# Language-Conditioned Latent Affordance Operators for Culturally Robust Physical Commonsense Reasoning

**Rizky Ananta**[1] **and Bagas Wicaksono**[2]
[1]Universitas Negeri Makassar, Jalan Daeng Tata Raya, Makassar 90222, Indonesia
[2]Universitas Jenderal Achmad Yani Yogyakarta, Jalan Ringroad Barat, Sleman 55292, Indonesia

**ABSTRACT** Large-scale language models increasingly mediate everyday decision-making, yet many deployed settings require physical commonsense judgments that are both linguistically diverse and culturally situated. Physical commonsense is not only about abstract plausibility; it also encodes community-specific conventions over tools, materials, foods, and routines, and such conventions interact with morphology, script, and discourse style. This paper proposes a technical framework for culturally robust physical commonsense reasoning that does not rely on translating all inputs into a single pivot language. The core contribution is a latent affordance operator model that maps utterances to a canonical action-object state space, then performs constrained probabilistic inference over feasible physical transformations. Language enters through a conditional operator generator that produces distributions over affordance parameters, while an invariance regularizer drives cross-lingual agreement at the level of predicted physical effects rather than surface form. The resulting system separates language-specific expression from shared physical structure and admits principled uncertainty estimates when cultural priors differ. We formalize the inference problem as energy-based posterior decoding with compositional operators, derive training objectives that couple contrastive alignment with calibration constraints, and analyze robustness under distribution shift in the language channel. We also outline an evaluation protocol spanning text-only plausibility, tool-use planning, and counterfactual intervention queries, emphasizing measurement that is sensitive to cultural specificity while remaining physically grounded.

**KEYWORDS**

## 1. Introduction

Physical commonsense reasoning is typically framed as the ability to choose actions that respect constraints imposed by objects, materials, and environments. In practice, the notion of what is physically reasonable is mediated by culturally contingent defaults: which container is used for a given liquid, whether a food is expected to be served hot or cold, which household surfaces are considered appropriate for preparation, and which improvised tools are normative. These conventions are not arbitrary; they compress repeated exposure to local environments and practices into fast priors that guide action selection. When reasoning is mediated through language, such priors are also entangled with lexicalization patterns, idioms, and the implicit assumptions that speakers omit because they are shared within a community. A robust system must therefore distinguish universal physical constraints from community-specific priors, and must do so across many languages, dialects, and scripts without assuming that translation preserves the relevant implicatures.

Current approaches often treat multilingual physical commonsense as a representation learning problem, aiming to align sentence embeddings across languages and then apply a single classifier or generative model. This can work when the decision boundary depends primarily on syntactic or semantic cues that are translated cleanly. However, physical commonsense queries frequently reference culturally local objects and procedures, or rely on underspecified descriptions where the missing details are filled in by local defaults [1]. In such cases, forcing all inputs into a pivot language can erase precisely the information that disambiguates the intended physical situation. What is needed is a model that keeps linguistic variation in the input channel, while projecting into a language-agnostic latent space

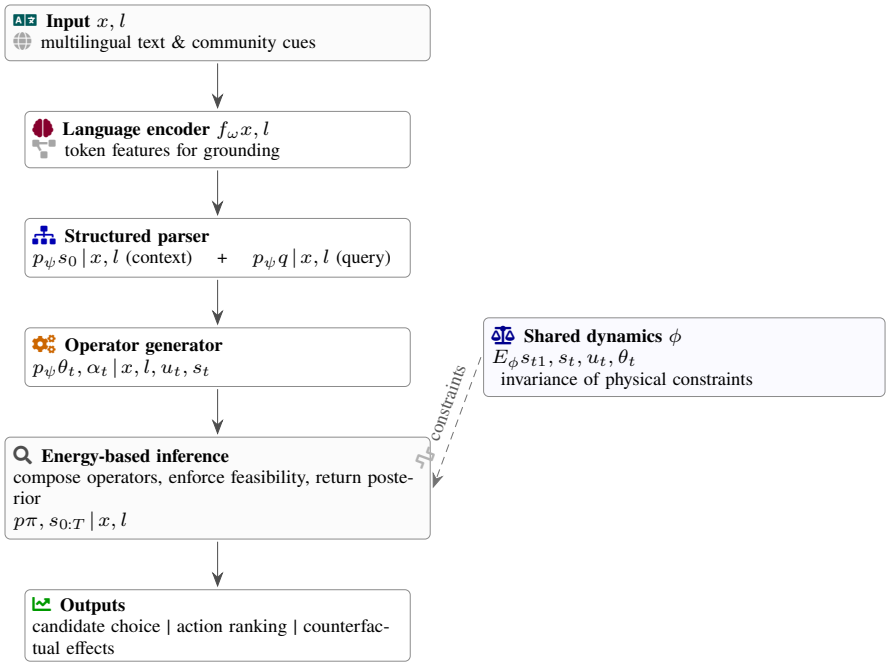


**Figure 1** End-to-end pipeline: multilingual text conditions a structured context/query posterior and an action-parameter generator, while a shared energy model enforces physically grounded feasibility and produces calibrated posteriors over plans and outcomes.

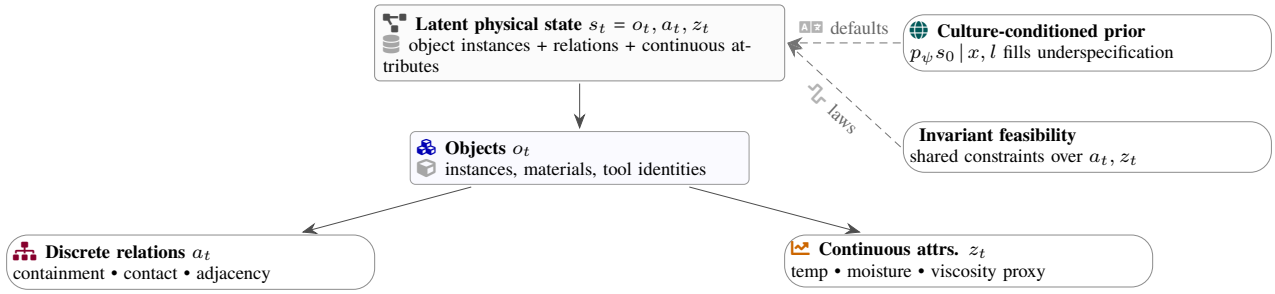


**Figure 2** Latent state factorization: object identities and relations support discrete feasibility, while continuous attributes summarize coarse physical dynamics. Culture- and language-conditioned priors infer typical initial contexts without altering the shared physical constraint set.

that explicitly represents physical state, affordances, and feasible transformations.

This paper develops a framework built around latent affordance operators, which are stochastic, compositional mappings from physical states to physical states. Each operator corresponds to an action type with parameters tied to objects and materials. Unlike purely symbolic planners, the operator parameters are generated from natural language using a learned conditional distribution that can vary with language and culture. Unlike purely neural classifiers, the inference is constrained by an explicit state representation and feasibility potentials that encode conservation-like and compatibility constraints. The key idea is to separate two forms of uncertainty: uncertainty about the physical dynamics, which should be largely invariant across languages for a fixed environment class, and uncertainty about the latent context and conventions, which can differ across communities even when the literal text is similar [2]. An observation is that multilingual evaluations of physical commonsense can expose large performance disparities across language communities, including substantial gaps between higher-resource and lower-resource settings, and such disparities persist even when the task is intended to test everyday knowledge rather than specialized expertise; for instance, Chang et al. (2025) report gaps as large as 37% accuracy between language groups on a culturally grounded physical commonsense benchmark. These findings motivate methods that explicitly model how language and culture modulate priors over actions and objects, rather than assuming that better scaling alone will close the gap [3].

The technical contribution of this work is not an attempt to optimize for any particular benchmark. Instead, it proposes a modeling and inference architecture that is suitable for a broad class of physically grounded queries, including choice questions, open-ended action proposals, and counterfactual queries of the form "what would change if an alternative action were taken." The contributions can be summarized as follows in narrative form: a canonical latent state space for household-scale physical reasoning; a language-conditioned operator generator that produces distributions over action parameters and affordance

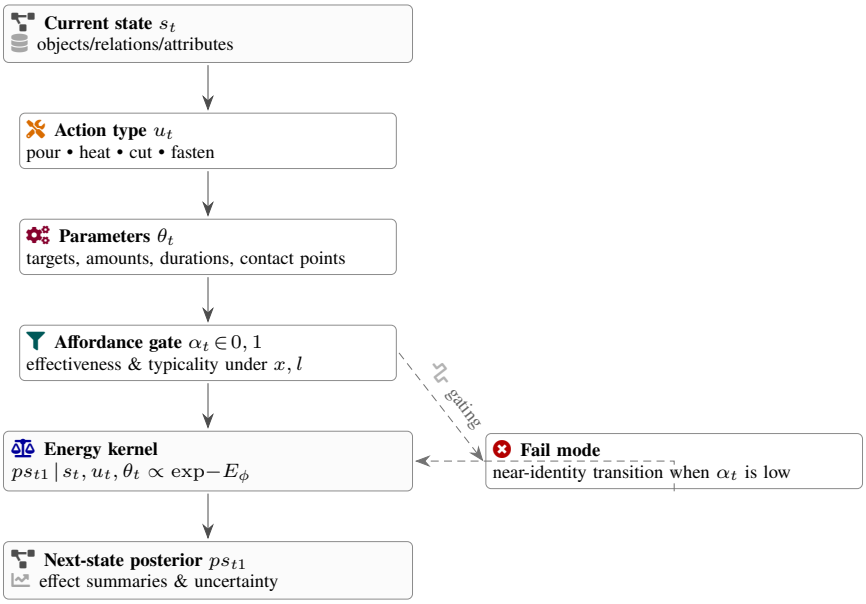


**Figure 3** Single-step operator application: language-conditioned parameters and an affordance gate modulate an energy-based transition that preserves hard feasibility while representing when an action is predicted ineffective or culturally atypical in the inferred context.

strengths; an energy-based inference procedure that composes operators and yields calibrated posteriors over candidate outcomes; and a training objective that encourages invariance of predicted physical effects across languages while allowing culture-specific priors to persist when supported by data. The remainder of the paper formalizes these components, derives learning and inference rules, and proposes an evaluation methodology and experimental design that can be instantiated in simulation and with curated text corpora.

## 2. Problem Formulation

We consider a family of tasks in which an input utterance describes a goal, a situation, or a question about the physical world, and the system must infer a plausible outcome, select a feasible action, or choose among candidate solutions [4]. Let $l$ denote a discrete language identifier that indexes writing system and linguistic conventions, while not presupposing any particular geopolitical boundary. Let $x$ denote the observed text string in language $l$. We assume that the text implicitly refers to a latent physical context $c$, which includes a set of objects, their material properties, and relevant environmental conditions. The system's output can take multiple forms: a binary choice among candidates, a ranked list of action sequences, or a predicted next state. For concreteness and to expose the underlying structure, we define a unified formulation in which the system produces a distribution over latent plans $\pi$ and resulting states $s_T$ after a bounded horizon $T$.

The latent physical state at time $t$ is denoted $s_t \in \mathcal{S}$, where $\mathcal{S}$ is a structured space comprising object identities, discrete attributes, and continuous attributes. A minimal but expressive instantiation partitions $s_t$ into $o_t, a_t, z_t$, where $o_t$ is a multiset of object instances, $a_t$ are discrete attributes such as containment relations and contact graphs, and $z_t$ are continuous attributes such as temperature, viscosity proxies, moisture level, and geometric configuration descriptors. We do not assume that all these quantities are observed; rather, they are latent and inferred from text and optional sensory side information [5]. The latent context $c$ induces a prior distribution over $s_0$, reflecting what objects are likely present and what default configurations are typical in a community.

Actions are modeled as operators that transform states. An action at time $t$ is $u_t \in \mathcal{U}$ with parameters $\theta_t \in \Theta$, and it induces a stochastic transition kernel $ps_{t1} \mid s_t, u_t, \theta_t$. The plan is $\pi = u_0, \theta_0, \ldots, u_{T-1}, \theta_{T-1}$. The text $x$ constrains the plan via a goal or question semantics, but it can do so indirectly. For example, a question may ask which of two candidate actions is more appropriate, or it may ask which outcome is more likely. We unify these by introducing a query variable $q$ derived from $x$ that defines a scoring function over terminal states and, when relevant, over intermediate action choices.

The key modeling challenge is multilingual and cultural variation. The same physical action can be described with different lexical granularity across languages, and some languages encode distinctions that others leave implicit [6]. Additionally, communities differ in which objects are salient defaults. We therefore introduce a language-conditioned semantic channel that maps $x, l$ to distributions over latent variables that connect language to physical state and action. Specifically, we define an operator generator distribution $p\theta_t \mid x, l, u_t$ that produces action parameters from text, and a context prior $ps_0 \mid x, l$ that infers the likely initial configuration given the utterance and language-conditioned conventions.

The overall inference goal is to compute a posterior over plans and outcomes given the text:

$$p\pi, s_{0:T} \mid x, l \propto ps_0 \mid x, l \prod_{t=0}^{T-1} pu_t, \theta_t \mid x, l, s_t \, ps_{t1} \mid s_t, u_t, \theta_t. \tag{2.1}$$

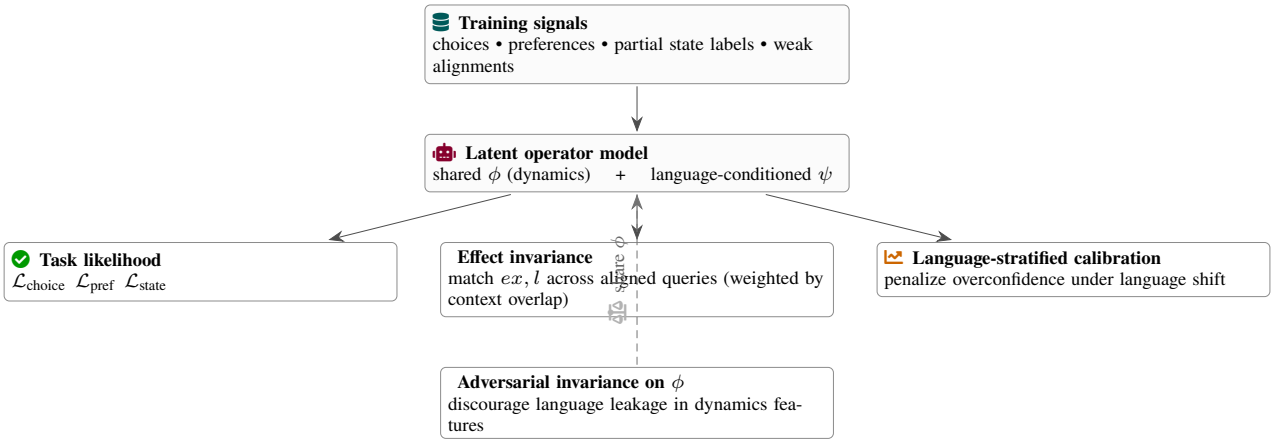


**Figure 4** Learning signals and regularization: task supervision trains decision accuracy, effect-level invariance promotes cross-lingual transfer in predicted physical changes, calibration reduces brittle confidence in sparse languages, and an adversarial term pushes dynamics parameters toward language-invariant structure.

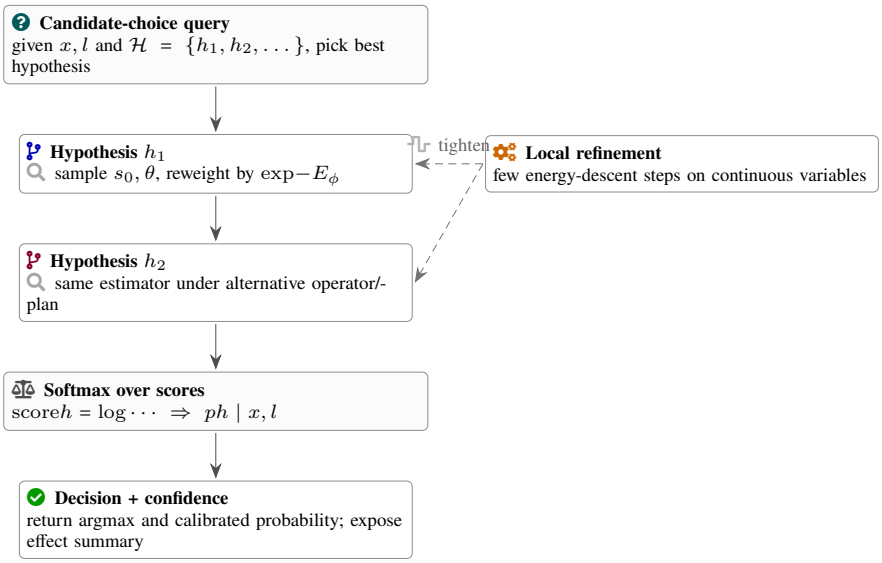


**Figure 5** Candidate-choice inference: each hypothesis induces a small latent model whose marginal score is approximated via sampling and energy reweighting, optionally refined in continuous subspaces, then normalized to produce both a selection and calibrated uncertainty.

The factor $pu_t, \theta_t \mid x, l, s_t$ is decomposed into a proposal distribution over action types and a conditional over parameters, and it may also incorporate feasibility constraints based on the current state. For tasks where candidate answers are provided, we restrict $\pi$ to a small set of candidate single-step actions or candidate terminal outcomes. For tasks requiring multi-step reasoning, we allow short sequences and marginalize accordingly [7].

A central design requirement is to permit cultural priors to differ while maintaining invariance of physical dynamics. We formalize this by decomposing the transition model into an invariant dynamics component and a culture-dependent affordance gating component. Let $\phi$ denote invariant dynamics parameters shared across languages, and let $\psi_l$ denote language-conditioned priors and lexical grounding parameters. We model the transition as an energy-based kernel:

$$ps_{t1} \mid s_t, u_t, \theta_t = \frac{1}{Zs_t, u_t, \theta_t} \exp\big(-E_\phi s_{t1}, s_t, u_t, \theta_t\big), \tag{2.2}$$

where $E_\phi$ encodes compatibility constraints and approximate physical laws at the granularity relevant to the tasks, such as containment feasibility, temperature monotonicity under heating, or viscosity reduction under dilution. The language- and culture-dependent components appear in the prior over $s_0$ and in the operator parameter generator. This separation enables the model to represent that different communities may choose different actions for the same goal because they have different typical tools or preferences, while still representing that the physical effect of adding water to oil or heating a substance follows the same underlying qualitative trends.

We assume access to training data of the form $x_i, l_i, y_i$, where

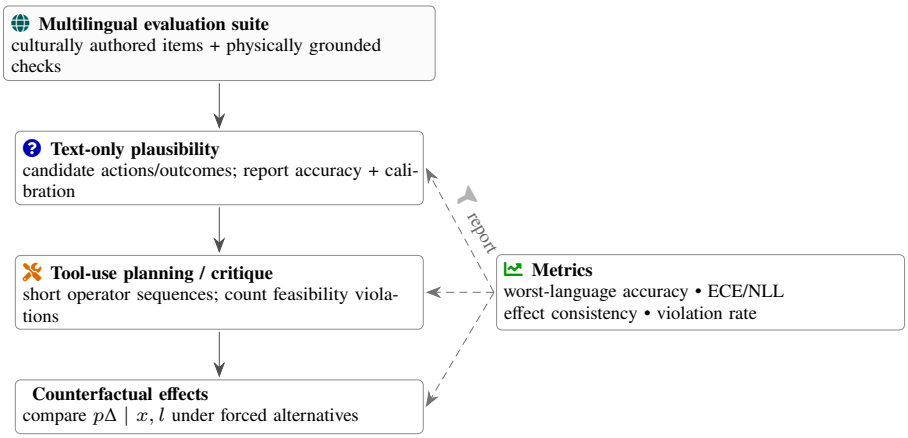


**Figure 6** Evaluation protocol: three complementary axes test multilingual physical commonsense—plausibility judgments, operator-level planning feasibility, and counterfactual effect prediction—while emphasizing worst-language performance and calibrated uncertainty on culturally specific scenarios.

$y_i$ is a supervision signal [8]. The signal can be a labeled correct option, a demonstrated action sequence, a partial annotation of the latent state, or a preference among outcomes. The formulation supports heterogeneous supervision by treating $y_i$ as evidence about a subset of variables. The training objective is then to maximize conditional likelihood or a margin-based surrogate while enforcing cross-lingual invariance at the level of predicted physical effects. The next section specifies the model class used to instantiate these components.

## 3. Model Architecture

The proposed model consists of three interacting modules: a language-conditioned parser that maps text to a distribution over latent contexts and queries, a latent affordance operator module that maps proposed actions to state transitions, and an inference engine that composes operators and produces posteriors over candidate outcomes. The design intentionally avoids treating the language model as the decision-maker. Instead, language modeling is used as a conditional generator of structured latent variables that parameterize a physically grounded inference problem [9].

The language-conditioned parser begins with an encoder $f_\omega x, l$ that produces a sequence of contextual embeddings. The encoder can be any multilingual sequence model, but its role here is limited: it produces features that support extraction of objects, action cues, and constraints. From these features, the parser defines distributions over a discrete object vocabulary $\mathcal{V}$ and a continuous attribute space. The context prior is modeled as a structured distribution

$$p_\psi s_0 \mid x, l = p_\psi o_0 \mid x, l \, p_\psi a_0 \mid o_0, x, l \, p_\psi z_0 \mid o_0, a_0, x, l, \tag{3.1}$$

where each factor can be implemented using neural conditional random fields or autoregressive structured prediction. The object factor selects a set of likely objects and materials. The attribute factor selects relations such as containment and adjacency. The continuous factor outputs distributions over relevant physical scalars. Importantly, these distributions are allowed to depend on $l$, capturing language- and community-specific defaults when the text is underspecified [10]. For example, an utterance that mentions a traditional cooking vessel could imply a material and shape that differ across regions even if the translation is the same in a pivot language.

The query variable $q$ is derived similarly, representing either a goal condition $gs_T$, a preference relation between candidate states, or an evaluation functional over action sequences. We model $q$ as a small program in a constrained domain-specific language that references state attributes and operator outcomes. This program is not produced via unconstrained code generation; rather, it is assembled from templates with learned slot filling, so that inference remains tractable and the meaning is inspectable. The parser outputs a distribution over templates and slot values, and the expected query is used in inference.

The latent affordance operator module defines a family of action types $u \in \mathcal{U}$, each associated with an operator function that maps a state to a distribution over next states. For each action type, the module defines a parameter vector $\theta$ and an affordance strength $\alpha \in 0, 1$ representing the feasibility and effectiveness of the action in the current context. Both $\theta$ and $\alpha$ are predicted from language and current state [11]. The key is that $\alpha$ can encode culturally contingent preferences insofar as they are grounded in typical tool availability and normative practice, while hard physical impossibilities remain enforced by the energy function. The conditional distribution is

$$p_\psi \theta, \alpha \mid x, l, u, s_t = \mathcal{N}\big(\theta \mid \mu_\psi x, l, u, s_t, \Sigma_\psi x, l, u, s_t\big) \tag{3.2}$$

$$\mathrm{Beta}\big(\alpha \mid r_\psi x, l, u, s_t, k_\psi x, l, u, s_t\big) \, . \tag{3.3}$$

where the mean and covariance functions are computed from a fusion of language features and state features. The state features are derived from a graph representation of $s_t$, with nodes for objects and edges for relations, and a graph neural network produces embeddings that are combined with the text embeddings. This design allows the same action phrase to yield different parameter distributions depending on the inferred

| Module | Input | Output | Role |
|---|---|---|---|
| Language encoder $f_\omega$ | Utterance $x$, language id $l$ | Token-level embeddings | Provides multilingual features for grounding objects, actions, and constraints. |
| Context prior $p_\psi s_0 \mid x, l$ | Encoded text | Distribution over initial state $s_0$ | Infers latent objects, relations, and attributes using culture-sensitive defaults. |
| Query parser $p_\psi q \mid x, l$ | Encoded text | Goal / scoring program $q$ | Maps text to a structured query over plans and terminal states. |
| Operator generator $p_\psi \theta_t, \alpha_t \mid x, l, u_t, s_t$ | Text and current state | Action parameters and affordance gate | Produces language-conditioned parameters for each action type. |
| Energy model $E_\phi$ | $s_t, s_{t1}, u_t, \theta_t$ | Transition energy | Encodes shared physical dynamics and feasibility constraints. |
| Inference engine | All factors above | $p\pi, s_{0:T} \mid x, l$ | Composes operators and returns posteriors over plans and outcomes. |

**Table 1** High-level modules in the latent affordance operator architecture and their roles in connecting language to physically grounded inference.

objects and relations.

The transition kernel uses an energy function that combines a dynamics prior, soft constraints, and affordance gating:

$$\begin{aligned} E_\phi s_{t1}, s_t, u_t, \theta_t = {} & E_{\mathrm{dyn},\phi} s_{t1}; s_t, u_t, \theta_t \\ & \lambda_{\mathrm{cons}} E_{\mathrm{cons}} s_{t1}, s_t \\ & \lambda_{\mathrm{aff}} 1 - \alpha_t \, E_{\mathrm{fail}} s_{t1}, s_t, u_t. \end{aligned} \tag{3.4}$$

where $E_{\mathrm{dyn},\phi}$ captures typical qualitative dynamics for the operator family, $E_{\mathrm{cons}}$ penalizes violations of generic constraints such as inconsistent containment, and $E_{\mathrm{fail}}$ concentrates mass on near-identity transitions when the action is predicted ineffective. The affordance gate $\alpha_t$ modulates whether an action is expected to have an effect in the given context [12]. This is useful for capturing that some actions, while physically possible, are rarely used in a community and therefore may be less likely to be the intended solution in a commonsense question. At the same time, because $E_{\mathrm{cons}}$ is independent of $\alpha_t$, hard constraints remain enforced even if a culturally favored action would otherwise be preferred.

To support compositional reasoning, operators are composable via marginalization over intermediate states. The inference engine must handle both candidate-choice tasks and open inference. For candidate-choice tasks, we define candidate hypotheses $h \in \mathcal{H}$, where each $h$ corresponds to either a single operator application or a short plan. The model computes a score

$$\begin{aligned} \mathrm{score} h \mid x, l = \log \; & p_\psi s_0 \mid x, l \; {}_t p_\psi \theta_t, \alpha_t \mid x, l, u_t, s_t \\ & \times \exp\big( - E_\phi s_{t1}, s_t, u_t, \theta_t \big) \, ds_{0:T} \, d\theta_{0:T-1}. \end{aligned} \tag{3.5}$$

and normalizes across candidates. For open-ended prediction, the engine samples or optimizes over $\pi$ and $s_{0:T}$ using amortized proposals and gradient-based refinement in the continuous components of the state.

A defining aspect of the architecture is that language-conditioning occurs at two points: in the context prior and in the parameter generator. The invariant dynamics parameters $\phi$ are shared across languages, and their training is regularized by cross-lingual consistency losses described in the next section [13]. This factorization is intended to yield a model that can adapt to different linguistic realizations without requiring translation, while still benefiting from shared physical structure learned from pooled data.

## 4. Learning and Inference

Learning proceeds from heterogeneous supervision, and inference must be tractable under a structured latent space. We present a unified objective that handles labeled choices, preference signals, and partial state annotations. The objective is composed of a task likelihood term, an invariance term, and a calibration term, each motivated by a distinct failure mode in multilingual physical commonsense.

For labeled choice data, each example provides a set of candidate hypotheses $\mathcal{H}_i$ and a correct label $h_i^\star$. The likelihood is modeled with a softmax over marginal scores:

$$\mathcal{L}_{\mathrm{choice}} \phi, \psi = - \; {}_i \log \frac{\mathrm{expscore} h_i^\star \mid x_i, l_i}{h \in \mathcal{H}_i \, \mathrm{expscore} h \mid x_i, l_i}. \tag{4.1}$$

For preference data, where supervision indicates that one hypothesis is more plausible than another, a logistic ranking loss

| Component | Symbol | Contents | Example instantiations |
|---|---|---|---|
| Full state | $s_t$ | Structured physical configuration | Objects, relations, and continuous attributes at time $t$. |
| Object set | $o_t$ | Multiset of entities and materials | Pots, bowls, knives, wooden spoon, cooking oil, rice, water. |
| Discrete relations | $a_t$ | Graph-level symbolic structure | Containment, support, contact, adjacency, tool–target links. |
| Continuous attributes | $z_t$ | Low-dimensional real-valued features | Temperature, moisture, viscosity proxy, coarse geometry. |
| Context $c$ | $c$ | Scenario-specific defaults | Typical kitchen layout, common vessels, household routines. |
| Prior over $s_0$ | $ps_0$ | Distribution induced by $c$ | Likelihood of objects and initial configurations before acting. |

**Table 2** Factorization of the latent physical state into object identities, discrete relations, and continuous attributes, together with scenario-level context that shapes priors over initial configurations.

can be used:

$$\mathcal{L}_{\text{pref}}\phi,\psi = \sum_i \log\big(1 \ \exp\big(-\big[\text{score}h_i \mid x_i, l_i - \text{score}h_i^- \mid x_i, l_i\big]\big)\big). \tag{4.2}$$

For partial state annotations, where certain attributes of $s_0$ or $s_T$ are labeled, the likelihood includes the corresponding marginal: [14]

$$\mathcal{L}_{\text{state}}\phi,\psi = -\sum_i \log p_\psi s_0 \mid x_i, l_i\, p_{\phi,\psi} s_T \mid s_0, x_i, l_i \times \mathbf{1}\big\{As_0, s_T = \hat{A}_i\big\}\, ds_0\, ds_T. \tag{4.3}$$

where $A\cdot$ extracts the annotated attributes.

These task losses alone are insufficient for robustness under language shift because they may encourage the model to encode language-specific shortcuts in $\psi$ that correlate with labels in high-resource settings but do not transfer. The core regularizer therefore enforces invariance of predicted physical effects across languages when the underlying physical situation is semantically matched. We assume access to weak alignment pairs $x_i, l_i$ and $x_j, l_j$ that describe similar physical queries, obtained via bilingual dictionaries, paraphrase mining, or teacher-based alignment, without requiring perfect translation equivalence. For such pairs, we define an effect representation $e$ extracted from the posterior over terminal states, such as expected changes in key continuous attributes and changes in relation graphs. Let $ex, l$ denote this effect representation. The invariance loss is then [15]

$$\mathcal{L}_{\text{inv}}\phi,\psi = \sum_{i,j} \|ex_i, l_i - ex_j, l_j\|_2^2 \quad \gamma \sum_{i,j} \mathrm{D}_{\mathrm{KL}}\big(p_{\phi,\psi}\Delta \mid x_i, l_i \,\|\, p_{\phi,\psi}\Delta \mid x_j, l_j\big). \tag{4.4}$$

where $\Delta$ denotes a discrete summary of the predicted physical change, such as whether temperature increases, whether a mixture becomes more dilute, or whether a container becomes more likely to leak. The Euclidean term enforces agreement in continuous summaries, while the divergence term enforces agreement in categorical outcomes. Crucially, this regularizer acts on the outputs of the physically grounded inference rather than on token embeddings. This makes it less sensitive to differences in lexicalization and more aligned with the intended invariance of physical effects.

However, strict invariance is not always desirable because communities can differ in which contexts are assumed. If the utterances are underspecified and the inferred $s_0$ differs legitimately across languages, forcing the effects to match can erase culturally appropriate priors. We therefore apply invariance conditionally, weighted by an estimated alignment confidence and by posterior overlap of inferred contexts [16]. Define $w_{ij} \in 0, 1$ as an alignment weight. Define an overlap measure $O_{ij}$ between the context posteriors, such as the expected Jaccard overlap of object sets and agreement of relation edges. The invariance loss becomes

$$\mathcal{L}_{\text{inv}}\phi,\psi = \sum_{i,j} w_{ij}\, O_{ij}\, \big(\|ex_i, l_i - ex_j, l_j\|_2^2 \quad \gamma\, \mathrm{D}_{\mathrm{KL}}\big(p\Delta \mid x_i, l_i \,\|\, p\Delta \mid x_j, l_j\big)\big). \tag{4.5}$$

When overlap is low, the model is permitted to maintain different priors, reflecting that the intended physical situation may differ across communities.

Calibration is a separate concern. In many applications, the system should be able to admit uncertainty rather than produce overconfident but culturally inappropriate outputs. We include a calibration term that encourages predicted probabilities over

| Element | Notation | Description |
|---|---|---|
| Action type | $u_t \in \mathcal{U}$ | Symbolic operator family such as *pour*, *heat*, *cut*, or *fasten*. |
| Parameters | $\theta_t \in \Theta$ | Targets, amounts, durations, contact points, and other operator-specific settings. |
| Affordance gate | $\alpha_t \in 0, 1$ | Degree to which the action is feasible, effective, and typical in the inferred context. |
| Transition kernel | $ps_{t1} \mid s_t, u_t, \theta_t$ | Energy-based distribution enforcing physical feasibility and qualitative dynamics. |
| Plan | $\pi = u_0, \theta_0, \ldots, u_{T-1}, \theta_{T-1}$ | Short sequence of operators composed over a bounded horizon. |

**Table 3** Structure of latent affordance operators that map a physical state to a distribution over next states, conditioned on action type, parameters, and an affordance gate.

candidate answers to be calibrated under language-conditioned shifts. Let $\hat{p}_i$ denote the predicted probability of the chosen hypothesis and let $y_i$ denote correctness. A differentiable calibration surrogate can be constructed by penalizing the squared difference between predicted confidence and empirical accuracy within mini-batches stratified by language:

$$\mathcal{L}_{\text{cal}}\phi, \psi = \underset{l\ b}{} \Big( \frac{1}{|B_{l,b}|} \underset{i \in B_{l,b}}{} \hat{p}_i \ - \ \frac{1}{|B_{l,b}|} \underset{i \in B_{l,b}}{} y_i \Big)^2, \quad (4.6)$$

where $B_{l,b}$ denotes a bin of examples in language $l$ with predicted confidence in an interval indexed by $b$. This term discourages the model from assigning high confidence in languages where it has not learned reliable mappings from text to context and operator parameters [17].

The full objective is

$$\begin{aligned} \min_{\phi,\psi} \quad & \mathcal{L}_{\text{choice}} \ \eta_{\text{pref}}\mathcal{L}_{\text{pref}} \ \eta_{\text{state}}\mathcal{L}_{\text{state}} \\ & \eta_{\text{inv}}\mathcal{L}_{\text{inv}} \ \eta_{\text{cal}}\mathcal{L}_{\text{cal}}, \end{aligned} \quad (4.7)$$

Optimization alternates between amortized inference and parameter updates. Because the scores involve integrals over latent states and operator parameters, exact marginalization is intractable in general. We therefore use a hybrid strategy. Discrete structure, such as object identity and relation edges, is handled with structured variational inference using mean-field factors and loopy belief propagation updates. Continuous attributes and operator parameters are handled with reparameterized sampling and gradient-based refinement.

For candidate-choice tasks, inference is simplified by restricting attention to short horizons and to candidate hypotheses [18]. Each hypothesis induces a small latent graphical model whose partition function can be approximated with importance sampling. The proposal distribution for $s_0$ is given by the context prior, while the proposal for $\theta$ is given by the operator generator. The importance weight is proportional to the exponentiated negative energy. When the action has continuous effects, a small number of refinement steps are applied to the sampled continuous variables by descending the energy while keeping discrete structure fixed, then reweighting accordingly. This yields a tighter approximation of the marginal score without requiring full simulation.

For open-ended planning, we use a beam of action types sampled from a language-conditioned policy $pu_t \mid x, l, s_t$, and for each beam element we sample parameters and propagate state distributions. Because the model is intended for commonsense scales rather than fine-grained robotics, the state representation remains coarse, and transitions can be computed efficiently [19]. The model can answer counterfactual questions by comparing posterior distributions under different forced actions. This is accomplished by conditioning on $u_t$ and recomputing the posterior over $s_{t1}$ and downstream outcomes, making explicit how different actions change the predicted physical effects.

A subtle but important aspect is how the model avoids collapsing culturally contingent priors into the invariant dynamics. We enforce this by constraining which parameters are shared and by adding a penalty that discourages $\phi$ from encoding language-specific signals. Concretely, we add an adversarial term in which a language classifier attempts to predict $l$ from intermediate dynamics features, while $\phi$ is trained to make this prediction difficult. This encourages physical dynamics features to be invariant. The language-conditioned parameters $\psi$ remain free to encode culture-specific priors in the context and operator distributions. The result is a model where the shared dynamics represent general physical trends, while the language-conditioned components capture how speakers and communities describe and prioritize actions.

## 5. Theoretical Properties

We analyze the proposed factorization through the lens of robustness to language shift and partial observability [20]. The aim is

| Loss term | Targets | Intuition | Supervision source |
|---|---|---|---|
| $\mathcal{L}_{\text{choice}}$ | Candidate labels | Fit softmax over marginal scores for discrete choices. | Multiple-choice physical commonsense questions. |
| $\mathcal{L}_{\text{pref}}$ | Pairwise preferences | Encourage preferred hypothesis to score higher than alternatives. | Human or model-generated preference judgments. |
| $\mathcal{L}_{\text{state}}$ | Latent states | Match partial annotations of $s_0$ or $s_T$ under the generative model. | Object/material labels, relation graphs, effect flags. |
| $\mathcal{L}_{\text{inv}}$ | Effect summaries | Align predicted physical effects across weakly aligned multilingual queries. | Cross-lingual paraphrases and dictionary-based pairs. |
| $\mathcal{L}_{\text{cal}}$ | Predicted confidences | Reduce overconfidence via language-stratified calibration penalties. | Validation accuracy grouped by language and confidence bin. |

**Table 4** Main training losses used to supervise tasks, encourage cross-lingual invariance of predicted effects, and maintain calibration under language shift.

not to claim universal guarantees, but to formalize conditions under which invariance at the level of physical effects improves generalization relative to purely embedding-level alignment.

Consider a distribution over examples indexed by language, where each example is generated by first sampling a latent physical scenario $\sigma$ from a scenario distribution, then sampling a language-conditioned utterance $x$ describing $\sigma$, and finally sampling a label $y$ according to a task rule that depends on the true physical effects of candidate actions in $\sigma$. The key assumption is that, conditional on $\sigma$, the physical transition rules are language-invariant, but the utterance channel may be biased and may omit details with probabilities that depend on $l$. This matches the intuition that the world does not change with language, but descriptions do.

Let $h_{\phi,\psi}x, l$ denote the model's predictive distribution over labels. We define a risk in each language $l$ as

$$R_l\phi, \psi = \mathbb{E}_{x,y\sim\mathcal{D}_l}\big[\ell h_{\phi,\psi}x, l, y\big], \tag{5.1}$$

and an overall objective that averages over languages. In a multilingual setting with imbalanced data, we care about the worst-language risk [21]. The model's factorization implies that $h_{\phi,\psi}$ depends on language primarily through the inferred context distribution and operator parameter distribution, and on shared physics through the energy function. The invariance loss encourages that, when two utterances describe comparable scenarios, the induced effect representation is similar.

We can view effect representations as a bottleneck. Let $e = e_{\phi,\psi}x, l$ and let the prediction be $hx, l = ge$ for some downstream mapping $g$ that reads off candidate scores. If $e$ is approximately sufficient for $y$ and is invariant across languages for matched scenarios, then learning $g$ on high-resource languages can transfer. A failure mode occurs if $e$ inadvertently encodes language identity, which can happen if the context prior uses language-specific shortcuts correlated with labels. The adversarial invariance in dynamics and the output-level invariance penalty aim to reduce this leakage.

A formal way to capture this is to bound the target-language risk in terms of source-language risk plus a divergence between effect distributions. Suppose we train on a mixture of source languages and evaluate on a target language $l_t$ [22]. Let $\mathcal{E}_s$ denote the distribution of effect representations under the source mixture, and $\mathcal{E}_t$ under the target. Under Lipschitz assumptions on the loss composed with $g$, one can express a domain adaptation style bound:

$$R_{l_t}\phi, \psi \le R_s\phi, \psi \quad L_g\, W_1\mathcal{E}_s, \mathcal{E}_t \quad \epsilon_{\text{approx}}, \tag{5.2}$$

where $W_1$ is the Wasserstein-1 distance between effect distributions, $L_g$ is a Lipschitz constant, and $\epsilon_{\text{approx}}$ captures the mismatch between the chosen effect representation and the true sufficient statistics for the task. The invariance regularizer directly targets the reduction of distances between effect representations for aligned utterances, which can be interpreted as reducing an empirical proxy for $W_1$. In contrast, embedding-level alignment operates on token features that may not be sufficient for the task and can be confounded by language-specific syntax.

Partial observability introduces another term. When text is underspecified, the model must infer a distribution over contexts. Let $\hat{s}_0$ be an inferred context sample and let $ps_0 \mid x, l$ be the true posterior. Errors in this posterior can change effect distributions. A useful decomposition separates error from dynamics and error from context inference [23]. Consider the expected effect under the model, $\mathbb{E}e \mid x, l$, and the true expected effect. Under smoothness of the dynamics mapping from contexts to effects, the error is bounded by a term proportional to the divergence between context distributions. This suggests that improving multilingual context inference is essential, and also suggests

| Aspect | Shared across languages? | Primary parameters | Examples |
|---|---|---|---|
| Physical dynamics | Yes | $\phi$ (energy function) | Containment feasibility, monotone heating effects, basic mixing trends. |
| Context prior | No (language-conditioned) | $\psi$ for $ps_0 \mid x, l$ | Typical cooking vessels, surface choices, default serving temperatures. |
| Operator parameters | No (language-conditioned) | $\psi$ for $p\theta_t, \alpha_t \mid x, l, u_t, s_t$ | Different cutting styles, typical tool–object pairings, plausible quantities. |
| Query semantics | Partially shared | Templates and slot predictors | Goal predicates, comparison functions, counterfactual operators. |

**Table 5** Separation between language-invariant physical dynamics and language-conditioned components that encode priors over contexts, operator parameters, and query forms.

| Evaluation axis | Input / output form | Physical focus | Typical metrics |
|---|---|---|---|
| Text-only plausibility | Multilingual description + candidate options | Everyday feasibility of actions and outcomes | Accuracy, negative log-likelihood, expected calibration error. |
| Tool-use planning / critique | Goal, objects, and short plans | Operator composition and feasibility violations | Feasibility violation rate, success rate, plan length statistics. |
| Counterfactual effects | Context with alternative actions | Changes in attributes and relations under hypothetical choices | Agreement on effect summaries, effect consistency across languages. |
| Language-robust calibration | Same task distribution across languages | Confidence under domain and language shift | Worst-language accuracy, overconfidence gap, entropy of posteriors. |

**Table 6** Planned evaluation axes spanning plausibility judgments, tool-use reasoning, counterfactual prediction, and calibration across languages.

why allowing language-conditioned priors can be beneficial: if the true context differs across communities, forcing identical priors can increase context error.

The model's conditional invariance weighting can be interpreted as a safeguard against negative transfer. If overlap $O_{ij}$ is low, the invariance penalty is small, allowing the model to represent that two utterances, while aligned at a superficial semantic level, might imply different default contexts and therefore different physical effects. In this sense, the method targets invariance of dynamics given context, not invariance of context itself. This aligns with the principle that physical laws are invariant, but typical initial conditions are not.

Another theoretical question concerns identifiability of the factorization between dynamics $\phi$ and language-conditioned priors $\psi$. Without additional constraints, one could shift explanatory burden between these components [24]. The architecture mitigates this by restricting $\phi$ to parameterize only the energy terms that correspond to generic physical constraints and operator families, while allowing $\psi$ to control context distributions and operator parameter priors. The adversarial language-invariance on intermediate dynamics features further discourages $\phi$ from encoding language-specific signals. While this does not guarantee identifiability, it restricts the space of degenerate solutions and empirically improves transfer in settings where language correlates with label frequencies.

Finally, the calibration term can be connected to uncertainty quantification under distribution shift. If the model's confidence is overly high in languages with sparse supervision, then downstream decisions can be unreliable. By penalizing miscalibration within each language, the model is encouraged to express higher entropy when evidence is weak. This is particularly important in culturally diverse settings, where the model may encounter objects and routines absent from training data [25]. A calibrated posterior over candidate answers can support safe abstention or request for clarification, without requiring that the model always commit to a single culturally loaded default.

Overall, the analysis supports the design choice to enforce invariance at the level of predicted physical effects while per-

| Type | Attribute / relation | Role in reasoning |
|---|---|---|
| Discrete relation | Containment (in/inside) | Determines which liquids or solids can be moved, mixed, or heated together. |
| Discrete relation | Support / contact | Constrains which objects can be stacked, balanced, or used as tools. |
| Discrete relation | Adjacency / reachability | Influences which actions are plausible without moving objects first. |
| Continuous attribute | Temperature | Governs heating, cooling, serving defaults, and safety constraints. |
| Continuous attribute | Moisture / wetness | Affects adhesion, dissolving, and drying behavior in household tasks. |
| Continuous attribute | Viscosity / geometry proxy | Determines whether pouring, spreading, or cutting will be effective. |

**Table 7** Representative discrete relations and continuous attributes used to capture coarse physical structure in household-scale commonsense scenarios.

mitting language-conditioned context priors. It also clarifies that robustness depends on both the fidelity of the dynamics energy function and the quality of context inference, and that improvements in multilingual performance need not come from scaling the language encoder alone if the bottleneck lies in structured physical inference.

## 6. Empirical Study

This section describes an experimental design to evaluate the proposed model across tasks that require physical commonsense, multilingual grounding, and sensitivity to culturally contingent defaults. The focus is on methodology rather than any single dataset. The key objective is to test whether the latent affordance operator factorization yields improved cross-lingual transfer and better-calibrated uncertainty compared to baselines that rely on direct text classification or pivot translation.

A first evaluation axis is text-only physical plausibility with candidate choices [26]. In this setting, each example presents a goal or situation description and two or more candidate completions or actions. The model must select the most plausible option. To ensure multilingual relevance, the evaluation set should include examples authored directly in each language rather than translated templates, and it should span scripts and morphological typologies. The examples should include both culturally neutral physical situations, such as generic container and temperature reasoning, and culturally specific situations involving local tools, foods, or household routines. The model is trained on a mixture of languages with imbalanced quantities, then evaluated on held-out languages and on languages with held-out culturally specific domains. The primary metric is accuracy, but calibration metrics such as expected calibration error and negative log-likelihood are also reported, because a model that is slightly less accurate but far better calibrated may be preferable in downstream applications.

A second evaluation axis is action proposal and critique [27]. Here the input is a goal and a short description of available objects, optionally extracted from text. The model must propose a short action sequence or judge whether a proposed sequence is feasible. This tests whether the operator composition is meaningful and whether the energy constraints prevent physically impossible proposals. Because the state representation is structured, feasibility can be evaluated against a simulator at the same granularity as the model, such as a symbolic physics engine that checks containment, contact, and qualitative temperature changes. The comparison baselines include sequence-to-sequence generation of actions and direct classification of feasibility. The expectation is that the operator model yields fewer gross physical violations and a more interpretable distribution over failure modes. Importantly, multilinguality enters not only through action descriptions, but also through object naming and implicit conventions about tool usage [28]. The evaluation therefore includes language-specific object lexicons and assesses whether the model can generalize from shared dynamics even when object names differ.

A third evaluation axis is counterfactual physical effects. The input specifies a context and asks what would change if an alternative action were taken, such as substituting a different material, changing a temperature-related step, or altering the order of operations. The model produces distributions over effect summaries, and performance is measured by agreement with a simulator or with expert annotations at the qualitative level. This axis is where output-level invariance is most directly testable.

| Signal type | Example annotation | Variables constrained | Notes |
|---|---|---|---|
| Labeled choices | Correct option among candidates | Relative scores of hypotheses $h \in \mathcal{H}$ | Most common supervision; drives end-task accuracy. |
| Preference pairs | One hypothesis preferred over another | Score differences $\text{score}h - \text{score}h^-$ | Captures fine-grained preferences without absolute labels. |
| Partial state labels | Objects, materials, or relations | Subsets of $s_0$ or $s_T$ | Improves grounding of latent state factors and operator effects. |
| Weak multilingual alignments | Semantically related $x_i, l_i, x_j, l_j$ | Effect summaries $ex, l$ | Used in effect-level invariance loss with overlap weighting. |
| Simulator or scripted traces | Short operator sequences with outcomes | Dynamics parameters $\phi$ | Provides dense supervision for qualitative physics. |

**Table 8** Sources of heterogeneous supervision used to train the model, ranging from direct labels to weak cross-lingual alignments and simulator-derived trajectories.

For aligned utterance pairs across languages that describe the same counterfactual, the model's predicted effect distributions should match closely even if the utterances differ substantially in surface form. In contrast, for utterance pairs that are superficially similar but correspond to different cultural defaults, the model should either represent a bimodal posterior over contexts or show increased uncertainty rather than forcing an invariant effect [29].

The training regimen uses heterogeneous supervision. Choice-labeled data supplies strong signals for the final decision, while a smaller amount of structured annotation supplies signals for context inference and operator parameterization. For example, some examples can be annotated with the key object and material, or with a categorical attribute such as whether the scenario involves heating, dilution, cutting, or fastening. These annotations improve identifiability and allow targeted evaluation of intermediate representations. Additionally, weak alignment pairs across languages can be mined from parallel corpora or paraphrase models, but the alignment need not be perfect because the invariance loss is weighted by overlap of inferred contexts.

Baselines are chosen to isolate the contribution of structured physical inference. One baseline is a multilingual encoder followed by a classifier over candidates, trained end-to-end on the same data [30]. Another baseline uses pivot translation into a high-resource language followed by a monolingual classifier. A third baseline uses a neural model that predicts effect summaries directly from text without explicit state or operator representations. Comparisons focus on performance under language shift, especially for low-resource languages and for culturally specific examples. A key hypothesis is that the operator model improves transfer by learning shared dynamics, while the context prior captures language-conditioned defaults. Another hypothesis is that calibration improves because uncertainty in context and operator parameters can be propagated through inference, yielding higher entropy when the text is underspecified.

Ablation studies test the role of each component. Removing the invariance loss should increase sensitivity to language-specific shortcuts, potentially improving high-resource accuracy while hurting transfer [31]. Removing the language-conditioned context prior and forcing a shared prior should harm performance on culturally specific examples, because the model cannot represent different default contexts. Removing the energy constraints and using an unconstrained transition predictor should increase physically implausible outcomes and reduce consistency under counterfactual queries. Removing the calibration term should increase overconfidence, especially in languages with sparse supervision, and this should be observable through higher calibration error even when accuracy is similar.

Because the framework is intended to be general, a simulator is useful for scalable evaluation of operator composition and counterfactual reasoning. The simulator need not be a high-fidelity physics engine; it can be a qualitative simulator that enforces the same constraints encoded in the energy function but with known rules. This allows controlled stress tests, such as varying the availability of objects, perturbing initial conditions, and introducing novel object-material combinations. One can also create synthetic multilingual utterances by sampling from language templates and substituting object lexicons, but such synthetic data should be used primarily for probing invariance and not as a substitute for culturally authored examples [32].

The expected outcomes of this empirical study are patterns rather than single headline numbers. One expects that the operator model yields improved worst-language accuracy compared to purely neural baselines, especially when training data is imbalanced. One expects smaller degradation when evaluating on languages unseen during training, because the shared dynamics parameters support transfer. One expects improved calibration, particularly on culturally specific items where the model should

| Error category | Description | Example consequence | Diagnostic signal |
|---|---|---|---|
| Feasibility violation | Predicted action contradicts hard physical constraints | Suggests container that leaks or impossible mixing step | High energy under $E_{\text{cons}}$, simulator failure, violation counts. |
| Context misinference | Latent objects or materials inferred incorrectly | Uses wrong vessel or tool for a described routine | Disagreement with annotated objects or human-written context. |
| Cultural prior mismatch | Action is feasible but atypical for a community | Proposes unusual serving temperature or preparation surface | Divergence between priors across languages or user feedback. |
| Calibration failure | Confidence misaligned with correctness | Overconfident predictions in sparse languages or novel domains | Elevated language-stratified calibration error and overconfidence gap. |

**Table 9** Qualitative error types that can arise in multilingual physical commonsense reasoning and the corresponding diagnostic signals for analysis.

express uncertainty if context is ambiguous. Finally, one expects that the model's intermediate variables, such as inferred objects and predicted effect summaries, provide interpretable diagnostics that can guide dataset curation and targeted improvement without requiring that every error be analyzed at the token level.

## 7. Discussion

The proposed framework is motivated by the observation that multilingual physical commonsense reasoning is fundamentally a problem of disentangling invariant physical structure from language- and culture-dependent priors. A purely text-based model can implicitly learn this disentanglement in some regimes, but it often does so in a brittle manner, because the latent variables remain unstructured and are optimized only to predict labels [33]. By introducing explicit latent states and compositional operators, the framework makes the inductive bias toward physical consistency explicit. This does not guarantee correctness, but it constrains the space of solutions and yields interpretable points of failure.

One implication of the framework is that scaling language encoders alone may not address the primary bottleneck in low-resource settings. If the model lacks robust mappings from language to context and action parameters, then better token representations may still fail to ground the relevant objects and conventions. The operator approach suggests a complementary pathway: learn shared dynamics from pooled data and simulations, then learn lightweight language-conditioned mappings into the shared physical space. This can reduce the amount of supervised data needed in each language because the expensive part of learning dynamics is shared. However, it also shifts attention to the quality of object lexicons, morphology-aware parsing, and the representation of culturally local artifacts, which may require community engagement and careful curation [34].

Another implication concerns evaluation. Benchmarks that emphasize culturally specific items are valuable for revealing gaps, but they also raise challenges: cultural specificity can blur the line between physical constraints and normative preference. For example, two actions can both be physically feasible, but only one is typical in a community. The framework addresses this by representing both feasibility via constraints and typicality via priors and affordance gating. Evaluations should therefore separate error types: violations of physical feasibility, misinference of context, and mismatch of cultural priors. A model that consistently violates feasibility is qualitatively different from a model that predicts a feasible but culturally atypical action. The latter may be acceptable in some applications if it remains calibrated and offers alternative feasible options [35].

The method also highlights the importance of uncertainty. In multilingual settings, ambiguity arises not only from the text but also from lack of training coverage. A system that always returns a single confident answer risks imposing a dominant culture's defaults on other communities. By maintaining distributions over contexts and operator parameters, and by calibrating confidence per language, the model can represent when multiple interpretations are plausible. This can support interaction patterns where the system asks for clarification or offers multiple feasible options. The design goal is not to encode cultural preferences as immutable rules, but to represent them as probabilistic priors that can be updated with evidence.

There are limitations inherent in the modeling choices [36]. The state space is necessarily coarse to keep inference tractable, which means fine-grained geometric reasoning and complex fluid dynamics are outside scope. The energy function encodes qualitative constraints, which can be insufficient for tasks requiring precise quantitative prediction. Additionally, the assumption that dynamics are invariant across languages is only approximately true because language communities can differ in the physical environments they commonly experience, such as climate and typical household infrastructure, which can change the distribution of scenarios. The framework partially addresses this by allowing context priors to differ and by applying invariance only when inferred contexts overlap, but there remains a gray area where environment variation and culture variation interact.

A further practical consideration is how to obtain training

signals for operator parameters and context inference. While labeled choice data is relatively easy to collect, structured annotations are more costly. Simulation can help by generating synthetic trajectories, but aligning them to natural language across many languages is nontrivial [37]. One path forward is to use multimodal data, such as videos with narrations, to ground operator effects. Another is to use weak supervision from procedural texts, where step sequences imply operator compositions. In all cases, care is needed to avoid encoding stereotypes or overgeneralizing from a small set of authors to an entire community. The framework's probabilistic representation can help by maintaining uncertainty, but data governance and evaluation design remain essential.

Despite these limitations, the latent affordance operator framework offers a coherent approach to building multilingual physical commonsense systems that are physically grounded, interpretable, and sensitive to cultural variation. It provides levers for controlling what is shared across languages and what is allowed to vary, and it enables evaluation and debugging at the level of inferred objects, actions, and effects rather than only at the level of final answers.

## 8. Conclusion

This paper introduced a language-conditioned latent affordance operator framework for culturally robust physical commonsense reasoning [38]. The central idea is to treat language as a conditional channel that specifies distributions over latent physical contexts and operator parameters, while keeping the qualitative dynamics and feasibility constraints largely invariant across languages. The model represents actions as compositional operators in a structured state space and performs energy-based inference to evaluate candidate actions and outcomes. A learning objective combining task likelihood, output-level invariance of predicted physical effects, and language-stratified calibration encourages transfer to low-resource languages without forcing inappropriate homogenization of cultural priors.

The approach is intended as an architectural and methodological contribution rather than a dataset-specific solution. It emphasizes that physical commonsense is best modeled as constrained inference over latent states, and that multilingual robustness benefits from invariance at the level of predicted physical effects rather than purely at the level of token embeddings. By explicitly representing uncertainty over contexts and action parameters, the framework can express ambiguity in culturally underspecified scenarios and reduce overconfidence under language shift. Future work can extend the state representation, improve operator learning through multimodal grounding, and develop evaluation suites that disentangle physical feasibility from culturally contingent typicality. The broader implication is that culturally diverse language interaction with the physical world can be supported by models that separate shared physical structure from language- and community-dependent priors, enabling more reliable and accountable commonsense reasoning across languages [39].

**Acknowledgments**

We would like to thank the reviewers of this research for their helpful comments and suggestions.

## References

[1] F. Lefèvre, A. M. C. Lefèvre, S. A. S. Scandar, and S. Yassumaro, "Social representations of the relationships between plant vases and the dengue vector," *Revista de saude publica*, vol. 38, pp. 405–414, 7 2004.

[2] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, and F. A. Batarseh, "Rationalization for explainable nlp: a survey.," *Frontiers in artificial intelligence*, vol. 6, pp. 1225093–, 9 2023.

[3] T. A. Chang, C. Arnett, A. Eldesokey, A. Sadallah, A. Kashar, A. Daud, A. G. Olanihun, A. L. Mohammed, A. Praise, A. M. Sharma, *et al.*, "Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures," *arXiv preprint arXiv:2510.24081*, 2025.

[4] G. Weikum, X. L. Dong, S. Razniewski, and F. Suchanek, "Machine knowledge: Creation and curation of comprehensive knowledge bases," *Foundations and Trends in Databases*, vol. 10, pp. 108–490, 7 2021.

[5] Y. Wang and J. Lee, "Handling uncertainty in answer set programming," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 3 2015.

[6] J. Allan, A. Clifford, P. Ball, M. Alston, and P. Meister, "'you're less complete if you haven't got a can in your hand': alcohol consumption and related harmful effects in rural australia: the role and influence of cultural capital," *Alcohol and alcoholism (Oxford, Oxfordshire)*, vol. 47, pp. 624–629, 7 2012.

[7] F. Cozman and H. Neri, "Pay attention," *Revista da Universidade Federal de Minas Gerais*, vol. 30, 12 2023.

[8] F. J. Pelletier, R. Elio, and P. P. Hanson, "Is logic all in our heads? from naturalism to psychologism," *Studia Logica*, vol. 88, pp. 3–66, 2 2008.

[9] X. Niu, S. Qin, Z. Haizhu, M. Wang, and R. Wong, "Exploring product design quality control and assurance under both traditional and crowdsourcing-based design environments," *Advances in Mechanical Engineering*, vol. 10, pp. 1–23, 12 2018.

[10] C. Hegde, "Automation of commonsense reasoning: A study on feasible knowledge representation techniques," *International Journal of Advanced Research in Computer Science*, vol. 8, pp. 601–604, 9 2017.

[11] Z.-Y. Dou and N. Peng, "Zero-shot commonsense question answering with cloze translation and consistency optimization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 10572–10580, 6 2022.

[12] M. A. El-Salam, A. K. Hussein, and A. A. Fahmy, "Understanding a simple arabic stories using event calculus," *American Journal of Applied Sciences*, vol. 10, pp. 1298–1306, 10 2013.

[13] C. Parker, C. M. Brunswick, and J. Kotey, "The happy hen on your supermarket shelf: what choice does industrial strength free-range represent for consumers?," *Journal of bioethical inquiry*, vol. 10, pp. 165–186, 5 2013.

[14] F. Rancourt, P. Vondrlik, D. Maupomé, and M.-J. Meurs, "Investigating self-rationalizing models for commonsense reasoning," *Stats*, vol. 6, pp. 907–919, 8 2023.

[15] S. Borgwardt, I. I. Ceylan, and T. Lukasiewicz, "Aaai - ontologymediated query answering over loglinear probabilistic data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 2711–2718, 7 2019.

[16] T. D. Noia, E. D. Sciascio, and F. M. Donini, "Semantic matchmaking as non-monotonic reasoning: A description logic approach," *Journal of Artificial Intelligence Research*, vol. 29, pp. 269–307, 7 2007.

[17] S. M. Sobhy and W. M. Khedr, "Developing of fuzzy logic decision support for management of breast cancer," *International Journal of Computer Applications*, vol. 147, pp. 1–6, 8 2016.

[18] R. Krishna, D. Lee, L. Fei-Fei, and M. S. Bernstein, "Socially situated artificial intelligence enables learning from human interaction.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, pp. e2115730119–, 9 2022.

[19] E. Ludwin-Peery, N. R. Bramley, E. Davis, and T. M. Gureckis, "Broken physics: A conjunction fallacy effect in intuitive physical reasoning," *Psychological science*, vol. 31, pp. 1602–1611, 11 2020.

[20] J. Blass and K. Forbus, "Analogical chaining with natural language instruction for commonsense reasoning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2 2017.

[21] M. Fasli, "Reasoning about knowledge and belief: A syntactical treatment," *Logic Journal of IGPL*, vol. 11, pp. 247–284, 3 2003.

[22] K. Čyras, "Rational versus intuitive outcomes of reasoning with preferences: Argumentation perspective," *Inteligencia Artificial*, vol. 20, pp. 70–81, 2 2017.

[23] E. Ludwin-Peery, N. R. Bramley, E. Davis, and T. M. Gureckis, "Limits on simulation approaches in intuitive physics.," *Cognitive psychology*, vol. 127, pp. 101396–101396, 6 2021.

[24] S. Sugawara, P. Stenetorp, K. Inui, and A. Aizawa, "Aaai - assessing the benchmarking capacity of machine reading comprehension datasets," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8918–8927, 4 2020.

[25] M. Forbes, A. Holtzman, and Y. Choi, "Cogsci - do neural language representations learn physical commonsense," *Cognitive Science*, pp. 1753–1759, 8 2019.

[26] K. Shen and M. Kejriwal, "An experimental study measuring the generalization of fine-tuned language representation models across commonsense reasoning benchmarks," *Expert Systems*, vol. 40, 2 2023.

[27] A. Bundy, "Smart machines are not a threat to humanity," *Communications of the ACM*, vol. 60, pp. 40–42, 1 2017.

[28] P. L. Elkin, G. Mehta, F. LeHouillier, M. Resnick, S. Mullin, C. Tomlin, S. Resendez, J. Liu, J. R. Nebeker, and S. H. Brown, "Semantic clinical artificial intelligence vs native large language model performance on the usmle.," *JAMA network open*, vol. 8, pp. e256359–e256359, 4 2025.

[29] F. B. Akman, "Turkophobia and islamophobia in the nineteenth century western/american popular fiction: An orientalist reading of her rescue from the turks (1896)," *Erdem*, pp. 23–46, 6 2019.

[30] X. Zhang, "Minimally inconsistent reasoning in semantic web.," *PloS one*, vol. 12, pp. 0181056–, 7 2017.

[31] G. He, A. Balayn, S. Buijsman, J. Yang, and U. Gadiraju, "It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 10, pp. 89–101, 10 2022.

[32] J. H. Lee, Y. Yao, O. Özdemir, M. Li, C. Weber, Z. Liu, and S. Wermter, "Spatial relation learning in complementary scenarios with deep neural networks.," *Frontiers in neurorobotics*, vol. 16, pp. 844753–, 7 2022.

[33] C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu, "A survey on machine reading comprehension: Tasks, evaluation metrics, and benchmark datasets," *Applied Sciences*, vol. 10, pp. 7640–, 10 2020.

[34] H. Prakken, "On direct and indirect probabilistic reasoning in legal proof," *Law, Probability and Risk*, vol. 13, pp. 327–337, 8 2014.

[35] E. Saad and E. Pontelli, "A new approach to hybrid probabilistic logic programs," *Annals of Mathematics and Artificial Intelligence*, vol. 48, pp. 187–243, 5 2007.

[36] Z. Aghahadi and A. Talebpour, "Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning," *Journal of Applied Non-Classical Logics*, vol. 32, pp. 55–71, 1 2022.

[37] J. Hou, X. Wu, X. Zhang, Y. Qi, Y. Jia, and J. Luo, "Aaai - joint commonsense and relation reasoning for image and video captioning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10973–10980, 4 2020.

[38] A. A. Aladeemy, A. Alzahrani, M. H. Algarni, S. N. Alsubari, T. H. H. Aldhyani, S. N. Deshmukh, O. I. Khalaf, W.-K. Wong, and S. Aqburi, "Advancements and challenges in arabic sentiment analysis: A decade of methodologies, applications, and resource development.," *Heliyon*, vol. 10, pp. e39786–e39786, 10 2024.

[39] M. Allouch, A. Azaria, and R. Azoulay, "Conversational agents: Goals, technologies, vision and challenges.," *Sensors (Basel, Switzerland)*, vol. 21, pp. 8448–8448, 12 2021.