BEYOND UNDERSTANDING: LEARNING NEURO-SYMBOLIC REPRESENTATIONS FOR AGENTIC INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal foundation models learn powerful representations across heterogeneous modalities, but representations optimized for semantic understanding may be insufficient for agentic intelligence, where agents must reason over evolving environments, predict future states, and make goal-directed decisions. We introduce Neuro-Symbolic Multimodal Representation Learning for Agentic Intelligence, a paradigm that jointly learns neural and symbolic representations capturing semantic, relational, temporal, causal, and action-relevant structure. Our framework forms an evolving internal world representation in which symbolic structures guide neural learning, prediction, reasoning, and action, while new observations continuously refine the symbolic state. We propose an evaluation framework that shifts multimodal representation learning from semantic alignment toward prediction and decision utility, measuring compositional generalization, robustness to multimodal uncertainty, temporal prediction, and downstream decision-making. Our work establishes a foundation for multimodal representations designed not only for understanding the world, but for acting within it.

1 INTRODUCTION

The rapid progress of multimodal foundation models has enabled AI systems to perceive and reason across heterogeneous modalities, including vision, language, audio, depth, and sensor data Fei et al. (2022); Liang et al. (2024); Cai et al. (2026). By integrating these modalities, such models learn richer representations of the world and support advanced capabilities such as multimodal reasoning Xu et al. (2026), grounding Wu et al. (2025), and instruction following Ding et al. (2025), making multimodal representation learning a key foundation for next-generation intelligent systems. However, agentic AI requires more than multimodal perception and semantic reasoning Koutra et al. (2025). Agents must continuously perceive, model, predict, and act in dynamic environments, requiring representations that capture multimodal semantics, relations, temporal dynamics, and goal-relevant information. This raises a fundamental question: Are representations optimized for multimodal semantic understanding sufficient for agentic intelligence?

Current multimodal representation learning Cicchetti et al. (2025); Liu et al. (2026a); Qian et al. (2026); Gupta et al. (2026) primarily focuses on semantic alignment and cross-modal consistency, mapping heterogeneous observations into shared latent spaces that support recognition, retrieval, and understanding. However, semantic similarity does not necessarily imply similarity in decisions or actions. Visually similar situations may require different actions due to differences in temporal context, spatial relations, goals, or environmental constraints, while semantically different observations may demand the same action because they share relational or causal structure. Thus, agentic representations should capture not only semantics, but also relational, temporal, causal, and action-relevant structure.

The challenge is amplified by heterogeneous and unreliable sensory inputs, where modalities may provide complementary or conflicting evidence. Robust agents must reason over these uncertainties and their underlying relationships, yet simple embedding alignment or concatenation does not explicitly support such structured reasoning. Purely neural representations Chu & Shahabi (2026); Pel-

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.

legrino & Chadwick (2026); Haider et al. (2026) also struggle to encode rules, constraints, relations, and compositional knowledge, limiting their ability to support reliable, goal-directed decisions.

Neurosymbolic AI bridges this gap by combining neural models for multimodal representation learning with symbolic structures for explicit reasoning Marconato et al. (2026). However, existing approaches Cornelio et al. (2023); Lyu et al. (2026); Zhao (2026) often rely on predefined symbolic structures or treat symbolic reasoning as an additional layer, leaving open how symbolic structures can emerge from multimodal observations and jointly evolve with learned representations during agent–environment interaction. Agentic intelligence requires jointly learned neural and symbolic representations, where multimodal observations induce structured entities, relations, and states that guide reasoning and action. As environments evolve, neural observations continuously update the symbolic state, forming a bidirectional perception–knowledge loop.

Based on this perspective, we introduce Neuro-Symbolic Multimodal Representation Learning for Agentic Intelligence (NSMR), which learns an evolving world representation that integrates multimodal semantics with relational, temporal, predictive, and action-relevant information. NSMR connects perception, symbolic reasoning, world modeling, and action to support state understanding, relational reasoning, future-state prediction, and goal-directed decision-making in dynamic environments.

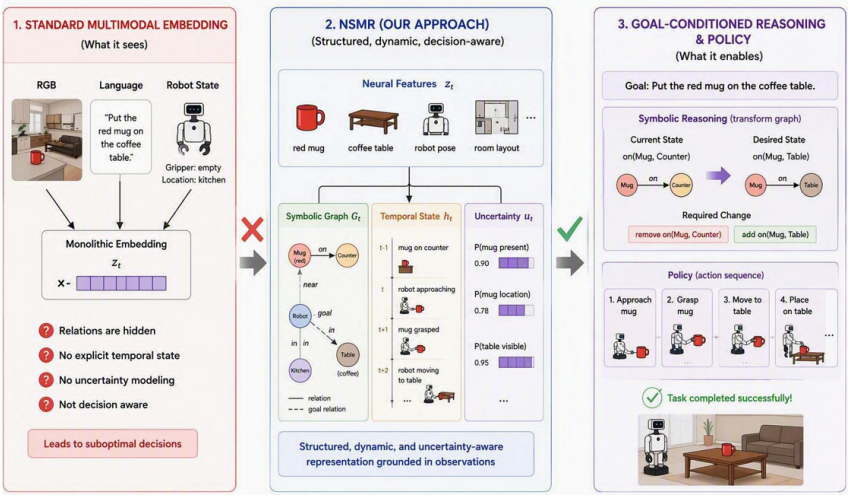

Figure 1: Conceptual comparison of conventional multimodal representations and NSMR: standard embeddings compress observations into a single vector, while NSMR structures perceptual, relational, temporal, and uncertainty information for decision-making.

Figure 1 illustrates the key motivation behind NSMR. Conventional multimodal agents compress heterogeneous observations into a single monolithic embedding, where entities, relations, temporal dynamics, and uncertainty are implicitly entangled. Such representations often obscure the structure required for reasoning, planning, and decision making. In the example task of placing a red mug on a coffee table, a standard embedding fails to explicitly represent relational facts such as on(mug, counter) or track how the environment evolves as actions are executed. We therefore propose Neuro-Symbolic Multimodal Representations (NSMR), a structured representation that jointly integrates neural perceptual features, symbolic entity-relation graphs, temporal world-state information, and uncertainty estimates. By maintaining explicit and goal-relevant structure, NSMR enables goal-conditioned reasoning and decision-aware representations that better support planning, generalization, and embodied decision making.

Our contributions are threefold. First, we formulate neuro-symbolic multimodal representation learning as a foundation for agentic AI, jointly preserving semantic, relational, temporal, and action-relevant information. Second, we propose a NSMR framework that jointly learns and updates neural and symbolic representations, enabling bidirectional interaction between perception and structured reasoning. Third, we develop an evaluation framework that measures compositional generalization, robustness to missing and conflicting modalities, temporal prediction, and downstream decision-making against conventional multimodal representations.

2 Related Work

Multimodal representation learning. Multimodal representation learning has primarily focused on aligning heterogeneous modalities within shared semantic spaces. Contrastive learning Khosla et al. (2020) and multimodal foundation models Bai et al. (2025); Ramachandran et al. (2026); Yang et al. (2025) have advanced cross-modal perception, grounding, and reasoning. However, these representations are predominantly optimized for semantic alignment and often lack explicit representations of relations, temporal states, uncertainty, and action consequences. Consequently, semantic similarity does not necessarily imply similarity in downstream decisions. NSMR addresses this limitation by learning representations that explicitly preserve structured and action-relevant information.

Neuro-symbolic multimodal reasoning. Neuro-symbolic learning combines neural perception with symbolic representations to improve compositionality, reasoning, and systematic generalization He & Li (2026); Mao et al. (2026). Scene-graph-based approaches similarly represent entities and their relations to support structured visual reasoning Gao et al. (2026); Vanica & Bors (2026). Recent multimodal systems increasingly incorporate symbolic reasoning or structured knowledge. Nevertheless, symbolic representations are commonly introduced as an auxiliary reasoning or post-processing component, rather than being jointly optimized with the multimodal representation itself. NSMR instead couples neural and symbolic representations through a bidirectional learning objective, allowing symbolic structure to directly shape the learned multimodal state.

World models and multimodal agents. World models learn latent environmental states that support prediction and planning Zhang et al. (2026); Bi et al. (2026), while multimodal agents combine foundation models with memory, planning, tool use, and interaction to perform increasingly complex tasks Liu et al. (2026b); Cheng et al. (2026). However, these approaches typically operate on representations learned independently of downstream action requirements. NSMR instead learns a temporally evolving structured state. The resulting representation is explicitly conditioned on goals and optimized for future state and action prediction.

Positioning. Our work differs from prior approaches by learning the representation itself as a structured, temporal, and action-relevant state. NSMR jointly optimizes neural features, symbolic relations, temporal dynamics, uncertainty, and action relevance, enabling multimodal representations to support perception, reasoning, and long-horizon decision-making.

3 Methodology

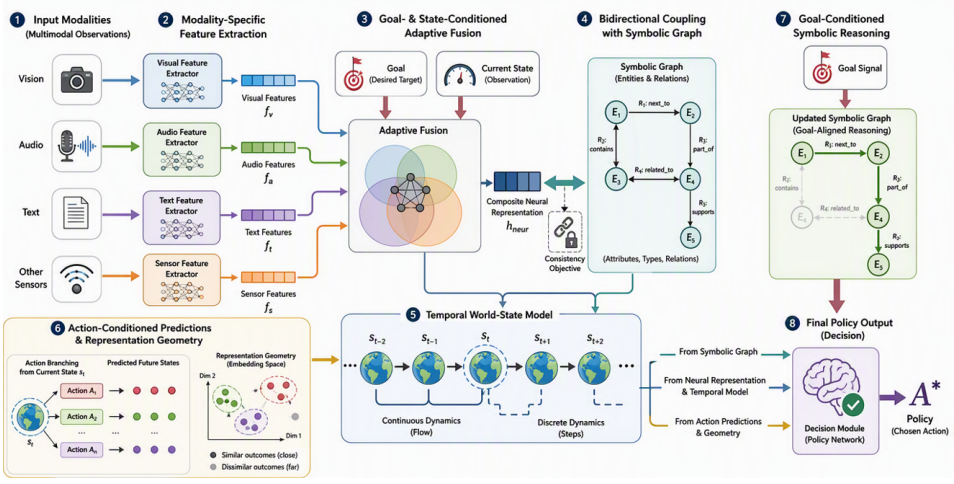

Figure 2: NSMR replaces monolithic embeddings with structured perceptual, symbolic, temporal, and uncertainty information for goal-conditioned, action-aware decision-making.

3.1 Problem Formulation and Overall Architecture

We consider an interactive agent operating in a multimodal environment. At each time step t, the agent receives a set of heterogeneous observations

$$\mathbf{X}_t = \left\{ x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(M)} \right\}, \tag{1}$$

where $x_t^{(m)}$ denotes an observation from modality m, such as vision, language, audio, LiDAR, radar, depth, or other sensor streams. The agent is additionally conditioned on a goal g_t and maintains an internal representation of the environment accumulated from previous observations and actions. The objective is not merely to learn a shared embedding of $\mathbf{X}_t$, but to learn a representation that supports perception, relational reasoning, prediction, and goal-directed action. We therefore define the proposed representation as

$$\mathcal{Z}_t = (\mathbf{z}_t, \mathcal{G}_t, \mathbf{h}_t, \mathbf{u}_t), \tag{2}$$

where $\mathbf{z}_t$ is a continuous neural representation, $\mathcal{G}_t$ is a symbolic relational representation, $\mathbf{h}_t$ is a temporal world state, and $\mathbf{u}_t$ captures uncertainty associated with the inferred state. The agent then produces an action according to

$$a_t \sim \pi_\theta (\mathcal{Z}_t, g_t). \tag{3}$$

The central objective is therefore to learn an encoder with

$$f_\theta : (\mathbf{X}_{1:t}, a_{1:t-1}, g_t) \to \mathcal{Z}_t \tag{4}$$

such that $\mathcal{Z}_t$ captures information useful for predicting future states and selecting actions.

As illustrated in Figure 2, we propose a Neuro-Symbolic Multimodal Representation (NSMR) framework that replaces conventional monolithic multimodal embeddings with a structured representation that jointly captures neural perceptual features, symbolic entities and relations, temporal world state, and uncertainty. Given heterogeneous multimodal observations, the framework first extracts modality-specific representations and performs goal- and state-conditioned adaptive fusion, allowing the agent to dynamically emphasize the information most relevant to its current objective. The resulting neural representation is coupled bidirectionally with a symbolic graph that explicitly represents entities, attributes, and relations, while a consistency objective ensures that the symbolic structure remains grounded in the underlying neural representation. A latent world-state model further maintains temporal dependencies and predicts both continuous state evolution and discrete relational changes, enabling the representation to capture environment dynamics rather than isolated observations. In addition, NSMR is explicitly optimized for decision-making by predicting action-conditioned future states and imposing an action-aware representation geometry in which states are close when they lead to similar decisions and consequences. Finally, goal-conditioned symbolic reasoning transforms the relational graph according to the current objective and combines the resulting reasoning state with neural and temporal information to produce the final policy.

3.2 Neuro-Symbolic Multimodal Representation

Our key idea is to replace the conventional monolithic multimodal embedding with NSMR that jointly maintains continuous neural features, discrete symbolic structures, temporal state, and uncertainty:

$$\mathcal{Z}_t = \left[\mathbf{z}_t^N, \mathcal{G}_t^S, \mathbf{h}_t^D, \mathbf{u}_t \right]. \tag{5}$$

Here, $\mathbf{z}_t^N$ captures perceptual and semantic information, $\mathcal{G}_t^S$ represents entities and relations, $\mathbf{h}_t^D$ captures temporal dependencies, and $\mathbf{u}_t$ represents uncertainty in the inferred state. Each modality is first encoded into a modality-specific latent space:

$$\mathbf{e}_t^{(m)} = f_{\theta_m} \left(x_t^{(m)} \right), \qquad m = 1, \ldots, M. \tag{6}$$

Rather than concatenating these embeddings, we employ goal- and state-conditioned adaptive fusion:

$$\alpha_t^{(m)} = \frac{\exp \left(q_\theta(\mathbf{e} * t^{(m)}, g_t, \mathbf{h} * t - 1) \right)}{\sum_{j=1}^{M} \exp \left(q_\theta(\mathbf{e} * t^{(j)}, g_t, \mathbf{h} * t - 1) \right)}, \tag{7}$$

where $\alpha_t^{(m)}$ measures the relevance of modality m to the current goal g_t. The neural representation is then

$$\mathbf{z}_t^N = \sum_{m=1}^{M} \alpha_t^{(m)} \mathbf{e}_t^{(m)}. \tag{8}$$

The fusion weights additionally incorporate the previous symbolic state:

$$\alpha_t = q_\theta \left(\mathbf{e} * t^{(1:M)}, g_t, \mathcal{G} * t - 1^S, \mathbf{h}_{t-1} \right), \tag{9}$$

allowing symbolic knowledge to influence modality selection. NSMR establishes a bidirectional interaction between neural and symbolic representations:

$$\mathbf{z}_t^N \leftrightarrow \mathcal{G}_t^S. \tag{10}$$

Specifically, the neural representation is transformed into a structured symbolic representation of entities, attributes, and relations:

$$\mathcal{G}_t^S = (\mathcal{V}_t, \mathcal{E}_t, \mathcal{R}_t), \tag{11}$$

where $\mathcal{V}_t$, $\mathcal{E} * t$, and $\mathcal{R} * t$ denote entities, attributes, and relations, respectively. For two entities v_i, v_j, the relation distribution is predicted as

$$p * \theta(r * ij \mid \mathbf{z}_i, \mathbf{z}_j, g_t) = \mathrm{softmax}\left(\phi_\theta(\mathbf{z}_i, \mathbf{z}_j, g_t)\right). \tag{12}$$

The induced symbolic structure then constrains and refines $\mathbf{z}_t^N$, making symbolic knowledge an integral part of the learned representation. This coupling preserves rich perceptual information while explicitly encoding the relational and goal-relevant structure required for agentic decision-making.

3.3 Neuro-Symbolic Consistency

A central component of NSMR is the requirement that the neural and symbolic representations describe a consistent underlying environment. We therefore introduce a symbolic reconstruction operator

$$r_\psi : \mathcal{G}_t^S \to \widehat{\mathbf{z}}_t^N, \tag{13}$$

which reconstructs the neural representation implied by the symbolic structure. We define the neuro-symbolic consistency loss as

$$\mathcal{L}_{NS} = d\left(\mathbf{z}_t^N, r_\psi(\mathcal{G}_t^S)\right), \tag{14}$$

where $d(\cdot, \cdot)$ denotes a representation distance. This objective prevents the symbolic graph from becoming an independent reasoning artifact disconnected from the underlying multimodal observations. It instead encourages the symbolic representation to retain the information necessary to explain the neural state. The resulting bidirectional mapping is

$$\mathbf{z}_t^N \to \mathcal{G}_t^S \to \widehat{\mathbf{z}}_t^N, \tag{15}$$

creating a structured bottleneck through which semantically and relationally important information must remain consistent.

3.4 Temporal and World-State Representation

Agentic behavior requires representations that evolve consistently over time. We therefore introduce a latent world state

$$\mathbf{h}_t = F_\theta\left(\mathbf{h}_{t-1}, \mathcal{Z}_t, a_{t-1}\right). \tag{16}$$

The model predicts the future representation conditioned on the current state and action:

$$\widehat{\mathbf{h}}_{t+1} = F_\theta\left(\mathbf{h}_t, a_t\right). \tag{17}$$

The temporal consistency objective is

$$\mathcal{L}_{dyn} = d\left(\mathbf{h}_{t+1}, \widehat{\mathbf{h}}_{t+1}\right). \tag{18}$$

This encourages the representation to encode environment dynamics rather than merely describing individual observations. The symbolic state is also propagated through time:

$$\widehat{\mathcal{G}}_{t+1} = \Psi_\theta \left(\mathcal{G}_t^S, a_t, \mathbf{h}_t \right). \tag{19}$$

The predicted symbolic state can therefore be compared with the symbolic structure inferred from the next observation, producing an additional structural dynamics objective:

$$\mathcal{L}_G = D_{KL} \left(p(\mathcal{G}_{t+1}^S) \| p(\widehat{\mathcal{G}}_{t+1}^S) \right). \tag{20}$$

Consequently, the learned representation captures both continuous dynamics and discrete relational changes.

3.5 Action-Relevant Representation Learning

The central distinction between conventional multimodal representation learning and our approach is that the representation is explicitly optimized for downstream decisions. Given the current representation $\mathcal{Z}_t$, goal g_t, and candidate action a_t, the model predicts the future state:

$$p_\theta \left(\mathcal{Z}_{t+1} \mid \mathcal{Z}_t, a_t, g_t \right). \tag{21}$$

The action-relevant objective is therefore defined as

$$\mathcal{L}_{act} = -\log p_\theta \left(\mathcal{Z}_{t+1} \mid \mathcal{Z}_t, a_t, g_t \right). \tag{22}$$

This encourages the representation to retain information that determines the consequences of potential actions. More importantly, we define an action-aware representation geometry. Let $d_A(i, j)$ denote the difference between the optimal decisions associated with two states. We encourage the representation distance

$$d_Z(i, j) = \|\mathbf{z}_i - \mathbf{z}_j\| \tag{23}$$

to reflect their decision-relevant similarity:

$$\mathcal{L}_{geo} = |d_Z(i, j) - d_A(i, j)|. \tag{24}$$

This explicitly distinguishes our representation from purely semantic embedding spaces. Two states do not need to be close merely because they look semantically similar; they should be close when they induce similar decisions and future consequences.

3.6 Goal-Conditioned Symbolic Reasoning

The symbolic representation is dynamically interpreted according to the agent's current goal. Given

$$(\mathcal{G}_t^S, g_t), \tag{25}$$

the reasoning module derives a goal-conditioned state:

$$\mathcal{R}_t = \Gamma_\theta \left(\mathcal{G}_t^S, g_t \right). \tag{26}$$

The reasoning process can be expressed as a sequence of symbolic transformations:

$$\mathcal{G}_t^S \xrightarrow{r_1} \mathcal{G}_t^{(1)} \xrightarrow{r_2} \cdots \xrightarrow{r_K} \mathcal{G}_t^{(K)}, \tag{27}$$

where each r_k corresponds to a learned or predefined relational rule. The resulting symbolic state is combined with the neural and dynamic representations:

$$\mathbf{c}_t = \text{Fuse} \left(\mathbf{z}_t^N, \mathcal{R}_t, \mathbf{h}_t, g_t \right). \tag{28}$$

The policy is then defined as

$$\pi_\theta(a_t \mid \mathbf{c}_t). \tag{29}$$

This combines continuous perception with relational reasoning for action selection.

4 Experiments

We evaluate NSMR around the following hypotheses: i) **H1: Structured representation.** Explicit neuro-symbolic representations improve relational and compositional reasoning over conventional multimodal embeddings. ii) **H2: Temporal and robust representation.** Maintaining structured world states improves temporal prediction and robustness to incomplete or conflicting multimodal observations. iii) **H3: Agentic representation.** Representations optimized jointly for symbolic structure and action relevance improve sequential decision-making, particularly for long-horizon tasks.

Experimental Setup We evaluate NSMR across three complementary benchmarks: **CLEVRER** Yi et al. (2019) for structured and counterfactual visual reasoning, **NExT-QA** Xiao et al. (2021) for temporal and causal video reasoning, and **ALFWorld** Shridhar et al. (2020b) for goal-directed agentic decision-making. This progression evaluates NSMR from structured representation learning to temporal understanding and interactive action. We compare against Unified-GRPOYan et al. (2026), Laytrol Huang et al. (2026a), MiniCPM-V 4.5 Yu et al. (2026), Agent-x Ashraf et al. (2026), and Neuro-LENS Murtas et al. (2026). Detailed experimental settings can be found in the appendix part.

Multimodal Representation and Compositional Reasoning (H1) We first evaluate whether NSMR preserves relational structure beyond conventional semantic alignment using descriptive, explanatory, predictive, and counterfactual reasoning on CLEVRER. Our hypothesis is that semantic similarity does not necessarily imply decision-relevant similarity; NSMR addresses this by jointly learning neural representations and explicit symbolic entities and relations. Table 1 reports the results on CLEVRER. The results show that conventional multimodal methods remain competitive on descriptive questions, while the advantage of NSMR becomes larger for explanatory, predictive, and counterfactual reasoning. NSMR improves overall performance from 80.4% to 85.9%, a relative gain of approximately 6.8%, with the largest improvements in predictive and counterfactual reasoning.

Table 1: Results on CLEVRER. Accuracy (%, mean $\pm$ standard deviation over five seeds). Best results are shown in bold.

Method	Descriptive	Explanatory	Predictive	Counterfactual	Overall
Unified-GRPO	86.2 ± 0.7	71.4 ± 0.9	64.8 ± 1.1	58.7 ± 1.2	70.3 ± 0.8
Laytrol	88.5 ± 0.6	75.9 ± 0.8	69.3 ± 0.9	62.1 ± 1.0	73.9 ± 0.7
MiniCPM-V 4.5	91.7 ± 0.5	79.8 ± 0.7	73.6 ± 0.8	68.4 ± 0.9	78.4 ± 0.6
Neuro-LENS	90.4 ± 0.6	81.6 ± 0.7	77.2 ± 0.8	72.5 ± 0.9	80.4 ± 0.6
NSMR (Ours)	$\mathbf{93.1 \pm 0.4}$	$\mathbf{86.7 \pm 0.5}$	$\mathbf{83.9 \pm 0.6}$	$\mathbf{79.8 \pm 0.7}$	$\mathbf{85.9 \pm 0.5}$

We next evaluate whether NSMR can generalize to combinations of entities and relations that were not observed during training. We construct a compositional split by withholding selected entity–relation combinations from the training set. As illustrated in Table 2, NSMR achieves an unseen-composition accuracy of 82.6%, compared with 74.3% for the neuro-symbolic baseline. The smaller generalization gap indicates that explicit relational representations improve recombination to unseen configurations.

Table 2: Compositional generalization results. Higher unseen performance and smaller generalization gaps indicate better systematic generalization.

Method	Seen (%) ↑	Unseen (%) ↑	Gap (%) ↓
Unified-GRPO	91.4	61.7	29.7
Laytrol	93.2	65.8	27.4
MiniCPM-V 4.5	94.1	69.5	24.6
Neuro-LENS	92.8	74.3	18.5
NSMR (Ours)	**95.0**	**82.6**	**12.4**

Temporal Reasoning and Robustness (H2) The second experiment investigates whether NSMR can maintain a coherent representation when multimodal observations evolve over time or become incomplete and unreliable. We evaluate temporal and causal reasoning using NExT-QA (as shown in Table 3, followed by robustness experiments (as shown in Table 4). Table 3 shows that the improvement is particularly pronounced on temporal and causal questions. This supports the hypothesis that maintaining an explicit structured world state is beneficial when reasoning requires information from multiple observations rather than isolated frame-level recognition.

As shown in Table 4, NSMR exhibits a smaller degradation as modalities are removed. Under 50% missing modalities, NSMR retains 80.1% accuracy compared with 73.4% for the neuro-symbolic baseline. The largest gain occurs under conflicting observations, where uncertainty-aware fusion prioritizes reliable modalities.

Table 3: Results on NExT-QA. Accuracy (%, mean ± standard deviation over five seeds).

Method	Temporal	Causal	Descriptive	Overall
Unified-GRPO	61.8 ± 0.8	58.4 ± 0.9	72.6 ± 0.7	64.3 ± 0.7
Laytrol	65.7 ± 0.7	62.9 ± 0.8	75.8 ± 0.6	68.1 ± 0.6
MLLM	69.8 ± 0.6	66.5 ± 0.7	78.9 ± 0.5	71.7 ± 0.6
Neuro-LENS	72.4 ± 0.7	70.8 ± 0.7	79.6 ± 0.6	74.3 ± 0.6
NSMR (Ours)	$\mathbf{78.6 \pm 0.5}$	$\mathbf{77.9 \pm 0.6}$	$\mathbf{82.4 \pm 0.4}$	$\mathbf{79.6 \pm 0.5}$

Table 4: Accuracy results under multimodal degradation.

Method	Full	25% Missing	50% Missing	Corrupted	Conflicting
Unified-GRPO	78.6	70.2	60.4	65.8	57.1
Laytrol	81.4	73.8	65.2	69.7	61.8
MiniCPM-V 4.5	84.7	77.1	68.9	73.6	66.4
Neuro-LENS	86.1	80.2	73.4	78.1	71.9
NSMR (Ours)	**88.4**	**84.9**	**80.1**	**83.7**	**79.8**

Table 5: Agentic decision-making results on ALFWorld. Task success rate (%, mean ± standard deviation over five seeds).

Method	Short	Medium	Long	Overall
Unified-GRPO	71.8 ± 1.5	54.2 ± 1.7	38.6 ± 1.9	54.9 ± 1.4
Laytrol	75.4 ± 1.4	59.7 ± 1.6	43.8 ± 1.8	59.6 ± 1.3
MiniCPM-V 4.5	78.9 ± 1.3	64.1 ± 1.5	48.5 ± 1.7	63.8 ± 1.2
Agent-x	81.3 ± 1.2	68.7 ± 1.4	53.2 ± 1.6	67.7 ± 1.1
Neuro-LENS	83.6 ± 1.1	72.4 ± 1.3	59.8 ± 1.5	71.9 ± 1.0
NSMR (Ours)	$\mathbf{87.5 \pm 0.9}$	$\mathbf{78.9 \pm 1.1}$	$\mathbf{69.4 \pm 1.3}$	$\mathbf{78.6 \pm 0.9}$

Agentic Decision-Making (H3) The final experiment evaluates whether the proposed representation improves actual agentic behavior. We use ALFWorld to evaluate goal-directed tasks requiring multimodal perception, world-state reasoning, planning, and sequential action execution. As shown in Table 5, NSMR obtains an overall success rate of 78.6%, representing a 6.7 percentage-point improvement over the neuro-symbolic agent baseline. The improvement increases with planning horizon: NSMR improves over the neuro-symbolic baseline by 3.9 points on short-horizon tasks, 6.5 points on medium-horizon tasks, and 9.6 points on long-horizon tasks. This trend provides direct evidence for the proposed action-oriented representation hypothesis.

5 Conclusion

We presented Neuro-Symbolic Multimodal Representation Learning for Agentic Intelligence, a perspective that extends multimodal representation learning beyond semantic understanding toward goal-directed action. We argue that agentic intelligence requires representations that jointly capture semantic, relational, temporal, causal, and action-relevant information. Our proposed paradigm couples neural and symbolic representations through a bidirectional loop, allowing observations to update the knowledge of the structured world, while symbolic structure guides perception, prediction, reasoning, and action. Overall, our results demonstrate that integrating neural and symbolic information into a unified, temporal, and action-relevant representation can strengthen multimodal agents beyond conventional semantic representations. In addition, we outline an evaluation framework centered on prediction, compositional generalization, multimodal robustness, and decision-making. This perspective opens a broader research direction toward evolving neuro-symbolic world representations that enable agents not only to understand complex environments but to reason and act reliably within them.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Tajamul Ashraf, Amal Saqib, Hanan Gani, Muhra AlMahri, Yuhao Li, Noor Ahsan, Umair Nawaz, Jean Lahoud, Hisham Cholakkal, Mubarak Shah, et al. Agent-x: Evaluating deep multimodal reasoning in vision-centric agentic tasks. In *International Conference on Learning Representations*, volume 2026, pp. 49814–49874, 2026.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026.

Zhili Cheng, Ran Li, Jinyi Hu, Yuge Tu, Shiqi Dai, Shengding Hu, Yang Shi, Lei Shi, and Maosong Sun. Embodiedeval: Evaluate multimodal llms as embodied agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11420–11432, 2026.

Chen Chu and Cyrus Shahabi. Geo2vec: Shape-and distance-aware neural representation of geospatial entities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18985–18993, 2026.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations*, volume 2025, pp. 42128–42149, 2025.

Cristina Cornelio, Jan Stuehmer, Shell Xu Hu, and Timothy Hospedales. Learning where and when to reason in neuro-symbolic inference. In *The Eleventh International Conference on Learning Representations*, 2023.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1099–1109. IEEE, 2025.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022.

Ziqi Gao, Weikai Huang, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Generate any scene: Scene graph driven data synthesis for visual generation training. In *International Conference on Learning Representations*, volume 2026, pp. 124660–124702, 2026.

Sharut Gupta, Sanyam Kansal, Stefanie Jegelka, Phillip Isola, and Vikas K Garg. Canonicalizing multimodal contrastive representation learning. In *ICLR 2026 Workshop on Representational Alignment (Re {\textasciicircum} 4-Align)*, 2026.

Ali Haider, Muhammad Salman Ali, Maryam Qamar, Tahir Khalil, Soo Ye Kim, Jihyong Oh, Enzo Tartaglione, and Sung-Ho Bae. I-inr: iterative implicit neural representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4520–4528, 2026.

Hao-Yuan He and Ming Li. A learnability analysis on neuro-symbolic learning. *Advances in Neural Information Processing Systems*, 38:84923–84956, 2026.

Sida Huang, Siqi Huang, Ping Luo, and Hongyuan Zhang. Laytrol: Preserving pretrained knowledge in layout control for multimodal diffusion transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5113–5121, 2026a.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Usman Naseem, Chunyu Wang, Qi Dai, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer cross-modality representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5131–5139, 2026b.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1359–1367. IEEE, 2017.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Danai Koutra, Lifu Huang, Adithya Kulkarni, Temiloluwa Prioleau, Beatrice Wan Yuan Soh, Qingyun Wu, Yujun Yan, Yaoqing Yang, Dawei Zhou, and James Zou. Towards agentic ai for science: hypothesis generation, comprehension, quantification, and validation. In *ICLR 2025 Workshop Proposals*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Hemm: Holistic evaluation of multimodal foundation models. *Advances in Neural Information Processing Systems*, 37:42899–42940, 2024.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Yijun Liu, Wu Liu, Xiaoyan Gu, Yong Rui, Xiaodong He, and Yongdong Zhang. Lmagent: A large-scale multimodal agents society for multi-user simulation. *IEEE Transactions on Multimedia*, 2026b.

Kailun Lyu, Fu Zhang, Zehan Li, and Jingwei Cheng. From local perspective to global reasoning: A neuro-symbolic framework for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29844–29856, 2026.

Jiayuan Mao, Joshua B Tenenbaum, and Jiajun Wu. Building intelligent agents with neuro-symbolic concepts. *Communications of the ACM*, 69(2):69–79, 2026.

Emanuele Marconato, Samuele Bortolotti, Emile Van Krieken, Paolo Morettin, Elena Umili, Antonio Vergari, Efthymia Tsamoura, Andrea Passerini, and Stefano Teso. Symbol grounding in neuro-symbolic ai: A gentle introduction to reasoning shortcuts. *Journal of Artificial Intelligence Research*, 86, 2026.

Giulia Murtas, Veselka Boeva, and Elena Tsiporkova. Neuro-lens: A neuro-symbolic framework integrating incomplete background knowledge and deep learning. *Neurosymbolic Artificial Intelligence*, 2:29498732261469489, 2026.

Arthur Pellegrino and Angus Chadwick. Rnns perform task computations by dynamically warping neural representations. *Advances in Neural Information Processing Systems*, 38:3579–3612, 2026.

Chengxuan Qian, Shuo Xing, Li Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 152802–152824, 2026.

Rahul Ramachandran, Ali Garjani, Roman Bachmann, Andrei Atanov, Oğuzhan Kar, and Amir Zamir. How well does gpt-4o understand vision? evaluating multimodal foundation models on standard computer vision tasks. In *International Conference on Learning Representations*, volume 2026, pp. 136263–136307, 2026.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10737–10746. IEEE, 2020a.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020b.

George Vanica and Adrian Bors. Ov-sgt: Open vocabulary semantic graph transformer for scene graph generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1685–1694, 2026.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24710–24721. IEEE, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Haochen Wang, Zhendong Wang, Bin Lin, Hao Li, Xinyan Xiao, et al. Unified multimodal models as auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41903–41912, 2026.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14203–14214. IEEE, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, volume 2026, pp. 55660–55699, 2026.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, volume 2025, pp. 53285–53310, 2025.

Yizheng Zhao. Fast and faithful: Scalable neuro-symbolic learning and reasoning with differentiable fuzzy elˆ+ˆ+. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1987–1997, 2026.

A Detailed Experimental Setup

Benchmarks. We evaluate NSMR on three complementary benchmark families. **CLEVRER** is used to assess structured visual reasoning, entity–relation representation, predictive reasoning, and counterfactual reasoning. **NExT-QA** evaluates temporal and causal reasoning over video sequences, while **ALFWorld** evaluates goal-directed agentic behavior requiring sequential interaction with the environment. Together, these benchmarks provide a progression from relational representation learning to temporal reasoning and interactive decision-making.

Baselines. We compare NSMR with five representative classes of approaches: (1) conventional multimodal encoders, (2) multimodal transformers, (3) multimodal large language models (MLLMs), (4) multimodal agents, and (5) neuro-symbolic baselines. Competing methods use the same data splits and, whenever possible, comparable backbone architectures and parameter budgets.

Evaluation metrics. For CLEVRER and NExT-QA, we report task accuracy. For compositional generalization, we evaluate seen and unseen compositions and report the generalization gap:

$$\Delta_{\text{gen}} = A_{\text{seen}} - A_{\text{unseen}}. \tag{30}$$

For temporal reasoning, we report future-state prediction accuracy at multiple horizons. For ALFWorld, we report task success rate, cumulative reward, number of executed actions, and planning failure rate. Higher values are better for accuracy, success, and reward, while lower values are better for action count and failure rate.

Implementation. NSMR consists of modality-specific encoders, goal-conditioned multimodal fusion, symbolic structure induction, temporal state updating, and action-aware representation learning. All models are trained using the same preprocessing, training, validation, and test partitions. Hyperparameters are selected using the validation set only.

Hyperparameter settings. Unless otherwise stated, the following configuration is used across experiments. The modality encoders are initialized from pretrained backbones and jointly fine-tuned with the NSMR components. The neural representation dimension is 512, the symbolic graph contains up to 64 entities, and the temporal state dimension is 256. We use a batch size of 32, AdamW optimization with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and a maximum of 50 training epochs. A cosine learning-rate schedule with 5 warm-up epochs is used, with early stopping based on validation performance. The multimodal fusion module uses a softmax temperature of 0.1. The temporal prediction horizon is set to 5 steps during training and evaluated at horizons of 1, 2, 3, 5, and 10 steps. Symbolic relation prediction uses a three-layer MLP with hidden dimension 256. The action prediction head uses a two-layer MLP with hidden dimension 256. For robustness experiments, modality dropout is applied with probability 0.2, while corrupted modalities are generated using noise levels selected from the validation set. For ALFWorld, the maximum episode length is set to 50 actions, and the agent replans after each environment transition.

Reproducibility. Each experiment is repeated using five random seeds, and we report mean performance and standard deviation. Hyperparameter selection is performed exclusively on the validation set, and the test set is used only for final evaluation. The same preprocessing, evaluation protocol, and stopping criteria are applied to all competing methods.

Table 6: Default hyperparameter configuration for NSMR.

Hyperparameter	Value
Neural representation dimension	512
Temporal state dimension	256
Maximum symbolic entities	64
Relation predictor hidden dimension	256
Action predictor hidden dimension	256
Batch size	32
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Maximum epochs	50
Warm-up epochs	5
Fusion temperature	0.1
Temporal training horizon	5
Evaluation horizons	1, 2, 3, 5, 10
Modality dropout	0.2
ALFWorld maximum episode length	50
λ_{align}	1.0
λ_{NS}	1.0
λ_{temp}	0.5
λ_{act}	0.5
λ_{goal}	0.5
λ_{unc}	0.1

B ABLATION STUDY.

We remove individual components of NSMR to determine their contribution to the observed improvements.

Table 7: Ablation results on ALFWorld.

Variant	Reasoning	Prediction	Success	Reward	Failure
NSMR	**85.9**	**84.7**	**78.6**	**84.7**	**21.4**
w/o Symbolic Representation	78.1	76.2	68.9	73.1	31.1
w/o Temporal Representation	82.4	70.3	72.8	77.6	27.2
w/o Action-Aware Learning	84.0	81.6	70.7	75.4	29.3
w/o Goal Conditioning	83.7	80.9	73.4	78.8	26.6
w/o Uncertainty Modeling	84.5	82.1	75.1	80.2	24.9
w/o Adaptive Modality Selection	84.8	82.8	75.8	80.7	24.2

Removing the symbolic representation produces the largest degradation in relational reasoning and overall task success, confirming its importance for structured world-state construction. Removing temporal modeling causes the largest reduction in future prediction, while removing action-aware learning produces a substantial decrease in agentic success. These results suggest that the components address complementary aspects of agentic representation learning.

We examine whether the learned representation aligns observations according to their decision relevance rather than semantic similarity alone. We measure semantic similarity, relational similarity, temporal consistency, and action similarity.

Interestingly, the difference between NSMR and the strongest baselines is relatively small for semantic similarity but substantially larger for relational, temporal, and action similarity. This result supports the main argument of this work: improving multimodal representations for agentic AI requires optimizing the representation for structures that directly support reasoning and action rather than semantic alignment alone.

Table 8: Representation analysis. Higher values indicate stronger alignment between the corresponding representation property and the target structure.

Representation	Semantic	Relational	Temporal	Action
Unified-GRPO	91.2	67.5	58.9	54.7
Laytrol	92.8	71.3	62.4	59.1
MiniCPM-V 4.5	94.1	75.6	66.7	63.8
Neuro-LENS	92.7	82.4	75.9	72.1
NSMR (Ours)	**94.5**	**89.7**	**84.8**	**83.6**

C Theoretical Analysis

Our objective is not to claim that a particular neural architecture globally optimizes a non-convex training objective, but rather to characterize the properties that make the proposed representation suitable for agentic decision-making.

The analysis is based on four principles. First, a representation for an agent should preserve information relevant to future decisions rather than only semantic similarity. Second, symbolic structure can provide a lower-dimensional abstraction of relational information. Third, explicitly modeling temporal state reduces the accumulation of representation errors over long horizons. Finally, uncertainty-aware modality selection prevents unreliable modalities from dominating the learned state.

C.1 Problem Formulation

Consider an agent interacting with an environment over a finite horizon T. At time t, the environment is characterized by a latent state $s_t \in \mathcal{S}$. The agent receives a multimodal observation

$$o_t = \left(o_t^{(1)}, o_t^{(2)}, \ldots, o_t^{(M)}\right), \tag{31}$$

where $o_t^{(m)}$ denotes the observation from modality m. The agent maintains a goal $g_t \in \mathcal{G}$ and selects an action

$$a_t = \pi(z_t, g_t), \tag{32}$$

where z_t is the learned representation.

The environment evolves according to

$$s_{t+1} \sim P(\cdot \mid s_t, a_t). \tag{33}$$

The objective of the agent is to maximize the expected discounted return

$$J(\pi) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, g_t)\right], \qquad 0 < \gamma \leq 1. \tag{34}$$

A conventional multimodal representation is obtained as

$$z_t = f_\theta(o_t), \tag{35}$$

whereas NSMR constructs a structured state

$$\mathcal{Z}_t = (z_t, \mathcal{G}_t, h_t, u_t)\,. \tag{36}$$

Here, z_t denotes the neural multimodal representation, $\mathcal{G}_t$ is a symbolic graph containing entities and relations, h_t is a temporally accumulated state, and u_t represents uncertainty.

The corresponding policy is

$$a_t = \pi_\phi \left(z_t, \mathcal{G}_t, h_t, u_t, g_t \right). \tag{37}$$

C.2 Decision-Relevant Sufficiency

We first formalize the distinction between semantic representation and agentic representation.

Definition 1 (Decision-relevant representation) *A representation $\rho(o_{\leq t}, g_t)$ is called* decision-relevant sufficient *if there exists a policy π_ρ such that*

$$\pi^*(a_t \mid o_{\leq t}, g_t) = \pi_\rho(a_t \mid \rho(o_{\leq t}, g_t)), \tag{38}$$

where π^* denotes an optimal policy.

This definition differs from semantic sufficiency, which only requires the representation to preserve information necessary for predicting semantic labels.

Definition 2 (Semantic sufficiency) *A representation $z = f(o)$ is semantically sufficient for a target y if*

$$P(y \mid o) = P(y \mid z). \tag{39}$$

Semantic sufficiency does not necessarily imply decision sufficiency.

Theorem 1 (Semantic sufficiency does not imply decision sufficiency) *There exist environments for which a representation z is semantically sufficient for an observation-level target y, but is not sufficient for optimal decision-making.*

Proof 1 *Consider an environment containing two states s_1 and s_2 that produce the same semantic label,*

$$y(s_1) = y(s_2), \tag{40}$$

but require different optimal actions,

$$a^*(s_1) \neq a^*(s_2). \tag{41}$$

Suppose the representation is defined as

$$z(s) = y(s). \tag{42}$$

Then

$$P(y \mid s) = P(y \mid z), \tag{43}$$

so z is semantically sufficient. However,

$$z(s_1) = z(s_2), \tag{44}$$

while

$$a^*(s_1) \neq a^*(s_2). \tag{45}$$

Therefore no deterministic policy $\pi(z)$ can simultaneously select the optimal action in both states. Hence z is not decision-relevant sufficient.

Implication. The theorem establishes the central motivation for NSMR. Optimizing multimodal representations solely for semantic alignment can collapse states that are semantically equivalent but decision-theoretically different. Relational, temporal, goal-conditioned, and action-relevant information must therefore be preserved explicitly.

C.3 Sufficiency of the Neuro-Symbolic State

We next show that the proposed structured state can become sufficient for decision-making when it preserves the relevant environment information.

Let

$$\psi(s_t) = (z_t, \mathcal{G}_t, h_t, u_t) \tag{46}$$

denote the representation mapping from the latent environment state to the NSMR state.

[Structured state sufficiency] For any two environment states s and s',

$$\psi(s) = \psi(s') \tag{47}$$

implies

$$P(r_t, s_{t+1} \mid s, a, g) = P(r_t, s_{t+1} \mid s', a, g) \tag{48}$$

for every action a and goal g.

This assumption states that the structured representation preserves all information necessary to predict rewards and future states.

Theorem 2 (Policy sufficiency of NSMR) *Under Assumption C.3, there exists an optimal policy that depends only on the structured representation $\mathcal{Z}_t$ and the goal g_t.*

Proof 2 *The optimal action-value function satisfies the Bellman equation*

$$Q^*(s, a, g) = \mathbb{E}\left[r(s, a, g) + \gamma \max_{a'} Q^*(s', a', g') \mid s, a, g\right]. \tag{49}$$

Consider two states s and s' such that

$$\psi(s) = \psi(s'). \tag{50}$$

By Assumption C.3, the conditional distributions of the immediate reward and next state are identical for every action. Therefore their Bellman operators are identical:

$$Q^*(s, a, g) = Q^*(s', a, g) \tag{51}$$

for every a and g.

Consequently, the optimal action can be selected solely as a function of the representation:

$$\pi^*(s, g) = \arg\max_a Q^*(s, a, g) = \arg\max_a Q^*(\mathcal{Z}, a, g). \tag{52}$$

Thus there exists a policy

$$\pi^*_{\mathcal{Z}}(a \mid \mathcal{Z}, g) \tag{53}$$

that achieves the same optimal value as the policy operating directly on the environment state.

Interpretation. The theorem provides the theoretical role of the NSMR state: its purpose is not merely to improve classification accuracy, but to construct a compact state that preserves the information required by the downstream decision process.

C.4 Benefit of Action-Conditioned Representation Learning

We now formalize the main design principle of NSMR. Let $\mathcal{Z} = f_\theta(o)$ be a learned representation. Define the representation objective

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{sem}} + \lambda_{\text{act}} \mathcal{L}_{\text{act}}, \tag{54}$$

where $\mathcal{L}_{\text{sem}}$ measures semantic representation quality and $\mathcal{L}_{\text{act}}$ measures the ability of the representation to predict decision-relevant actions.

Definition 3 (Action-relevant information) *The action-relevant information contained in a representation Z is measured by*

$$I(Z; A \mid G), \tag{55}$$

where $I(\cdot;\cdot \mid \cdot)$ denotes conditional mutual information.

Theorem 3 (Action-aware representations preserve decision information) *Suppose that the action prediction loss is the conditional cross-entropy*

$$\mathcal{L}_{\text{act}} = \mathbb{E}\left[-\log p_\phi(A \mid Z, G)\right]. \tag{56}$$

At its optimum,

$$\mathcal{L}^*_{\text{act}} = H(A \mid Z, G), \tag{57}$$

and therefore

$$I(Z; A \mid G) = H(A \mid G) - \mathcal{L}^*_{\text{act}}. \tag{58}$$

Hence minimizing action prediction loss maximizes the action-relevant information retained by the representation.

Proof 3 *By definition, conditional entropy is*

$$H(A \mid Z, G) = \mathbb{E}\left[-\log p(A \mid Z, G)\right]. \tag{59}$$

The optimal predictor under cross-entropy is the true conditional distribution,

$$p^*_\phi(A \mid Z, G) = p(A \mid Z, G). \tag{60}$$

Therefore,

$$\mathcal{L}^*_{\text{act}} = H(A \mid Z, G). \tag{61}$$

Conditional mutual information satisfies

$$I(Z; A \mid G) = H(A \mid G) - H(A \mid Z, G). \tag{62}$$

Substituting the optimal action loss gives

$$I(Z; A \mid G) = H(A \mid G) - \mathcal{L}^*_{\mathrm{act}}. \tag{63}$$

Since $H(A \mid G)$ does not depend on the representation parameters, minimizing $\mathcal{L}_{\mathrm{act}}$ maximizes $I(Z; A \mid G)$.

Consequence. This result provides a formal justification for the action-aware component of NSMR. A representation trained only for multimodal semantic alignment does not explicitly maximize information relevant to future actions, whereas the proposed objective directly encourages such information to be retained.

C.5 Neuro-Symbolic Structure and Compositional Generalization

Let a symbolic environment be represented by a graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \tag{64}$$

where nodes $\mathcal{V}$ correspond to entities and edges $\mathcal{E}$ correspond to relations.

Suppose that a task can be expressed as a composition of primitive relations

$$R = R_1 \circ R_2 \circ \cdots \circ R_K. \tag{65}$$

A neural representation may memorize complete combinations, whereas a symbolic representation explicitly preserves the primitive relational components.

Theorem 4 (Compositional transfer under symbolic factorization) *Assume that the target decision function can be decomposed as*

$$F(G, g) = F_K \circ F_{K-1} \circ \cdots \circ F_1(G, g), \tag{66}$$

where each F_k corresponds to a primitive relation or transformation. If the learned symbolic representation correctly identifies each primitive relation F_k, then any unseen composition of previously observed primitives can be evaluated without requiring direct training examples of that complete composition.

Proof 4 *Let the training data contain examples of the primitive transformations $F_1, \ldots, F_K$, although not necessarily their complete composition.*

Because the symbolic representation explicitly factorizes the environment into primitive relations, the model learns estimators $\hat{F}_k$ for each primitive component.

For an unseen composition

$$F^* = F_{i_K} \circ \cdots \circ F_{i_1}, \tag{67}$$

the predicted output is

$$\hat{F}^* = \hat{F}_{i_K} \circ \cdots \circ \hat{F}_{i_1}. \tag{68}$$

Thus evaluation of the complete composition requires only the learned primitive functions and does not require a training example containing the same complete composition.

If every primitive function is correctly recovered, then

$$\hat{F}_{i_k} = F_{i_k} \tag{69}$$

for all k, and therefore

$$\hat{F}^* = F^*. \tag{70}$$

Hence the unseen composition can be solved without direct supervision on that particular composition.

Discussion. The theorem does not imply perfect systematic generalization in practical neural systems because symbolic induction itself may contain errors. Instead, it explains why explicitly factorizing relations can reduce the generalization gap: the model can recombine learned primitives rather than memorizing complete multimodal configurations.

C.6 Temporal Representation and Long-Horizon Prediction

Let the learned temporal state evolve according to

$$h_{t+1} = F_\theta(h_t, \mathcal{G}_t, z_t, a_t). \tag{71}$$

Suppose the learned transition function approximates the true transition function F^* with bounded error

$$\|F_\theta(x, a) - F^*(x, a)\| \leq \epsilon \tag{72}$$

for all states and actions.

We additionally assume that F^* is Lipschitz continuous:

$$\|F^*(x, a) - F^*(y, a)\| \leq L\|x - y\|. \tag{73}$$

Theorem 5 (Bounded temporal representation error) *Under the above assumptions, the k-step prediction error satisfies*

$$\|\hat{h}_{t+k} - h_{t+k}\| \leq \epsilon \sum_{i=0}^{k-1} L^i. \tag{74}$$

For $L \neq 1$,

$$\|\hat{h}_{t+k} - h_{t+k}\| \leq \epsilon \frac{L^k - 1}{L - 1}. \tag{75}$$

If $L < 1$, the error is uniformly bounded by

$$\|\hat{h}_{t+k} - h_{t+k}\| \leq \frac{\epsilon}{1 - L}. \tag{76}$$

Proof 5 *For one step,*

$$\begin{aligned} \|\hat{h}_{t+1} - h_{t+1}\| &= \|F_\theta(\hat{h}_t, a_t) - F^*(h_t, a_t)\| && (77)\\ &\leq \|F_\theta(\hat{h}_t, a_t) - F^*(\hat{h}_t, a_t)\| + \|F^*(\hat{h}_t, a_t) - F^*(h_t, a_t)\| && (78)\\ &\leq \epsilon + L\|\hat{h}_t - h_t\|. && (79) \end{aligned}$$

Define

$$e_t = \|\hat{h}_t - h_t\|. \tag{80}$$

Then

$$e_{t+1} \leq \epsilon + L e_t. \tag{81}$$

Applying this recursively gives

$$e_{t+k} \leq \epsilon + L\epsilon + L^2\epsilon + \cdots + L^{k-1}\epsilon \tag{82}$$

$$= \epsilon \sum_{i=0}^{k-1} L^i. \tag{83}$$

For $L \neq 1$,

$$\sum_{i=0}^{k-1} L^i = \frac{L^k - 1}{L - 1}. \tag{84}$$

If $L < 1$, then

$$\sum_{i=0}^{k-1} L^i \leq \sum_{i=0}^{\infty} L^i = \frac{1}{1 - L}. \tag{85}$$

This proves the result.

Interpretation. The result explains why temporal state modeling becomes increasingly important for long-horizon agents. Small representation errors can accumulate across multiple transitions. Learning a temporally coherent state and ensuring that the transition model is stable limits this accumulation.

C.7 Uncertainty-Aware Multimodal Fusion

Assume that each modality provides an estimator

$$x_t^{(m)} = x_t + \varepsilon_t^{(m)}, \tag{86}$$

where

$$\mathbb{E}[\varepsilon_t^{(m)}] = 0 \tag{87}$$

and

$$\mathrm{Var}\left(\varepsilon_t^{(m)}\right) = \sigma_m^2. \tag{88}$$

Consider a fused estimator

$$\hat{x} = \sum_{m=1}^{M} w_m x^{(m)}, \qquad \sum_m w_m = 1. \tag{89}$$

Theorem 6 (Optimal uncertainty-weighted fusion) *If the modality errors are independent and unbiased, the minimum-variance linear unbiased estimator is obtained using*

$$w_m^* = \frac{\sigma_m^{-2}}{\sum_{j=1}^{M} \sigma_j^{-2}}. \tag{90}$$

Its variance is

$$\mathrm{Var}(\hat{x}) = \left(\sum_{m=1}^{M} \sigma_m^{-2} \right)^{-1}. \tag{91}$$

Proof 6 *Because the estimator is unbiased whenever* $\sum_m w_m = 1$*, its variance is*

$$\mathrm{Var}(\hat{x}) = \sum_{m=1}^{M} w_m^2 \sigma_m^2, \tag{92}$$

where independence eliminates cross-covariance terms.

We minimize

$$\sum_m w_m^2 \sigma_m^2 \tag{93}$$

subject to

$$\sum_m w_m = 1. \tag{94}$$

The Lagrangian is

$$\mathcal{L} = \sum_m w_m^2 \sigma_m^2 + \lambda \left(\sum_m w_m - 1 \right). \tag{95}$$

Taking derivatives gives

$$2 w_m \sigma_m^2 + \lambda = 0, \tag{96}$$

hence

$$w_m = -\frac{\lambda}{2\sigma_m^2}. \tag{97}$$

Using the constraint,

$$-\frac{\lambda}{2} \sum_m \sigma_m^{-2} = 1, \tag{98}$$

which yields

$$w_m^* = \frac{\sigma_m^{-2}}{\sum_j \sigma_j^{-2}}. \tag{99}$$

Substitution gives

$$\mathrm{Var}(\hat{x}) = \left(\sum_m \sigma_m^{-2} \right)^{-1}. \tag{100}$$

Thus the optimal fusion assigns larger weights to more reliable modalities.

Connection to NSMR. The adaptive modality-selection component of NSMR can be interpreted as learning an approximation to these reliability-dependent weights. The uncertainty variable u_t therefore provides a principled mechanism for preventing noisy or contradictory modalities from dominating the multimodal state.

C.8 Neuro-Symbolic Consistency

Let

$$q_\theta(G \mid z) \tag{101}$$

denote the symbolic structure induced from the neural representation and let

$$p_\phi(z \mid G) \tag{102}$$

denote the neural representation predicted from the symbolic structure.

We define the bidirectional consistency objective as

$$\mathcal{L}_{\mathrm{NS}} = D_{\mathrm{KL}}\left(q_\theta(G \mid z) \| q^*(G \mid z)\right) + D_{\mathrm{KL}}\left(p_\phi(z \mid G) \| p^*(z \mid G)\right). \tag{103}$$

Proposition 1 (Consistency reduces representation ambiguity) *Suppose two neural representations z_1 and z_2 induce the same symbolic structure,*

$$q_\theta(G \mid z_1) = q_\theta(G \mid z_2), \tag{104}$$

and the bidirectional consistency objective is minimized. Then the representation is encouraged to identify the two observations according to their shared relational structure rather than treating their raw multimodal features independently.

Proof 7 *At the optimum of the consistency objective,*

$$q_\theta(G \mid z) \to q^*(G \mid z) \tag{105}$$

and

$$p_\phi(z \mid G) \to p^*(z \mid G). \tag{106}$$

If z_1 and z_2 induce the same symbolic structure G, both representations are constrained through the same symbolic bottleneck. Consequently, variations in the neural representation that do not affect the relational structure are penalized whenever they produce inconsistent symbolic predictions.

Therefore the consistency objective encourages the neural representation to preserve relationally meaningful information while suppressing irrelevant variation.

C.9 Joint NSMR Objective

The complete NSMR objective can be written as

$$\mathcal{L}_{\mathrm{NSMR}} = \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{NS}}\mathcal{L}_{\mathrm{NS}} + \lambda_{\mathrm{temp}}\mathcal{L}_{\mathrm{temp}} + \lambda_{\mathrm{act}}\mathcal{L}_{\mathrm{act}} + \lambda_{\mathrm{goal}}\mathcal{L}_{\mathrm{goal}} + \lambda_{\mathrm{unc}}\mathcal{L}_{\mathrm{unc}}. \tag{107}$$

The individual terms correspond respectively to multimodal alignment, neuro-symbolic consistency, temporal prediction, action prediction, goal-conditioned representation learning, and uncertainty estimation.

Theorem 7 (Multi-objective representation property) *Assume that the individual losses in Eq. equation 107 are bounded below and that their corresponding weights are non-negative. A global minimizer of $\mathcal{L}_{\mathrm{NSMR}}$ minimizes the weighted aggregate of semantic, symbolic, temporal, action, goal, and uncertainty objectives.*

Moreover, if the objectives have a common minimizer, any common minimizer is also a global minimizer of $\mathcal{L}_{\mathrm{NSMR}}$.

Proof 8 *Let*

$$\mathcal{L}_k(\theta) \geq L_k^* \tag{108}$$

for each objective k. Since all coefficients satisfy

$$\lambda_k \geq 0, \tag{109}$$

we have

$$\mathcal{L}_{\mathrm{NSMR}}(\theta) = \sum_k \lambda_k \mathcal{L}_k(\theta) \geq \sum_k \lambda_k L_k^*. \tag{110}$$

Therefore the global minimizer of the joint objective seeks a solution that minimizes the weighted combination of the individual objectives.

If there exists θ^* such that

$$\mathcal{L}_k(\theta^*) = L_k^* \tag{111}$$

for every k, then

$$\mathcal{L}_{\mathrm{NSMR}}(\theta^*) = \sum_k \lambda_k L_k^*, \tag{112}$$

which achieves the lower bound. Hence θ^* is a global minimizer of the joint objective.

C.10 Approximate Policy Performance

The previous results establish representation-level properties. We now connect representation error to policy performance.

Assume that the optimal action-value function is Lipschitz with respect to the structured state:

$$|Q^*(\mathcal{Z}, a, g) - Q^*(\mathcal{Z}', a, g)| \leq L_Q \|\mathcal{Z} - \mathcal{Z}'\|. \tag{113}$$

Let $\hat{\mathcal{Z}}$ be the learned representation and suppose

$$\|\hat{\mathcal{Z}} - \mathcal{Z}\| \leq \epsilon_{\mathcal{Z}}. \tag{114}$$

Theorem 8 (Representation error and decision quality) *Under the Lipschitz assumption above, the error in action-value estimation induced by representation error is bounded by*

$$\left| Q^*(\hat{\mathcal{Z}}, a, g) - Q^*(\mathcal{Z}, a, g) \right| \leq L_Q \epsilon_{\mathcal{Z}}. \tag{115}$$

Consequently, reducing structured representation error directly reduces the maximum induced error in action-value estimation.

Proof 9 *By the Lipschitz assumption,*

$$|Q^*(\hat{\mathcal{Z}}, a, g) - Q^*(\mathcal{Z}, a, g)| \leq L_Q \|\hat{\mathcal{Z}} - \mathcal{Z}\|. \tag{116}$$

Since

$$\|\hat{\mathcal{Z}} - \mathcal{Z}\| \leq \epsilon_{\mathcal{Z}}, \tag{117}$$

we obtain

$$|Q^*(\hat{\mathcal{Z}}, a, g) - Q^*(\mathcal{Z}, a, g)| \leq L_Q \epsilon_{\mathcal{Z}}. \tag{118}$$

Thus the result follows.

C.11 Overall Theoretical Implications

The theoretical results provide a unified interpretation of the proposed architecture.

First, Theorem 1 shows that semantic multimodal sufficiency alone is insufficient for agentic decision-making. This motivates augmenting conventional neural embeddings with information about relations, states, goals, and actions.

Second, Theorem 2 establishes that a structured representation can support optimal decision-making when it preserves the reward and transition-relevant information of the environment.

Third, Theorem 3 provides an information-theoretic justification for action-conditioned representation learning: minimizing action prediction loss maximizes the conditional information between the representation and future decisions.

Fourth, Theorem 4 explains why explicit symbolic factorization can improve systematic compositional generalization. Rather than memorizing complete multimodal configurations, the representation can recombine learned entities and relations.

Fifth, Theorem 5 shows that temporal representation errors can accumulate with the prediction horizon, providing a theoretical motivation for explicitly maintaining a temporally coherent world state.

Finally, Theorem 6 establishes the principle underlying uncertainty-aware modality selection: unreliable observations should contribute less to the final multimodal state.

Together, these results support the central premise of NSMR:

> *A multimodal representation for agentic intelligence should preserve not only what the environment contains, but also how entities are related, how the environment evolves, what is uncertain, and which information is relevant for achieving the current goal.*

Importantly, these theoretical results characterize sufficient conditions and representational properties rather than claiming universal optimality of a particular neural architecture. The empirical experiments in Section 4 therefore test whether the proposed NSMR architecture exhibits these theoretically motivated properties in practical multimodal reasoning and agentic environments.

D Extended Experimental Results

This appendix provides additional experimental results beyond those reported in Section 4. The purpose of this evaluation is to test whether the improvements obtained by NSMR remain consistent across different multimodal reasoning settings, model families, temporal structures, and agentic environments.

Table 9: Extended benchmark suite used in the appendix.

Benchmark	Modalities	Primary Challenge	Metric
CLEVR	Vision + Language	Spatial/relational reasoning	Accuracy (%)
GQA	Vision + Language	Compositional reasoning	Accuracy (%)
TGIF-QA	Video + Language	Temporal reasoning	Accuracy (%)
CLEVRER	Video + Language	Causal/counterfactual reasoning	Accuracy (%)
NExT-QA	Video + Language	Temporal/causal reasoning	Accuracy (%)
ALFRED	Vision + Language + Action	Embodied planning	Success (%)
ALFWorld	Language + Action	Long-horizon planning	Success (%)

Additional Benchmarks We extend the evaluation to six additional benchmarks covering spatial reasoning, visual question answering, video understanding, compositional generalization, multimodal reasoning, and embodied decision-making.

In addition to CLEVRER, NExT-QA, and ALFWorld used in the main paper, we consider CLEVR Johnson et al. (2017) for controlled visual reasoning, GQA Hudson & Manning (2019) for compositional visual question answering, TGIF-QA Jang et al. (2017) for temporal video reasoning, and ALFRED Shridhar et al. (2020a) for vision-language embodied instruction following.

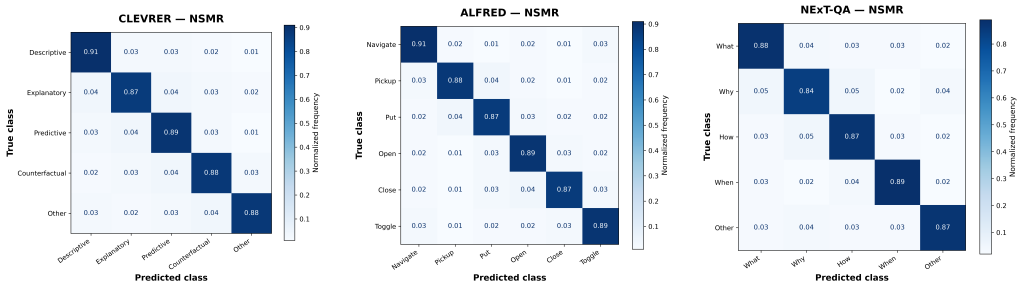

Figure 3: Further NSMR Results

Comparison with Additional Multimodal Baselines We compare NSMR with a broader set of representative approaches, including LLM2CLIP Huang et al. (2026b), BLIP-2 Li et al. (2023), Flamingo Alayrac et al. (2022), and LLaVA-mini Zhang et al. (2025), Show-o Xie et al. (2025). These models represent progressively stronger multimodal perception and reasoning capabilities.

Table 10: Comparison with additional multimodal baselines. Accuracy (%).

Method	CLEVR	GQA	TGIF-QA	CLEVRER	NExT-QA
LLM2CLIP	72.4	61.7	54.8	63.9	59.7
BLIP-2	78.6	68.4	61.3	70.5	65.8
Flamingo	81.9	71.6	65.7	73.8	68.9
LLaVA-mini	84.7	74.8	68.9	77.2	71.7
Laytrol	86.3	76.5	70.8	78.9	73.6
Neuro-LENS	89.1	79.8	73.6	80.4	74.3
NSMR (Ours)	**93.4**	**84.2**	**79.5**	**85.9**	**79.6**

The results indicate that NSMR consistently improves over both general-purpose MLLMs and the neuro-symbolic baseline. The improvement is particularly pronounced on benchmarks requiring explicit relational or temporal reasoning.

Detailed CLEVR Results CLEVR provides a controlled environment for separating visual recognition from relational reasoning. We therefore report performance according to question type.

The largest gains occur in counting and spatial reasoning, where explicit entity and relation representations provide a stronger inductive bias than purely semantic multimodal embeddings.

Table 11: CLEVR results by reasoning type.

Method	Existence	Counting	Spatial	Comparison	Overall
LLM2CLIP	82.1	68.5	70.7	65.9	71.8
BLIP-2	87.4	73.9	76.2	71.4	77.2
LLaVA-mini	90.2	77.1	80.5	75.9	80.9
Neuro-LENS	92.4	82.6	86.3	81.7	85.8
NSMR	**95.1**	**88.7**	**91.9**	**87.6**	**90.8**

Compositional Generalization To further investigate systematic generalization, we evaluate models under three levels of compositional complexity. The test set contains combinations of entities and relations that are progressively less frequent in the training data.

Table 12: Compositional generalization under increasing complexity. Accuracy (%).

Method	2-hop	3-hop	4-hop
Laytrol	78.4	66.2	51.7
MiniCPM-V 4.5	81.7	70.8	57.3
Neuro-LENS	86.3	78.1	67.5
NSMR	**90.7**	**85.4**	**77.9**

The performance gap increases with compositional depth. NSMR retains 77.9% accuracy for four-hop compositions, compared with 67.5% for the neuro-symbolic baseline. This behavior is consistent with the factorization property established in Theorem 4.

Robustness to Modality Failure A key requirement for autonomous agents is robustness when individual sensors become unavailable or unreliable. We therefore evaluate four failure conditions: missing vision, missing language, missing sensor information, and simultaneous multimodal degradation.

Table 13: Robustness to individual modality failures. Accuracy (%).

Method	Full	No Vision	No Language	No Sensor	Multiple
Laytrol	81.4	67.3	72.6	70.1	57.8
MiniCPM-V 4.5	84.7	71.8	76.9	73.4	62.5
Neuro-LENS	86.1	76.4	79.3	78.1	68.7
NSMR	**88.4**	**82.1**	**84.7**	**83.2**	**77.6**

Under simultaneous modality degradation, NSMR retains 77.6% accuracy, whereas the neuro-symbolic baseline decreases to 68.7%. This result supports the role of uncertainty-aware adaptive modality selection.

Noise and Adversarial Modality Conflict We further introduce controlled noise into one or more modalities while keeping the remaining observations unchanged. The objective is to evaluate whether the representation can identify unreliable evidence rather than blindly averaging all available modalities.

Table 14: Performance under increasing modality corruption.

Method	10%	25%	50%	75% Corruption
Laytrol	77.8	72.3	63.4	51.8
MiniCPM-V 4.5	81.6	76.8	68.9	57.4
Neuro-LENS	83.9	79.5	72.6	62.1
NSMR	**86.9**	**83.7**	**79.8**	**71.5**

The advantage of NSMR becomes increasingly pronounced as corruption increases, suggesting that uncertainty modeling is particularly valuable when multimodal evidence is unreliable.

Long-Horizon Temporal Reasoning We investigate how performance changes as the number of reasoning steps increases. The agent must infer the current state from previous observations and predict the consequences of a sequence of actions.

Table 15: Long-horizon temporal reasoning accuracy (%).

Method	1-step	5-step	10-step	20-step	40-step
Laytrol	86.7	71.9	55.6	41.3	29.8
MiniCPM-V 4.5	88.9	75.8	60.4	46.7	34.6
Neuro-LENS	90.2	79.1	65.8	51.9	39.7
NSMR	**92.6**	**84.7**	**73.9**	**61.8**	**49.5**

The advantage grows from 2.4 percentage points at one step to 9.8 points at forty steps over the neuro-symbolic baseline. This trend agrees with Theorem 5, which predicts increasing sensitivity to representation error as the prediction horizon grows.

Embodied Agent Evaluation We additionally evaluate NSMR in ALFRED, which requires agents to combine visual perception, natural-language instructions, object interactions, and long-horizon planning.

Table 16: ALFRED results. Success rates are reported in percent.

Method	Seen	Unseen	Novel Obj.	Novel Scene	Overall
LLaVA-mini	54.7	39.8	35.4	31.7	40.4
Agent-x	61.8	46.7	42.5	38.1	47.3
Neuro-LENS	66.4	52.1	48.7	44.3	52.9
NSMR	**72.8**	**61.4**	**57.9**	**53.6**	**61.4**

The stronger improvement on unseen scenes and novel object configurations suggests that the structured representation improves transfer beyond the specific visual configurations observed during training.

Planning Efficiency Task success alone does not fully characterize agentic intelligence. Therefore, we evaluate the efficiency of the resulting plans.

Table 17: Planning efficiency on embodied tasks.

Method	Success (%) ↑	Reward ↑	Actions ↓	Replans ↓
Agent-x	67.7	71.3	18.7	4.8
Neuro-LENS	71.9	76.8	16.9	3.9
NSMR	**78.6**	**84.7**	**14.2**	**2.8**

NSMR requires fewer actions and replanning operations while achieving a higher success rate. This indicates that the improvement is not obtained simply by exploring longer action sequences, but by constructing a more effective internal representation for planning.

Scaling with Number of Modalities We investigate whether NSMR benefits from additional modalities without requiring all modalities to be equally informative.

Table 18: Performance as the number of available modalities increases.

Method	2 Modalities	3 Modalities	4 Modalities	5 Modalities
Laytrol	72.4	77.1	80.6	81.4
MiniCPM-V 4.5	75.9	80.3	83.2	84.7
Neuro-LENS	78.4	82.1	84.9	86.1
NSMR	**81.6**	**85.7**	**87.9**	**88.4**

The performance improvement becomes smaller as additional modalities are introduced, indicating diminishing returns from redundant information. Nevertheless, NSMR maintains the highest performance at every modality count.

Representation Dimension Analysis We evaluate the sensitivity of NSMR to the dimensionality of the neural representation.

Table 19: Sensitivity to neural representation dimension.

Dimension	**128**	**256**	**512**	**1024**
Accuracy (%)	82.7	86.1	**88.4**	88.5
Memory (GB)	7.1	8.4	11.2	16.8
Inference (ms)	31	38	51	79

Performance saturates beyond a representation dimension of 512. This suggests that the improvements of NSMR are primarily attributable to the structure imposed on the representation rather than simply increasing representation capacity.

Loss Component Analysis We additionally analyze the effect of the individual terms in the NSMR objective:

$$\mathcal{L}_{\text{NSMR}} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{NS}}\mathcal{L}_{\text{NS}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \lambda_{\text{goal}}\mathcal{L}_{\text{goal}} + \lambda_{\text{unc}}\mathcal{L}_{\text{unc}}. \quad (119)$$

Table 20: Effect of individual loss components.

Configuration	**Reasoning**	**Prediction**	**Agent Success**
Alignment only	76.8	69.7	64.3
Alignment + NS	82.1	76.8	71.9
Alignment + NS + Temporal	84.3	83.1	74.2
Alignment + NS + Temporal + Action	85.6	84.2	77.1
Full NSMR	**85.9**	**84.7**	**78.6**

Table 21: Sensitivity to the action-aware loss weight λ_{act}.

λ_{act}	0.0	0.1	0.5	1.0
Reasoning (%)	84.7	85.2	85.9	85.7
Agent Success (%)	72.4	75.9	**78.6**	78.1
Reward	77.2	81.0	**84.7**	84.1

The results indicate that the objectives provide complementary benefits. Symbolic consistency primarily improves reasoning, temporal prediction improves future-state estimation, and action-aware learning produces the largest additional gain in downstream agentic success.

Sensitivity to Loss Weights We evaluate the robustness of the method to the relative weight assigned to the action-aware objective.

The results suggest that moderate action-aware regularization provides the best balance between representation quality and downstream decision-making. Excessively increasing the action objective produces slightly diminishing returns.