18] Z. Han, J. Yang, G. Wang, J. Li, Q. Xu, M. Z. Shou, and C. Zhang, "DOTA: Distributional test-time adaptation of vision-language models," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=2T6QXSP8Cf