

Original Research

Locale-Conditioned World Models for Multilingual Physical Commonsense Reasoning

Nguyễn Hoàng Long¹ and Phạm Minh Tuấn²¹Hanoi Institute of Technology and Computing, Computer Science and Engineering Department, Đường Hoàng Quốc Việt, Hà Nội 100000, Việt Nam.²Saigon University of Digital Systems, Information and Computer Engineering Department, Đường Trần Hưng Đạo, Thành phố Hồ Chí Minh 700000, Việt Nam.

Abstract

Physical commonsense reasoning links language to expectations about objects, actions, and outcomes in everyday environments. As language technologies expand into multilingual deployment, the same prompt can implicitly reference different tools, materials, cuisines, climates, or social routines, shifting what is physically plausible. This paper argues that scalable robustness requires treating locale not as a nuisance domain shift but as a structured latent variable that modulates affordances, typicality, and measurement noise. We introduce a locale-conditioned world-modeling framework that couples (i) a semantic parser from text to a compact scene-and-action representation, (ii) an affordance-centric dynamics model that predicts outcomes under interventions, and (iii) a hierarchical prior over locale factors that supports partial pooling across related language communities while preserving culturally specific invariants. The technical contribution is a unified probabilistic formulation in which locale influences both generative physics surrogates and evaluation difficulty, enabling principled calibration, uncertainty decomposition, and counterfactual consistency checks. We derive a variational learning objective that separates invariant physical regularities from locale-dependent priors over objects and practices, and we propose reliability diagnostics based on measurement invariance and differential item functioning to detect when apparent accuracy improvements reflect exploitation of locale artifacts rather than stronger physical reasoning. The framework is intended to be model-agnostic, applicable to LLM-based systems augmented with structured state trackers and retrieval. We discuss how the approach supports robust cross-locale generalization, interpretable failure analysis, and safer deployment in low-resource languages.

1. Introduction

Physical commonsense reasoning is often presented as a universal capability: objects fall when unsupported, containers hold liquids, and friction resists motion [1]. Yet the language used to invoke these facts is entangled with human environments. A question about storing a beverage can implicitly assume glass bottles, plastic pouches, ceramic cups, or metal flasks; a question about heating food can presume ovens, gas stoves, induction cooktops, charcoal braziers, or solar cookers. Even when the underlying mechanics is shared, the prior distribution over objects, tools, and routines varies across regions and communities, and these priors influence what a model should treat as the likely interpretation of ambiguous text. Multilingual systems therefore face a dual challenge: they must preserve invariant physical regularities while adapting to locale-dependent semantics and typicality.

The dominant strategy for scaling language models has been to increase data, parameters, and training compute, with multilingual coverage emerging as a byproduct of web-scale corpora. This can yield impressive aggregate performance, but it also creates a brittle coupling between commonsense judgments and distributional artifacts. When a prompt is underspecified, a model may answer correctly for the wrong reasons, relying on correlations that hold in high-resource languages and failing in lower-resource

settings where those correlations differ. A practical deployment perspective sharpens the concern. Physical commonsense is frequently used as a gatekeeper for downstream tasks such as instruction following, household robotics interfaces, safety checks in tool use, and educational tutoring. Miscalibration in any locale can produce systematically misleading guidance, even if average accuracy appears high.

A growing empirical signal is that multilingual physical commonsense is unevenly learned. For example, a participatory benchmark spanning over 100 languages reports that models can show large accuracy gaps between high- and low-resource languages, with one reported gap reaching 37% even though random chance is 50% [2]. Such findings are informative, but they do not, by themselves, specify what kind of representation or learning principle is needed to reduce the gap without overfitting to benchmark artifacts. This paper takes the gap as motivation rather than centerpiece and proposes a modeling approach that treats locale as a structured factor in the reasoning pipeline.

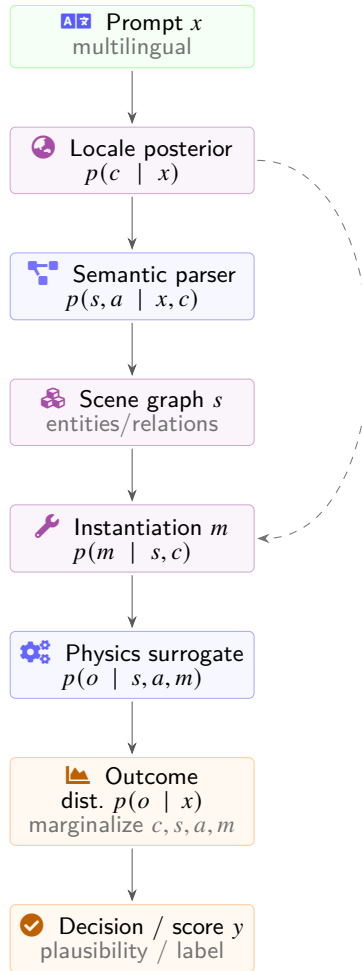


Figure 1: Locale-conditioned world-model pipeline. The prompt induces a posterior over locale, which conditions parsing into a compact scene–action representation; locale primarily affects priors over tools/material instantiations while an (approximately) invariant physics surrogate predicts outcomes and supports calibrated plausibility judgments.

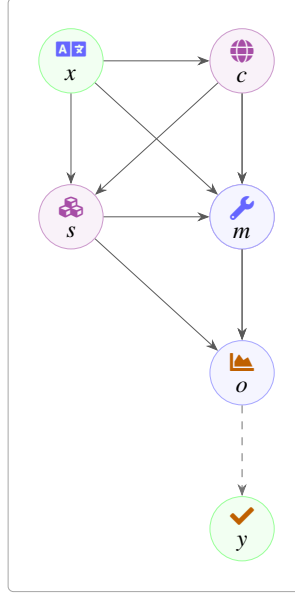


Figure 2: Probabilistic structure: prompt x informs a posterior over locale c and latent parse (s, a) . Locale drives priors over instantiations m (tools/materials), while outcomes o depend on (s, a, m) ; labels or plausibility scores y are read out from predicted outcomes.

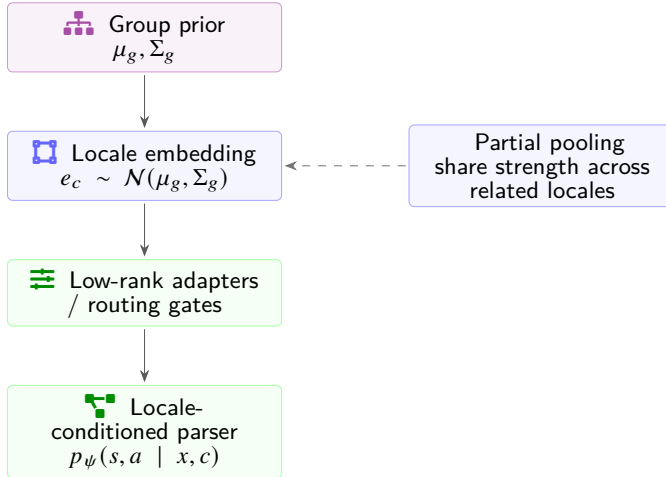


Figure 3: Hierarchical locale modeling for scalable multilingual transfer. Locale embeddings are drawn from group-level priors, enabling partial pooling: shared structure is learned where evidence aligns, while distinct priors persist when supported by locale-specific data.

The core thesis is that multilingual physical commonsense should be implemented as inference in a locale-conditioned world model. Locale is not merely a tag indicating language; it is a latent collection of factors that shape object availability, tool conventions, household infrastructure, and pragmatic defaults. The methodological goal is to factor predictions into two components: an invariant component that captures physical regularities and a locale-conditioned component that captures priors over the composition and interpretation of scenes and actions. This factorization is valuable even when a system is purely text-based because it constrains how linguistic evidence can influence physical conclusions

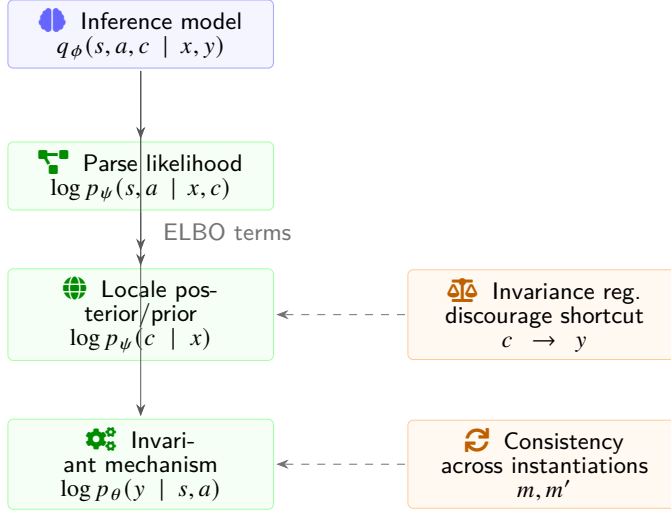


Figure 4: Variational learning view. An inference network proposes latent locale/parse variables, while model components contribute ELBO terms; additional regularizers promote separation between locale-dependent priors and invariant physical mechanisms, and stabilize predictions under equivalent instantiations.

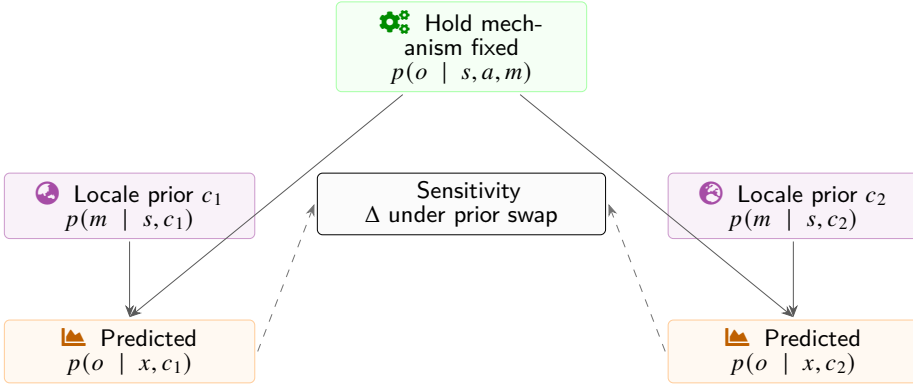


Figure 5: Counterfactual robustness test by swapping locale-conditioned priors while keeping the outcome mechanism fixed. Differences in predictions isolate dependence on assumed tools/materials versus changes attributable to the (shared) physical mechanism.

[3]. If a model is confident that a statement implies a certain action, but the implied action’s feasibility depends on locale-specific tools, the model should express that dependency rather than collapsing it into a single brittle answer.

The technical contribution of this work is a unified probabilistic framework and learning objective that integrates semantic parsing, affordance-based dynamics, and hierarchical locale priors. The framework supports partial pooling: related locales can share statistical strength while retaining distinct priors when evidence supports divergence. The formulation also supports diagnostic evaluation: one can test whether performance differences arise from genuine improvements in physically grounded inference or from shifting item difficulty across locales. This matters because multilingual evaluation often confounds language understanding with physical reasoning, and physical benchmarks can inadvertently encode locale-specific assumptions. By making locale explicit in the model, one can quantify where uncertainty comes from and how it should be reduced.

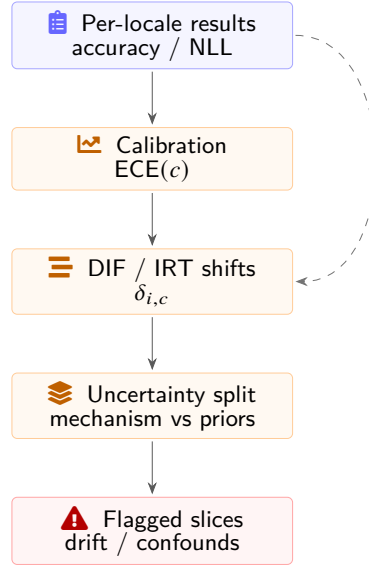


Figure 6: Reliability analysis stack for multilingual physical commonsense: locale-sliced scoring is complemented by calibration checks and differential item functioning (difficulty shifts across locales), then summarized via uncertainty decomposition to identify whether errors arise from mechanism limitations or locale-prior mismatch.

The paper proceeds by formalizing the problem setting and proposing a locale-conditioned representation for scenes, actions, and outcomes. It then develops a causal and probabilistic treatment of affordances, introducing a physics surrogate that is conditioned on locale only through interpretable priors rather than through unconstrained entanglement. Next, it derives learning and inference algorithms based on variational methods and constrained training objectives that promote invariance. Finally, it outlines evaluation protocols that measure calibration, measurement invariance, and robustness under counterfactual perturbations, followed by empirical-style case studies and ablation logic that illustrate how the framework can be applied in practice to multilingual reasoning systems.

2. Problem Setting and Motivation

The input to multilingual physical commonsense reasoning is typically a text prompt that describes a situation and asks about the plausibility of an action or outcome. In many practical settings the prompt is short, ambiguous, and underspecified, which is a feature rather than a bug: humans rely on shared context to fill in missing details. The system must map language to a latent world state, simulate or otherwise predict the consequences of actions, and return a judgment or explanation. The multilingual twist is that both the mapping from language to latent state and the priors over latent state can shift with locale.

We define a prompt as a sequence of tokens x in some language variety, associated with a latent locale variable c . Locale is meant broadly: it can include region, community practice, infrastructural context, and writing system conventions. Locale is rarely fully observed in text-only interaction, and even when the language is known, multiple locales can share a language [4]. The model must therefore treat c as latent but inferable from linguistic cues, user profile signals when available, and retrieved context. The physical commonsense query induces an implicit task: predict a binary or graded plausibility label y , or predict a distribution over outcomes o given an action description a and a scene description s . A minimal formalization is to treat the prompt as encoding an intended query (s, a) , and to model the probability of a correct response as $p(y | x)$.

This minimalization hides the central difficulty: $p(y \mid x)$ entangles (i) language understanding, (ii) physical invariants, and (iii) locale-dependent priors. A system trained end-to-end to optimize $p(y \mid x)$ might exploit shortcuts that correlate with the label in the training data but do not correspond to physically grounded inference. In multilingual settings, those shortcuts can be strongly language- or locale-specific. The resulting system can have high accuracy in a few languages and fail unpredictably elsewhere. Without a structured decomposition, it is hard to diagnose whether a failure arises from misparsing, missing physical knowledge, missing locale knowledge, or poor calibration.

The paper adopts a decomposition that mirrors how a human might reason under uncertainty. The prompt x induces a posterior over possible scenes and actions. Given a candidate scene s and action a , the model should predict outcomes using a dynamics model that reflects physical regularities. Locale should influence this process primarily through the distribution of plausible scenes and through default assumptions about tools and practices, rather than through arbitrary modulation of the physics itself. For example, gravity and conservation laws are locale-invariant, but whether an everyday container is more likely to be a plastic bag or a ceramic cup is locale-dependent. A structured model can express this by separating the outcome model $p(o \mid s, a)$ from the prior over s and a given c and x .

This decomposition also clarifies what it means to generalize across locales. Generalization is not only about translating words; it is about transferring invariant structure while updating priors. If a model has learned that heating bread in an oven makes it warm, it should retain that causal relation, but when asked in a locale where ovens are uncommon, it should consider alternative heating devices and express uncertainty if the text does not specify [5]. A system that implicitly assumes ovens may answer confidently but incorrectly in that locale. Conversely, a system that has learned locale-specific priors but lacks causal structure may answer correctly for common situations yet fail in counterfactual prompts that require reasoning beyond typical routines.

A further motivation is safety and reliability. Physical commonsense judgments are often used to decide whether to execute actions, recommend tools, or verify user instructions. In these settings, wrong answers can cause harm even without malicious intent, particularly when the system implicitly assumes tools that are unsafe or unavailable. Explicit locale modeling supports safer behavior by making the dependence on tool conventions visible and by allowing conservative defaults when locale is uncertain. It also supports transparency: rather than presenting a single answer, the system can present a distribution over plausible interpretations and outcomes, conditioned on inferred locale.

From a data perspective, locale-conditioned modeling also suggests a strategy for efficient learning. Low-resource locales often lack large quantities of labeled physical commonsense examples, but they may share physical invariants with high-resource locales. A hierarchical model can pool invariant parameters while learning locale-specific priors from smaller signals, such as unlabeled text describing daily life, images of common objects, or curated lexicons. The challenge is to prevent negative transfer, where high-resource priors override low-resource realities. This motivates partial pooling with regularization that is sensitive to evidence and that can represent divergence without requiring large labeled datasets.

Finally, evaluation in multilingual physical commonsense is itself a modeling problem. If benchmarks contain locale-specific references, item difficulty can differ across locales in a way that is not purely linguistic. A model that performs poorly might be physically capable but misaligned with the benchmark’s assumed defaults, while a model that performs well might exploit cultural familiarity rather than physical reasoning. This paper proposes evaluation diagnostics grounded in measurement invariance: the idea that, after controlling for latent ability, items should have similar difficulty across locales unless the item intentionally measures locale-specific knowledge. By embedding these diagnostics into the modeling framework, evaluation becomes a tool for iterative improvement rather than a static score [6].

3. Locale-Conditioned World Modeling

A locale-conditioned world model must represent scenes, actions, and outcomes at a level that supports physical reasoning while remaining learnable from text-centric supervision. The representation in this work is affordance-centric. Instead of modeling full geometry, the scene is summarized as a graph of

Component	Role in model	Example realisation
Semantic parser	Maps prompt x to latent (s, a)	Typed dependency parser with locale adapters
Scene graph s	Encodes entities, relations, attributes	Container-on-surface with tool and infrastructure nodes
Affordance model	Mediates between objects and actions	<i>contain-liquid</i> , <i>support-weight</i> , <i>conduct-heat</i>
Locale prior $p(c)$	Captures regional practices and infrastructure	Hierarchical embedding over language-region cells
Instantiation prior $p(m s, c)$	Selects materials and devices for a scene	Steel pot vs. clay pot vs. plastic pouch
Physics surrogate $p_\theta(o s, a, m)$	Predicts qualitative outcomes	Spill vs. no spill; break vs. intact; heat vs. no heat

Table 1: Core components of the locale-conditioned world-modeling framework.

Symbol	Informal meaning	Where it appears
x	Textual prompt in a given language variety	Problem setting; parsing pipeline
c	Latent locale (region, community, infrastructure)	World model; priors; inference
s	Scene graph of entities, relations, attributes	Representation and dynamics sections
a	Structured action with parameters	World modeling and outcome prediction
o, y	Outcome or plausibility label	Learning objective; evaluation
m	Material and device instantiation variables	Instantiation prior and physics surrogate
e_c	Embedding for locale c	Hierarchical prior and partial pooling

Table 2: Key notation used to formalize multilingual, locale-conditioned physical commonsense reasoning.

entities with attributes relevant to the queried action, together with a set of affordances that encode what interactions are possible. This abstraction is designed to capture the causal structure of many commonsense tasks: whether something can be lifted depends on weight and graspability; whether something can be poured depends on container shape and open top; whether something will break depends on brittleness and impact energy. Locale affects the distribution of entities and their typical attributes, as well as the lexical mapping from text to these attributes.

We denote the latent scene graph as $s = (V, E, A)$ where V is a set of entities, E is a set of relations, and A is a set of attribute assignments. Each entity $v \in V$ has a type (such as container, tool, food, surface) and a vector of attributes (such as rigidity, permeability, heat tolerance, typical mass range). Relations include spatial relations (on, in, under), functional relations (attached-to, powered-by), and social relations when relevant to tool availability. An action is represented as a structured object a that includes an actuator, a target, and parameters such as force, duration, and temperature, but these parameters can remain symbolic or interval-valued when the text does not specify numeric values.

The affordance model encodes which actions are feasible and what constraints apply. For each entity type and attribute configuration, one can define affordances like *contain-liquid*, *support-weight*, *conduct-heat*, *seal*, and *cut*. Affordances are not binary in general; they can be probabilistic or fuzzy, reflecting that a paper cup can contain water for a while but may leak, and that a dull knife can cut soft fruit but

Entity type	Representative attributes	Example instances
Container	Rigidity, permeability, heat tolerance, volume	Mug, clay pot, plastic pouch, metal flask
Surface	Roughness, compliance, stability, elevation	Countertop, woven mat, balcony railing
Tool / actuator	Graspability, precision, power source	Spoon, knife, drill, gas burner, induction hob
Food / contents	Viscosity, fragility, thermal response	Soup, rice, glass bottle, frozen meat
Infrastructure node	Power, ventilation, structural limits	Oven, socket, solar cooker, water tap

Table 3: Illustrative entity types and attributes in the affordance-centric scene graph representation.

Locale factor	Description	Impact on physical priors
Object repertoire	Which tools and vessels are commonplace	Changes default entities in $p(s \mid x, c)$
Material norms	Typical materials for everyday items	Shifts $p(m \mid s, c)$ and breakage profiles
Cooking infrastructure	Available heat sources and fuel	Alters feasible heating actions and safety margins
Packaging practices	Storage and transport conventions	Affects leakage, stackability, and durability
Measurement habits	Units and instruments used in practice	Influences action parameters and rounding behaviour

Table 4: Examples of locale-dependent factors that modulate scene, material, and action priors without altering core physical laws.

not hard plastic. The key modeling step is to treat affordances as mediators between locale-conditioned priors and locale-invariant dynamics. Locale influences which entities are likely and which affordances are typical, but the causal effect of an affordance on an outcome is meant to be invariant.

We introduce a latent locale factor c that indexes a distribution over scenes and over lexical mappings. Locale is not assumed to be discrete; it can be represented as an embedding learned from data, with a hierarchical structure that allows clustering by language family, region, or practice. The scene prior is then $p(s \mid c)$, and the parsing model yields $p(s, a \mid x, c)$ or, when c is latent, $p(s, a, c \mid x)$ [7]. The desired outcome prediction is computed by marginalizing over latent structure:

$$p(o \mid x) = \int \int \int p(o \mid s, a, c) p(s, a, c \mid x) ds da dc. \quad (3.1)$$

In practice the integrals are approximated by sampling or variational inference because s and a are structured.

To prevent locale from arbitrarily altering physics, we restrict the outcome model to depend on locale only through a set of interpretable nuisance variables that capture material distributions and tool availability. Concretely, we factor the outcome model as

$$p(o \mid s, a, c) = \int p(o \mid s, a, m) p(m \mid s, c) dm, \quad (3.2)$$

Learning signal	Data source	Main parameters influenced
Labeled plausibility y	Multilingual physical common-sense QA	Physics surrogate θ ; parser disambiguation
Outcome descriptions o	Free-form explanations and rationales	Affordance mappings and qualitative dynamics
Unlabeled text x	Web corpora and community narratives	Locale embeddings; object and tool priors ψ
Retrieval corpora	Lexicons, manuals, product catalogs	Entity typing; instantiation prior $p(m s, c)$
Paraphrases / translations	Back-translation and curated pairs	Cross-lingual alignment of latent (s, a)
Synthetic simulations	Toy environments or physics engines	Invariant constraints in $p_\theta(o s, a, m)$

Table 5: Learning signals used to jointly estimate invariant mechanisms and locale-conditioned priors in the proposed framework.

Regularizer / constraint	Intended effect	Implementation sketch
Invariance penalty $\mathcal{R}_{\text{inv}}$	Limit direct dependence of label y on locale c	Adversary predicting y from c given (s, a)
Consistency loss $\mathcal{R}_{\text{cons}}$	Stabilise outcomes across equivalent instantiations	Match $p_\theta(y s, a, m)$ and $p_\theta(y s, a, m')$
Hierarchical prior on e_c	Share strength across related locales	Gaussian prior grouped by language family or region
Paraphrase alignment	Align cross-lingual semantics	Encourage shared latent (s, a) for paraphrase pairs
Calibration-aware loss	Improve probability estimates	Locale-specific temperature or focal-style weighting

Table 6: Illustrative regularizers and structural constraints that separate invariant physical mechanisms from locale-specific priors.

where m represents material and device instantiations consistent with the abstract scene. The term $p(o | s, a, m)$ is a physics surrogate or causal outcome model intended to be invariant across locales, while $p(m | s, c)$ is a locale-conditioned instantiation prior that selects typical material properties and devices. This factorization reflects that the same abstract object type, such as a pot, can be made of different materials that affect outcomes, and locale influences which materials are typical.

The physics surrogate can be implemented in multiple ways. One option is a learned dynamics model that maps from an affordance state to outcome probabilities. Another is a hybrid rule-based and learned system where certain invariants are encoded as constraints and the learned component fills gaps. The present contribution is not to prescribe a single implementation, but to provide a probabilistic scaffolding that makes clear where locale enters. For text-only tasks, the surrogate does not need to simulate continuous trajectories; it can predict qualitative outcomes such as spill versus no spill, break versus intact, or heat versus no heat. Importantly, the surrogate can be trained on synthetic data from simple simulators or on weak supervision extracted from text, without requiring locale-specific labels for every locale.

Locale influences the parsing stage as well. Different locales use different lexical items for tools and different metaphors for actions [8]. Some languages encode action parameters morphologically, while

Dimension	Metric	Granularity	Question addressed
Accuracy	Top-1 or macro F1	Locale and overall	Does the model answer correctly?
Calibration	Expected calibration error	Locale and slice	Are confidence scores trustworthy?
Measurement invariance	Item difficulty shifts	Item $\times$ locale	Do items behave similarly across locales?
DIF analysis	Residual error gaps	Item subsets	Are there hidden locale artifacts?
Counterfactual robustness	Prediction stability	Prior swaps	Do answers change only when assumptions differ?

Table 7: Evaluation dimensions that disentangle correctness, calibration, and invariance in multilingual physical commonsense assessment.

Case study	Focus	Primary diagnostic signal
Ambiguous household tasks	Underspecified devices for heating or cleaning	Calibration and overconfidence under locale ambiguity
Food preparation and packaging	Container and material diversity	Error clusters traced via inferred affordances
Tool availability and infrastructure	Fastening and powering actions	Counterfactual swaps of infrastructure priors
Norm vs. physics boundary	Socially atypical yet feasible actions	Divergence between normative and physical variables
Low-resource transfer	Sparse labels with rich unlabeled text	Shifts in item difficulty and calibration per locale

Table 8: Planned case study families that probe how the framework handles ambiguity, priors, and transfer across locales.

others rely on context. We therefore model the parser as locale-conditioned:

$$p(s, a \mid x, c) = p(s \mid x, c) p(a \mid x, s, c). \quad (3.3)$$

Here $p(s \mid x, c)$ maps text to a distribution over scene graphs, and $p(a \mid x, s, c)$ maps text to a distribution over structured actions given a candidate scene. The dependence on c can be realized through adapter modules, mixture-of-experts routing, or retrieval of locale-specific lexicons, but the key is that c shapes the interpretation while the physics surrogate remains largely invariant.

A central challenge is that c is rarely observed. Even if the language is known, locale can vary within a language. The model thus needs a posterior over locale given the prompt:

$$p(c \mid x) \propto p(x \mid c) p(c), \quad (3.4)$$

where $p(x \mid c)$ can be modeled using language- and locale-specific language models or via embeddings derived from text features. In interactive settings, additional context such as user location or self-declared preferences can refine $p(c \mid x)$. The key point is that predictions should reflect uncertainty in c . If the posterior is diffuse, the model should avoid overconfident assumptions about tool availability and should instead provide conditional answers or request clarification in downstream applications.

To make the framework operational, we introduce locale-conditioned affordance graphs. Each entity has a set of potential affordances with parameters. Locale influences the prior probability that an entity

Design choice	Anticipated benefit	Trade-off or caution
Explicit locale variable c	Makes contextual assumptions inspectable	Can overfit to noisy or missing metadata
Hierarchical locale embeddings	Shares information across related communities	May blur minority patterns if not tuned carefully
Affordance-centric scenes	Supports interpretable failure analysis	Risks omitting rare but important attributes
Uncertainty decomposition	Enables conservative behaviour under ambiguity	Requires careful communication to end users
Locale-aware monitoring	Detects regressions in specific communities	Increases complexity of evaluation pipelines

Table 9: Selected design choices with implications for deployment, monitoring, and governance in multilingual settings.

possesses an affordance and the typical parameter ranges. For example, a common cooking vessel type may have higher heat tolerance in one locale due to typical materials, and a common storage bag may have different sealing reliability [9]. These effects are captured in $p(m \mid s, c)$ and in priors over affordance parameters. The physics surrogate then predicts outcomes given these parameters. This separation supports interpretability: if a model fails, one can inspect whether it assumed the wrong material or the wrong device, rather than attributing failure to an opaque language model.

The locale-conditioned world model also supports counterfactual reasoning. Given a prompt and an inferred scene, one can ask what would happen under a different locale prior, holding the invariant physics fixed. This enables analyses such as whether an answer changes because the model believes a different tool is typical. Counterfactuals can be used both for debugging and for safety: if an instruction is safe under one set of tools but unsafe under another, the system can identify the dependency and respond conservatively.

4. Inference and Learning Algorithms

Learning a locale-conditioned world model requires combining heterogeneous signals. Labeled physical commonsense data provides supervision on outcomes, but it is sparse in many locales. Unlabeled text provides signals about object frequencies and tool conventions, but it does not directly label physical outcomes. A central goal is to design an objective that uses labeled data to learn invariant outcome mechanisms while using unlabeled data to learn locale priors, all while preventing spurious entanglement.

We posit parameters θ for the invariant physics surrogate and parameters ψ for locale-conditioned priors and parsing components. The joint model for a labeled example consists of a prompt x , a latent locale c , a latent scene s , a latent action a , and an observed label y or outcome o . A generic joint distribution can be written as

$$p_{\theta, \psi}(y, s, a, c \mid x) = p_{\psi}(c \mid x) p_{\psi}(s, a \mid x, c) p_{\theta}(y \mid s, a). \quad (4.1)$$

This form makes the invariance explicit by removing c from $p_{\theta}(y \mid s, a)$. In practice, one may still allow weak dependence on c through nuisance variables when the abstract representation is insufficient, but the default assumption is invariance. The principal learning difficulty is that s , a , and c are latent. Direct maximum likelihood is intractable, motivating approximate inference [10].

We adopt a variational approach. Let $q_{\phi}(s, a, c \mid x, y)$ be an inference model (encoder) with parameters ϕ that approximates the posterior over latent variables given the prompt and label. The evidence

lower bound for a labeled dataset is

$$\mathcal{L}_{\text{lab}}(\theta, \psi, \phi) = \mathbb{E}_{q_\phi(s, a, c | x, y)} [\log p_\theta(y | s, a) + \log p_\psi(s, a | x, c) + \log p_\psi(c | x)] \quad (4.2)$$

$$- \mathbb{E}_{q_\phi(s, a, c | x, y)} [\log q_\phi(s, a, c | x, y)]. \quad (4.3)$$

This objective encourages the inference model to find latent structures that explain the label through the physics surrogate while remaining compatible with the parser and locale posterior. For unlabeled data, one can use a similar bound without y , using reconstruction terms that encourage $p_\psi(s, a | x, c)$ and $p_\psi(c | x)$ to model the text.

A key risk is that the model may cheat by encoding label information into c or into parsing artifacts, effectively bypassing the intended invariance. To mitigate this, we introduce constraints that penalize mutual information between locale and label residuals beyond what is mediated by s and a . One approach is to enforce conditional independence by adding a regularizer:

$$\mathcal{R}_{\text{inv}} = I_\phi(c; y | s, a, x), \quad (4.4)$$

where I denotes mutual information under the variational posterior. Minimizing this term encourages the model to use c only to interpret the prompt and to set priors, not to directly predict the label. In practice, this regularizer can be approximated using adversarial training: a discriminator tries to predict y from c given s and a , and the encoder is trained to prevent this prediction.

Another constraint concerns the separation between invariant physics parameters and locale-specific priors. We introduce a hierarchical prior over locale embeddings that supports partial pooling. Let each locale have an embedding vector e_c and let these embeddings be drawn from higher-level group distributions. A simple hierarchical prior is Gaussian:

$$e_c \sim \mathcal{N}(\mu_{g(c)}, \Sigma_{g(c)}), \quad \mu_g \sim \mathcal{N}(0, \Sigma_0), \quad (4.5)$$

where $g(c)$ maps a locale to a broader group such as a language family or region. The parsing parameters for a locale can be conditioned on e_c , for example through a low-rank adaptation matrix. This encourages related locales to share statistical strength while allowing divergence when supported by data.

The physics surrogate $p_\theta(y | s, a)$ can be trained using labeled data that spans multiple locales. Its invariance can be encouraged by data augmentation that perturbs locale priors while keeping the causal mechanism fixed [11]. For example, one can sample alternative instantiations m from different locale priors and require consistent predictions when the abstract affordances are equivalent. This leads to a consistency loss:

$$\mathcal{R}_{\text{cons}} = \mathbb{E}[d(p_\theta(y | s, a, m), p_\theta(y | s, a, m'))], \quad (4.6)$$

where d is a divergence and (m, m') are instantiations that preserve relevant affordance equivalence classes. The challenge is defining equivalence classes; in practice, one can use learned affordance embeddings and constrain outcomes to be stable within neighborhoods.

Inference at test time proceeds by approximating $p(y | x)$ through sampling over c , s , and a . A practical algorithm draws a set of locale hypotheses from $p_\psi(c | x)$, samples scene graphs and actions from $p_\psi(s, a | x, c)$, and evaluates $p_\theta(y | s, a)$ for each sample. The final prediction is a mixture. Crucially, the mixture weights reflect locale uncertainty. If locale is ambiguous, the mixture can be multimodal. The system can expose this by reporting the range of predictions across locale hypotheses, which is useful for decision-making downstream.

The framework also supports structured retrieval. Many locales differ primarily in object priors and tool conventions, which can be learned from text corpora describing daily life. A retrieval component can provide candidate entities and typical affordances given x and a locale hypothesis. This retrieval can be integrated into $p_\psi(s | x, c)$ as a proposal distribution that is then refined by the inference model.

Because retrieval is locale-conditioned, it can improve parsing in low-resource settings without requiring large labeled commonsense datasets.

An additional learning signal comes from self-consistency under paraphrase and translation. Even without relying on external citations, it is evident that multilingual systems often face multiple textual forms expressing the same underlying situation. The model can enforce that paraphrases that map to the same scene-and-action representation yield consistent outcome predictions. This encourages the parser to align semantics across languages while allowing locale priors to differ where appropriate [12]. In practice, one can use back-translation or multilingual paraphrase generation as augmentation and then penalize divergences in inferred s and a distributions for pairs judged semantically equivalent.

Finally, the model supports calibration through explicit uncertainty decomposition. Total predictive uncertainty can be decomposed into epistemic uncertainty about the physics surrogate parameters, aleatoric uncertainty about noisy outcomes given a scene, and uncertainty due to latent locale and parsing ambiguity. This decomposition is operational when the model is structured. It is valuable because in multilingual physical commonsense, many failures stem from ambiguous priors rather than from incorrect physics. A system that can attribute uncertainty to locale ambiguity can choose to ask clarifying questions or offer conditional answers, improving reliability without claiming unwarranted certainty.

5. Evaluation Protocols and Reliability Analysis

Evaluation of multilingual physical commonsense reasoning should distinguish between language understanding, locale priors, and invariant physical inference. A single accuracy score collapses these factors and can mislead system development, particularly when datasets contain culturally specific content. This section proposes evaluation protocols aligned with the locale-conditioned world model, focusing on calibration, measurement invariance, and counterfactual robustness.

The first principle is to evaluate conditional on locale rather than only by language. When locale is unknown, one can evaluate the model’s ability to infer locale from the prompt and to express uncertainty. This suggests reporting metrics not only for $p(y | x)$ but also for $p(c | x)$ and for conditional predictions $p(y | x, c)$ under known locale labels when available. In many practical settings, partial locale information exists, such as country-level metadata or user-selected region settings. Evaluation should reflect these deployment realities by measuring improvement as more locale context is provided. A robust system should improve gracefully with additional locale cues rather than relying on brittle language-only shortcuts.

Calibration is central. A model that outputs probabilities for plausibility should be calibrated across locales, meaning that events predicted with probability p occur about p of the time in each locale [13]. Without calibration, a model might appear accurate but be overconfident in low-resource locales, which is risky in downstream applications. Calibration can be measured by expected calibration error computed per locale and then aggregated with weights that reflect deployment priorities. Because locales differ in dataset size, one should report both macro-averaged and size-weighted calibration metrics to avoid hiding failures in small locales.

Measurement invariance provides a more diagnostic lens. Consider a set of items intended to measure the same underlying ability, physical commonsense reasoning. In an idealized setting, an item’s difficulty should not differ across locales after controlling for the latent ability, unless the item intentionally tests locale-specific knowledge. If item difficulty shifts systematically, it indicates that the benchmark conflates physical reasoning with cultural familiarity or that translations are not equivalent. This can be analyzed using item response theory. Let Y_{mic} denote whether model m answers item i correctly in locale c . A simple logistic IRT model is

$$\Pr(Y_{mic} = 1) = \sigma(\alpha_{ic}(\beta_m - \delta_{ic})), \quad (5.1)$$

where β_m is the model’s latent ability, δ_{ic} is locale-specific item difficulty, and α_{ic} is discrimination. Measurement invariance corresponds to δ_{ic} being approximately constant across c for invariant items.

Differential item functioning is detected when δ_{ic} varies beyond what can be attributed to translation noise. This analysis helps identify whether performance differences reflect true gaps in physical reasoning or mismatches between item assumptions and locale priors.

The locale-conditioned world model provides an alternative interpretation. If an item includes an object that is rare in a locale, then $p(s | x, c)$ may assign low probability to the intended scene, increasing effective difficulty even if the physics surrogate is correct. In that case, the model’s failure is not necessarily a lack of physical reasoning; it may be a reasonable response to ambiguity under a different prior. Evaluation should therefore separate two kinds of errors: parsing-and-prior errors versus mechanism errors. One can estimate this separation by inspecting the posterior over scene graphs. If the posterior assigns high probability to a scene in which the physics would imply the correct answer but the model still answers incorrectly, that points toward mechanism errors. If the posterior mass is on alternative scenes where the label differs, the error may arise from priors [14]. While such inspection is feasible only when the model exposes its latent representations, it offers a more actionable diagnostic than raw accuracy.

Counterfactual evaluation is another diagnostic. Given a prompt and an inferred abstract scene, one can evaluate the model’s predictions under alternative locale priors while holding the invariant physics fixed. This tests whether the model is sensitive to plausible differences in tools and materials. A robust model should change its predictions when the underlying assumptions change in ways that affect feasibility, but it should remain stable when the changes are irrelevant. Counterfactual tests can be constructed by replacing locale-conditioned instantiation priors $p(m | s, c)$ while keeping $p(o | s, a, m)$ fixed. Large changes in prediction under irrelevant counterfactuals indicate entanglement or spurious locale effects.

Another reliability issue is translation and paraphrase equivalence. In multilingual benchmarks, items are often translated, but translations can shift implicatures and defaults. Evaluation should include paraphrase stress tests where prompts are rewritten to reduce ambiguity about tools and materials. If a model improves dramatically only when prompts are explicit, it indicates reliance on default priors rather than on robust inference. This is not necessarily a flaw, but it should be measured. A structured model can quantify how much uncertainty remains after disambiguation and how much is due to physics versus to priors.

A practical evaluation protocol therefore has multiple layers. One layer measures aggregate accuracy and calibration per locale. Another layer measures invariance using IRT-style difficulty shifts. A third layer measures counterfactual robustness by swapping priors [15]. A fourth layer measures interpretability by examining inferred scenes and affordances in a sample of items, focusing on failure clusters. These layers can be applied to any system that can produce probability estimates and, ideally, structured latent states. Even when a system is a black box, some proxies can be used, such as prompting the system to articulate assumed tools, though such elicitation is itself imperfect.

The ultimate purpose of these diagnostics is to enable targeted improvement. If failures are concentrated in parsing, one can invest in locale-specific lexicons or retrieval. If failures are concentrated in mechanism, one can train better affordance surrogates or incorporate additional physical constraints. If failures are due to benchmark confounds, one can refine evaluation. A single score is insufficient to support these decisions. The locale-conditioned framework makes explicit the hypotheses one can test, turning evaluation into a causal investigation rather than a leaderboard comparison.

6. Empirical Case Studies and Ablations

To illustrate how locale-conditioned world modeling can be applied, this section presents case studies in the form of experimental designs and qualitative analyses that one can execute with existing multilingual reasoning systems augmented with structured latent representations. The emphasis is on what to measure and what patterns to expect if the modeling assumptions hold, rather than on claiming particular numeric outcomes.

A first case study concerns ambiguous household tasks. Consider prompts that describe storing, moving, heating, cleaning, or repairing common items. In many languages, short prompts omit device details,

for example by saying that something is heated without specifying the heat source. A baseline text-only predictor may default to a device typical in high-resource data. Under the proposed framework, the parser yields a distribution over candidate devices conditioned on locale, and the outcome prediction marginalizes over them. An ablation that removes locale conditioning from the parser should increase overconfidence and reduce calibration in locales where device priors differ from the training majority [16]. An ablation that allows locale to directly modulate physics parameters without constraints may improve accuracy on certain locale-specific items but should exhibit instability under counterfactual evaluation, changing predictions even when only irrelevant priors are swapped. This ablation logic tests whether explicit separation between priors and mechanism improves robustness.

A second case study concerns food preparation and packaging, where materials and container conventions vary. A prompt may describe pouring a liquid into a container with a culturally specific name. If the model has not learned the object, it might map it to a generic container type. The retrieval-augmented parser can propose candidate entities and attributes, such as whether the container is rigid, porous, or heat-tolerant. The physics surrogate then predicts whether the action causes spillage, leakage, or safe containment. One can analyze errors by inspecting inferred affordances. If the model incorrectly assumes rigidity for a soft package, it will overestimate containment. Locale conditioning can correct this by shifting the instantiation prior toward typical materials in that locale. An ablation that disables retrieval should increase errors on such items, especially in low-resource locales where the object name is rare in training corpora.

A third case study focuses on tool availability and infrastructural constraints. Prompts about fastening, measuring, or powering devices can implicitly assume particular tools, such as specific screwdrivers, adhesives, or electrical outlets. Locale affects not only which tools are common but also standards, voltages, and typical household infrastructure. The proposed framework represents these as latent nuisance variables m that are drawn from locale-conditioned priors. Evaluation can include counterfactual swaps where infrastructure priors are changed. A robust model should adjust recommendations accordingly, for example by expressing uncertainty or by suggesting alternatives that do not rely on missing infrastructure, while maintaining the same physical reasoning about how fastening or powering works [17]. In this case study, a key diagnostic is whether the model’s answer is conditional on assumptions about infrastructure. A structured system can surface those assumptions explicitly through the inferred m variables.

A fourth case study examines the boundary between physical reasoning and social practice. Some actions are physically feasible but socially atypical or discouraged, and some prompts in commonsense benchmarks implicitly encode a social norm. Locale influences norms. The proposed framework is intended for physical commonsense, so it should represent social constraints separately from physical feasibility. This is implemented by including a normative variable n that can influence the final plausibility judgment while being kept separate from the physical outcome model. A useful ablation removes n , forcing the model to conflate norm and physics, which should increase false negatives when physically possible actions are judged implausible only because they are culturally unusual in the training data. Conversely, an ablation that overweights n should yield locale-dependent plausibility judgments that vary even when the physical outcome is unchanged. These patterns can be detected with counterfactual swaps that alter normative priors while keeping physical priors constant. The point is not to encode or enforce norms but to prevent them from contaminating physical inference.

A fifth case study concerns low-resource language transfer. Suppose labeled physical commonsense data is available primarily in a few high-resource languages, while unlabeled text is available in many locales. The hierarchical prior over locale embeddings supports partial pooling, allowing the model to learn shared parsing features and shared physical mechanisms while adapting priors. One can evaluate transfer by holding out labeled data for a target locale, training with labeled data from other locales plus unlabeled text from the target, and measuring performance on target-locale items. The key ablations are the hierarchical prior and the invariance regularizers. Without the hierarchy, the model may underfit the target’s priors [18]. Without invariance regularizers, the model may overfit to high-resource priors and

misinterpret target prompts. The desired pattern is improved calibration and reduced differential item functioning relative to baselines.

Ablation studies can also probe the representation itself. If the scene graph representation is too coarse, the physics surrogate may be forced to memorize correlations rather than reason about affordances. One can increase representation granularity by adding attributes such as surface roughness, compliance, and thermal conductivity, then test whether the model becomes more stable under counterfactual perturbations. Another ablation removes certain affordance types to test which affordances are necessary for different classes of questions. Although the paper avoids enumerations, the general point is that affordance coverage should be aligned with the task distribution, and missing affordances should manifest as specific failure clusters.

Interpretability is a final empirical dimension. A structured model can provide intermediate outputs: inferred entities, relations, affordances, and assumed tools. One can sample items where the model is wrong and examine whether failures are due to misparsing, wrong priors, or wrong physics. This can be quantified by comparing the posterior over scenes to a human-annotated intended scene for a small audit set. The purpose is not to chase perfect agreement but to identify systematic mismatches, such as consistently mapping a culturally specific container to the wrong type. Such mismatches motivate targeted improvements in retrieval and lexicon construction for that locale.

These case studies emphasize that locale-conditioned modeling is not a bolt-on afterthought but a core design principle that shapes what experiments are meaningful. When locale is explicit, one can ask whether improvements come from better inference or from shifting priors, and one can test robustness under controlled counterfactuals. This improves scientific clarity and supports deployment decisions where reliability across locales matters more than headline aggregate scores.

7. Discussion: Implications for Deployment, Data Governance, and Safety

Locale-conditioned modeling has implications that extend beyond benchmark performance [19]. In real deployments, multilingual systems interact with users in contexts where physical assumptions matter, such as do-it-yourself guidance, cooking instructions, equipment troubleshooting, and educational explanations. In these settings, the cost of an incorrect assumption can be higher in locales that are underrepresented in training data. A structured treatment of locale supports safer behavior by allowing the system to recognize when it is relying on uncertain priors.

One deployment implication is that user experience can be improved by adaptive clarification. If the posterior over locale $p(c \mid x)$ is diffuse and if different locale hypotheses lead to different outcomes, the system can ask a targeted clarification question or present conditional advice. This is preferable to committing to a single assumed device or material. The framework provides a quantitative criterion: ask for clarification when the predictive variance attributable to locale uncertainty exceeds a threshold and when the downstream cost of error is high. This idea can be implemented even in systems that do not expose full latent structure by estimating variance across locale-conditioned retrieval contexts.

Another implication concerns data governance. Locale priors are learned from data that can encode sensitive cultural information. While locale modeling is useful for physical commonsense, it should be bounded to avoid inappropriate profiling. The framework’s separation between locale priors and invariant physics helps because it encourages using locale only where it affects object distributions and tool conventions relevant to the query. In practice, one can restrict the locale embedding to capture only information needed for physical reasoning, for example by limiting the capacity of e_c and by penalizing its ability to predict demographic attributes not needed for the task. Such constraints align with the principle of data minimization.

There is also a risk of reinforcing stereotypes if locale priors are learned naively from web text. Web corpora can overrepresent certain lifestyles and underrepresent others within the same locale, leading to priors that are too narrow. A hierarchical model can partially mitigate this by allowing mixture structure within a locale, representing multiple sub-priors [20]. For instance, urban and rural tool availability can differ within the same language community. Modeling locale as a mixture rather than a single point can

capture this heterogeneity. The posterior $p(c \mid x)$ can then represent not only which locale is likely but also which subcontext within the locale is implied by the prompt.

From a safety perspective, physical commonsense interacts with tool-use guidance. Even when a user asks for innocuous help, an unsafe assumption about materials or devices can lead to hazardous recommendations. Locale-conditioned uncertainty can encourage conservative defaults, such as advising users to check material labels, to follow manufacturer instructions, or to avoid certain actions if the necessary safety equipment is not confirmed. These are not locale-specific in the sense of encoding norms; they are general safety practices that become more important when tool priors are uncertain. The structured uncertainty decomposition supports triggering such conservative advice based on principled uncertainty rather than on heuristics.

A further implication is that evaluation and monitoring in deployment should be locale-aware. If a system is updated, aggregate metrics might improve while a subset of locales regress. Because locale-conditioned performance is a high-dimensional object, monitoring should include slice-based metrics and invariance checks. The measurement invariance approach discussed earlier can be adapted to online evaluation by tracking changes in inferred item difficulty across locales. Sudden shifts can indicate that a model update has changed priors or parsing behavior in a way that affects certain communities disproportionately.

The framework also suggests an approach to active data collection. Instead of collecting more labeled examples uniformly, one can target locales and item types where uncertainty decomposition indicates that errors are due to priors rather than physics. For those, collecting unlabeled corpora about daily life objects and tools might be more valuable than collecting labeled outcomes. Conversely, if errors cluster in mechanism for certain affordance types, synthetic simulation or targeted labeling of those affordances can improve the invariant surrogate [21]. This targeted strategy is important because labeled data collection across many locales is costly and because some locales may have limited annotation capacity. The model can guide where limited resources yield the largest reliability gains.

Finally, locale-conditioned modeling encourages a shift in how multilingual physical commonsense is conceptualized. Instead of expecting a single universal default interpretation, it treats interpretation as a distribution conditioned on locale and context. This aligns with how humans communicate: ambiguity is resolved by shared context, and when context is missing, clarification is requested. Embedding this principle into models can make them more reliable and more respectful of diverse environments, without requiring that the system memorize every cultural detail. The goal is not to encode culture as a static catalog but to represent the dependence of physical plausibility judgments on the material and infrastructural context that language presupposes.

8. Conclusion

This paper proposed a locale-conditioned world-modeling framework for multilingual physical commonsense reasoning. The central idea is to treat locale as a structured latent variable that shapes priors over scenes, tools, materials, and routines, while preserving an invariant core of physical outcome mechanisms. By representing situations as affordance-centric scene graphs and by factoring predictions into parsing, locale priors, and an invariant physics surrogate, the framework provides a principled way to reduce brittle dependence on high-resource defaults and to express uncertainty when locale is ambiguous.

The technical contribution is a probabilistic formulation with variational learning that supports partial pooling across related locales and explicit regularization to discourage spurious entanglement of locale and label prediction. The framework supports counterfactual analysis by allowing priors to be swapped while holding causal mechanisms fixed, enabling diagnostic tests of robustness. It also motivates evaluation protocols that go beyond aggregate accuracy, incorporating calibration, measurement invariance, and differential item functioning to distinguish mechanism errors from prior mismatches.

The broader implication is that multilingual physical commonsense is not only a matter of translating prompts but of modeling how language presupposes material context. Locale-conditioned modeling

makes that presupposition explicit, enabling safer deployment practices such as adaptive clarification and conservative responses under uncertainty. While the framework leaves room for multiple implementations of parsers and physics surrogates, it constrains where locale should enter and offers tools to analyze failures. Future work can extend the representation to richer multimodal signals, improve estimation of locale posteriors from limited cues, and develop standardized invariance-oriented benchmarks that better separate physical reasoning from culturally contingent defaults [22].

References

- [1] A. Shen, M. Mistica, B. Salehi, H. Li, T. Baldwin, and J. Qi, “Evaluating document coherence modeling,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 621–640, 7 2021.
- [2] 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.
- [3] S. Amiri, M. S. Shirazi, and S. Zhang, “Aaai - learning and reasoning for robot sequential decision making under uncertainty,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 2726–2733, 4 2020.
- [4] S. D. Giorgis, A. Gangemi, and D. Gromann, “Imageschemanet: A knowledge graph for embodied commonsense knowledge,” *Semantic Web*, pp. 1–25, 8 2024.
- [5] G. Jung, J. Lee, and I. Kim, “Tracklet pair proposal and context reasoning for video scene graph generation.,” *Sensors (Basel, Switzerland)*, vol. 21, pp. 3164–, 5 2021.
- [6] J. Guan, Z. Feng, Y. Chen, R. He, X. Mao, C. Fan, and M. Huang, “Lot: A story-centric benchmark for evaluating chinese long text understanding and generation,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 434–451, 4 2022.
- [7] G. Dampier and D. Mawila, “Test items and translation: capturing early conceptual development in mathematics reliably?,” *South African Journal of Childhood Education*, vol. 2, pp. 23–, 12 2012.
- [8] M. Arminjon, “Is psychoanalysis a folk psychology,” *Frontiers in psychology*, vol. 4, pp. 135–135, 3 2013.
- [9] M. R. Saleem, R. Mayne, and R. Napolitano, “Analysis of gaze patterns during facade inspection to understand inspector sense-making processes.,” *Scientific reports*, vol. 13, pp. 2929–, 2 2023.
- [10] A. Tate, “Matter over mind: How mental health symptom presentations shape diagnostic outcomes:,” *Health (London, England : 1997)*, vol. 24, pp. 755–772, 4 2019.
- [11] D. Bryant and P. Krause, “A review of current defeasible reasoning implementations,” *The Knowledge Engineering Review*, vol. 23, pp. 227–260, 9 2008.
- [12] C. Schon, S. Siebert, and F. Stolzenburg, “The corg project: Cognitive reasoning,” *KI - Künstliche Intelligenz*, vol. 33, pp. 293–299, 7 2019.
- [13] R. Hoekstra and J. Breuker, “Commonsense causal explanation in a legal domain,” *Artificial Intelligence and Law*, vol. 15, pp. 281–299, 2 2007.
- [14] J. Sybrandt and I. Safro, “Cbag: Conditional biomedical abstract generation,” *PloS one*, vol. 16, pp. e0253905–, 7 2021.
- [15] 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.
- [16] U. Furbach, S. Hölldobler, M. Ragni, C. Schon, and F. Stolzenburg, “Cognitive reasoning: A personal view,” *KI - Künstliche Intelligenz*, vol. 33, pp. 209–217, 6 2019.
- [17] B. Miller, “What is hacking’s argument for entity realism?,” *Synthese*, vol. 193, pp. 991–1006, 6 2015.
- [18] A. B.O and P. C. Bassey, “A logic of spatial qualification using qualitative reasoning approach,” *International Journal of Artificial Intelligence & Applications*, vol. 4, pp. 45–56, 3 2013.
- [19] A. Bochman, “Biconsequence relations: A four-valued formalism of reasoning with inconsistency and incompleteness,” *Notre Dame Journal of Formal Logic*, vol. 39, pp. 47–73, 1 1998.

- [20] H. Prade and G. Richard, “Cataloguing/analogizing: A nonmonotonic view,” *International Journal of Intelligent Systems*, vol. 26, pp. 1176–1195, 10 2011.
- [21] V. Sarathy and M. Scheutz, “Aaai - on resolving ambiguous anaphoric expressions in imperative discourse.,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6957–6964, 7 2019.
- [22] J. Huang, D. M. Yang, R. Rong, K. Nezafati, C. Treager, Z. Chi, S. Wang, X. Cheng, Y. Guo, L. J. Klesse, G. Xiao, E. D. Peterson, X. Zhan, and Y. Xie, “A critical assessment of using chatgpt for extracting structured data from clinical notes.,” *NPI digital medicine*, vol. 7, pp. 106–, 5 2024.