Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084877

ARTICLE

Teaching LLMs to Infer Real-World Consequences through Embodied Semantic Grounding

Manaswi Kulahara[1], Khadija Parwez[2], Faisal Alhwikem[3,*] and Fawwad Hassan Jaskani[4]

[1]Department of Geoinformatics, TERI School of Advanced Studies, Delhi, India
[2]Department of Computing and Technology, IQRA University Islamabad Campus, H-9, Islamabad, Pakistan
[3]Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
[4]Department of Computer Systems Engineering, The Islamia University of Bahawalpur, Pakistan

*Corresponding Author: Faisal Alhwikem. Email: faisal.alhwikem@qu.edu.sa

Received: 07 May 2026; Accepted: 01 July 2026; Published: 13 August 2026

ABSTRACT: Large Language Models (LLMs) have recently advanced in real-world commonsense reasoning, including understanding everyday object behaviors and inferring their attributes from text. However, they remain limited in reasoning about the *real-world consequences* of events, such as how object failures, obstructions, or structural changes affect the surrounding environment-especially without visual or sensorimotor input. Existing works like PIQA and NEWTON evaluate narrow sub-skills, such as whether an object action makes sense and whether object properties can be inferred, providing valuable benchmarks for commonsense and physical reasoning but offering limited evaluation of how events alter environmental functionality and downstream conditions. To address this gap, we propose Embodied Semantic Grounding (ESG), a framework that equips LLMs with consequence-aware text representations. ESG learns a consequence-grounded space by aligning event descriptions with affordance maps-structured representations of how an environment can be used or traversed after an event-capturing how structural changes modify environmental functionality. A Flan-T5-XL model is trained with a contrastive alignment objective to encode event descriptions into this space, for coherent prediction of consequences such as collapses, blockages, and environmental changes. Rather than introducing a new language-model architecture, ESG extends affordance-grounding with consequence-level representations of post-event environmental functionality. We evaluate ESG on a unified benchmark comprising PIQA, NEWTON, LIBERO-derived affordance text, and 2400 synthetic scenario-based tasks. Results show that ESG improves performance over baseline language models across commonsense reasoning and consequence-prediction benchmarks. Under structured affordance-map supervision, ESG improves zero-shot accuracy on PIQA and NEWTON and demonstrates improved performance on synthetic consequence-prediction scenarios designed to evaluate post-event environmental reasoning.

KEYWORDS: Environment; Flan-T5-XL; text; large language models; real-world

1 Introduction

Large Language Models (LLMs) have recently shown progress in real-world commonsense reasoning, including understanding everyday object behaviors and inferring their properties directly from text [1,2]. These findings indicate that text alone can encode fragments of real-world knowledge once thought to require perceptual or embodied grounding [3,4]. Despite this progress, LLMs remain limited in reasoning about the *real-world consequences* of events. Humans can readily infer how object failures, obstructions, or structural changes affect accessibility, stability, or the surrounding environment-even without observing the scene [2].

In contrast, LLMs often produce linguistically plausible yet physically inconsistent interpretations when asked to infer such consequences from text alone. As illustrated in Fig. 1, standard models frequently overlook how an event modifies the functional state of the environment.

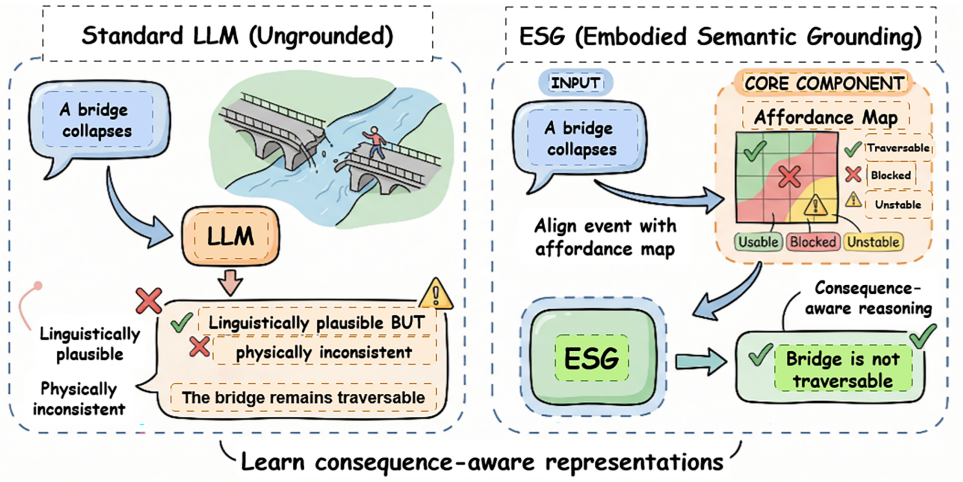

Figure 1: Illustration showing that standard LLMs often produce linguistically plausible but physically inconsistent interpretations of a bridge collapse, failing to infer its *real-world consequences*. This limitation motivates ESG, which aligns event descriptions with affordance maps to learn consequence-aware representations for coherent consequence prediction.

Existing works reveal why this gap persists. Works such as PIQA [1] test whether an object action makes sense, and NEWTON [5] assesses whether object properties can be inferred, while LIBERO-derived text benchmarks [6] examine affordance-related knowledge. While these works provide valuable benchmarks for commonsense reasoning, physical properties, and affordance-related knowledge, they provide only limited evaluation of how events alter environmental functionality-that is, how an event changes what becomes usable, traversable, blocked, or unstable, and how such changes influence connected parts of an environment without visual or sensorimotor input. Moreover, prior affordance-learning and contrastive-grounding approaches have demonstrated the value of grounding language in action-relevant representations, but are typically designed around object-level semantics, action feasibility, or static affordances rather than representations of post-event environmental functionality. This missing capability is essential for applications such as disaster assessment [7–9], accessibility analysis, environmental planning [10,11], and autonomous decision-making [12,13].

To address this gap, we introduce **Embodied Semantic Grounding (ESG)**, a framework that equips LLMs with consequence-aware text representations. Although ESG does not use physical embodiment, simulation, or multimodal sensory inputs, the term "embodied" refers to grounding textual event descriptions in functional environment-level affordance changes that approximate how agents reason about real-world environmental interactions and constraints. ESG learns a consequence-grounded space by aligning event descriptions with affordance maps [14]-structured representations of how an environment can be used or traversed after an event-capturing how structural or environmental changes modify functional conditions. Unlike existing grounding or affordance-learning approaches that focus on static affordances or action feasibility, ESG models event-induced transitions in environmental functionality through consequence-level affordance representations and structured post-event reasoning. A Flan-T5-XL[1] model is trained with a

[1] https://huggingface.co/google/flan-t5-xl

contrastive alignment objective to encode event descriptions into this consequence-grounded space, for coherent prediction of outcomes such as collapses, blockages, and environmental changes. Importantly, ESG operates entirely in the textual domain and uses affordance-map supervision as a proxy representation of post-event environmental functionality rather than direct embodied interaction or multimodal grounding. Viewed from this perspective, ESG extends existing affordance-grounding and structured-reasoning paradigms toward consequence-oriented reasoning by focusing on how events transform the functional state of an environment rather than solely on plausible actions or object attributes. In summary, our contributions are summarized as follows:

- We propose ESG, a text-only framework that learns consequence-aware representations by aligning event descriptions with affordance maps, for LLMs to infer how structural and environmental changes affect the functional state of the environment.
- We introduce consequence-level affordance grounding, which models post-event environmental state transitions for structured reasoning about usability, traversability, blockage, and stability changes beyond conventional affordance prediction or commonsense plausibility estimation.
- We evaluate ESG on PIQA, NEWTON, LIBERO-derived affordance text, and synthetic disaster scenarios. Experimental results indicate that ESG improves performance relative to baseline language models under both zero-shot and few-shot settings and supports transfer across multiple consequence-reasoning domains.

The remainder of this work is organized as follows. Section 2 provides a review of related literature. Section 3 formulates the problem and shows the key objectives. Section 4 introduces the proposed ESG framework. Section 5 describes the datasets, evaluation metrics and baseline configurations. Section 6 presents quantitative analyses of ESG across different tasks. Section 7 provides ablation studies showing the contribution of individual components. Finally, Section 8 concludes the work and discusses potential future directions.

2 Related Works

Our work is interconnected to two major research directions: (1) works examining what LLMs understand about real-world events and object behavior, and (2) works that use text in affordance-based or structured representations. These lines of work highlight why existing models cannot infer *real-world consequences* and motivate the design of ESG. In particular, existing approaches largely focus on static affordance understanding, action plausibility, or generic structured reasoning, whereas ESG targets structured modeling of how events transform the functional state of an environment through consequence-aware grounding.

2.1 Real-World Consequence Reasoning

Prior work has explored various aspects of commonsense and real-world reasoning in LLMs, but none directly target the ability to infer *real-world consequences* from text. Works such as PIQA [1] evaluate whether an object action makes sense, and SWAG [15] focuses on everyday motion prediction, revealing that LLMs capture fragments of intuitive knowledge from language alone. NEWTON [5] extends this direction by testing whether object properties-such as rigidity or brittleness-can be inferred from text. LIBERO-derived affordance text [6] examines whether models recognize affordance-related hints about how objects may be used.

Other than that, a growing body of work studies whether models can form internal world representations that support prediction of future states [7] and physical dynamics [5]. World-model research argues that intelligent systems benefit from representations that capture how environments evolve under

actions and events [7]. This perspective is exemplified by world-model approaches [16], which learn latent environment dynamics to predict future states and support internal simulation. However, their primary focus is future-state prediction rather than modeling how environmental functionality changes following an event. Related efforts in causal [13] and physical [5] reasoning investigate whether models can predict outcomes of object interactions [1], reason about temporal event chains [15], and perform forms of physical simulation from observations or textual descriptions [5]. Benchmarks such as CLEVRER [17] examine causal, explanatory, and counterfactual reasoning about physical events, while recent studies such as [5] evaluate physical commonsense [1] and environment dynamics [7] through structured prediction tasks [5]. These works provide evidence that language models acquire fragments of causal and physical knowledge from data.

Although these works provide valuable insights, they each assess narrow sub-skills-object action sense-making [1], property inference [5], affordance semantics [18], physical plausibility [15], or causal prediction [13]-rather than requiring models to infer how an event changes the functional state of an environment. Most prior benchmarks evaluate whether an outcome can be predicted [13] or explained [17], whereas comparatively little attention has been given to representing the resulting environmental state and its functional implications [7]. As a result, current LLMs can describe a collapsed bridge yet often fail to infer whether traversal becomes impossible, whether stability decreases, or whether downstream functionality is affected. This limitation becomes particularly evident when reasoning requires multi-step consequence propagation, where local changes must be tracked across multiple entities and environmental regions rather than treated as isolated outcomes. Existing benchmarks such as PIQA [1], NEWTON [5], and related reasoning tasks provide valuable supervision for action plausibility, physical properties, and causal outcomes. However, they only indirectly capture how events modify environmental functionality, leaving limited evaluation of what becomes usable, traversable, blocked, or unstable following an event, particularly without visual or sensorimotor input.

Similarly, recent structured reasoning frameworks, including graph-based reasoning approaches [13], reasoning-trajectory approaches, and structured grounding approaches, have demonstrated benefits for multi-hop inference [19], relational reasoning [17], and semantic alignment. However, their primary objective is to improve reasoning quality, planning [20], or semantic consistency rather than to model event-driven environmental state transitions. These approaches generally represent relations among entities [13] or reasoning steps, but do not capture how functional properties propagate through an environment after an event occurs [7]. These representations provide useful mechanisms for modeling entities, relations, and reasoning processes. ESG builds upon these ideas by extending them to consequence-level reasoning, where graph structures are associated with event-induced functional state transitions and their propagation across connected entities. Rather than predicting whether an event is plausible or identifying its immediate outcome, ESG represents how an event alters environmental functionality through structured consequence graphs that model state changes and their propagation across connected entities.

2.2 Affordance Maps

Another line of work studies how environments can be represented through affordances-i.e., what they allow an object to use, traverse, or interact with. Classical affordance modeling in robotics [21] and embodied AI [22] focuses on predicting which actions are possible in a given state, such as grasping an object [21], supporting weight [5], or navigating through space [20]. These representations are often encoded as affordance maps or structured descriptions of how objects or regions can be used. Related work has also explored structured representations of environments through scene graphs [23], relational world models [7], and graph-based environment representations [24] that encode entities and their interactions. Such representations provide a foundation for reasoning about object relationships [13], navigation [20], and

environment structure [24], but typically focus on describing the current state of an environment rather than modeling how that state evolves after an event. Recent text-based works attempt to extract affordance information from corpora [18,25], such as identifying which actions are associated with which objects [25] or which interactions are physically plausible [18]. However, these works primarily capture action-level affordances (e.g., what an object can be used for) rather than event-level affordances-how an event such as a collapse, blockage, or deformation changes what the environment affords afterward. As a result, affordance representations are typically anchored to a static environmental state and provide limited support for reasoning about how affordances evolve across successive state transitions following disruptive events.

Recent studies have also examined affordances in the context of language-guided agents [20], embodied planning [22], and environment understanding, where affordance representations are used to support action selection [22] and task completion [20]. While these approaches model what actions are possible in a given state, they generally do not represent how affordances change across a sequence of events or how functional consequences propagate through interconnected environmental entities following a disruptive event.

Therefore, most existing affordance maps primarily focus on how environments support actions within a given state and provide limited mechanisms for modeling how structural or environmental changes propagate into altered functionality. They do not typically specify, for example, how a collapsed bridge affects downstream traversal or how debris changes which regions remain usable. Similarly, prior graph-based and affordance-based representations [13] primarily encode object-action relations [18], spatial relations [11], or navigation constraints within a static environment [20]. They generally do not represent event-triggered transitions between functional states [7] or the cascading [13] effects that such transitions may have on connected entities and regions. Rather than introducing an entirely new form of affordance representation, ESG extends existing affordance-oriented and graph-based formulations by associating entities and relations with event-induced functional state transitions. Unlike prior affordance-learning approaches such as [18] that primarily encode static object-action relations, ESG introduces consequence-level affordance maps that represent event-induced functional transitions for structured prediction of post-event environmental states. This extension helps modeling of how functional changes propagate across connected environmental components following an event. ESG directly addresses this limitation by learning consequence-level affordance maps for inference about *real-world consequences* from text alone.

2.3 Event and Causal Reasoning in LLMs

Another related direction investigates whether language models construct internal mental models of environments and use these representations to perform forms of implicit simulation [2]. Studies on predictive world modeling [7] and simulation-based reasoning [13] examine whether models can anticipate future states by internally modeling interactions among objects [5], agents [19], and environmental conditions [7]. While such work provides evidence that language models can approximate aspects of physical and causal dynamics, the inferred representations are typically latent and are not structured around environmental functionality or consequence propagation [13]. These representations provide useful mechanisms for modeling environment dynamics and future-state prediction, but are primarily optimized for anticipating future states or observations rather than representing how functional properties of an environment change following an event.

Recent work has also explored multi-step reasoning over event sequences [13], where the objective is to infer intermediate events [17], predict future developments [15], or maintain temporal consistency across long reasoning chains. These approaches improve the ability of models to connect events across time and perform extended causal inference. However, they generally reason over sequences of events rather than over transformations of environmental state. Their primary objective is to maintain temporal consistency between

events, whereas consequence reasoning requires tracking how local state changes influence interconnected environmental components over time. As a result, the effects of an event on usability [18], accessibility [24], stability [5], or traversability [20] are rarely represented as reasoning targets.

A central distinction between prior event-reasoning research such as [13] and ESG lies in the representation of consequences themselves. Existing approaches typically formulate consequences as textual predictions [15], generated explanations [17], or future event hypotheses [13]. In contrast, ESG treats consequences as structured functional state changes associated with entities and environmental regions. Rather than introducing a completely new reasoning paradigm, ESG builds upon existing event- and causal-reasoning formulations by extending the reasoning target from future-event prediction to structured representations of post-event functionality. This formulation allows reasoning over how local event effects influence downstream components and environmental conditions, rather than focusing solely on prediction of subsequent events.

2.4 Structured Affordance and Graph Representations

Structured representations have been widely used to support reasoning through modeling of entities [3], relations [13], and environmental structure [24]. Prior work has explored knowledge graphs [3], scene graphs [23], relational world models [7], and graph-based reasoning frameworks [13] as mechanisms for organizing information and supporting multi-step inference. These representations provide structured abstractions that facilitate relation tracking [13], compositional reasoning [19], and information aggregation across multiple entities [3]. Their success has motivated extensive use of graph-based representations [13] for reasoning tasks that require modeling of complex relational dependencies.

A related line of work focuses on affordance-oriented representations [18], which describe the actions that objects [21], agents [22], or environments [20] support under particular conditions. Affordance representations have been employed in robotics [21], embodied planning [22], and environment understanding to support action selection [22], navigation [20], and task execution [6]. Recent efforts have also investigated extracting affordance information from text [18,25] and grounding affordance knowledge [26] in structured representations. These approaches characterize what actions are possible in a given state, but generally do not model how affordances evolve following disruptive events or environmental changes [7].

Despite their effectiveness, prior graph-based [13] and affordance-based [18] representations primarily encode static relations [13], object attributes [5], navigation constraints [20], or action possibilities [18] within a particular environmental state. These representations provide effective mechanisms for organizing relational knowledge, modeling entity interactions, and supporting structured reasoning. However, they are primarily designed to characterize environments in their current state rather than represent how environmental functionality evolves following an event. The resulting structures are largely descriptive and are not designed to represent event-induced functional transitions [7] or the propagation of consequences across interconnected entities [13]. Rather than introducing an entirely new graph formalism, ESG extends graph-based and affordance-oriented representations by associating nodes and relations with event-induced functional state transitions. In contrast, ESG employs a structured consequence graph in which nodes and relations are associated with event-driven functional changes. This extension helps reasoning over how local functional changes influence connected entities and environmental regions after an event. This formulation supports consequence-level reasoning over post-event environmental states rather than reasoning solely over static structure or action affordances.

3 Problem Statement

We formalize *real-world consequence* reasoning as the task of predicting how an event changes the functional state of an environment, using text alone. Let X denote a natural text description of an environment, and let E denote a textual description of an event that alters it (e.g., a collapse, blockage, or structural deformation). The model must infer how the event affects what in the environment becomes usable, traversable, blocked, or unstable. We represent these functional changes using a consequence-level affordance map encoded as a graph via $G = (V, A)$, where V is the set of entities in the environment and A is the set of affordance relations describing how each entity can be used or traversed after the event. Each graph is constructed by extracting environment entities and associating them with post-event affordance transitions inferred from the event description, producing a structured representation of how functional conditions change after the event. Each affordance relation is expressed as a triple via Eq. (1), where o_i is an entity, r_k is a functional relation (e.g., SUPPORTS, BLOCKS, TRAVERSABLE), and a_j is the state of that affordance after the event (e.g., INTACT, BLOCKED, UNSTABLE). *The goal is to learn a function* $f_\theta : (X, E) \rightarrow \hat{G}$, *that maps an environment description and an event description to a predicted consequence-level affordance map* $\hat{G}$ that reflects the environment's functional state after the event.

$$(o_i, r_k, a_j) \tag{1}$$

Given a training dataset $\mathcal{D} = \{(X_i, E_i, G_i^*)\}_{i=1}^N$ containing environment descriptions, event descriptions, and reference consequence graphs, the objective is to learn model parameters θ that minimize the discrepancy between the predicted graph $\hat{G}$ and the reference graph G^*. Formally, the learning objective is defined as per Eq. (2), where $\mathcal{L}$ denotes the training objective used to align predicted consequence representations with reference affordance structures.

$$\theta^* = \arg\min_\theta \mathcal{L}(G^*, \hat{G}) \tag{2}$$

At inference time, the model receives only (X, E) and predicts a consequence graph $\hat{G} = f_\theta(X, E)$ that captures the post-event functional state of the environment. The construction of consequence graphs and affordance transitions from textual descriptions is described in Section 5. Given a reference affordance map G^*, performance is evaluated by measuring how accurately the model predicts the correct affordance relations and overall graph structure (see Section 5.2).

4 Methodology

The ESG framework allows an LLM[2] to infer the *real-world consequences* of an event from text alone. ESG operates in two stages (see Fig. 2). Stage 1 (Grounding) learns a consequence-grounded representation by aligning textual event descriptions with consequence-level affordance maps. Stage 2 (Prediction) uses this grounded representation to predict an updated affordance map that reflects how the event changes the functional state of the environment.

4.1 Stage 1: Grounding

The goal of Stage 1 is to embed textual event descriptions into a space that reflects their consequence-level affordances-that is, how an event changes what is usable, traversable, blocked, or unstable in the environment. This stage establishes a shared representation space in which events with similar *real-world consequences* are placed near one another. To begin, ESG constructs a unified textual input that describes the

[2] ESG is model-agnostic and can be instantiated with any encoder-decoder language model; we use Flan-T5-XL in our experiments.

environment and the event. Given an environment description X and an event description E, we form Eq. (3), and encode it using the LLM[3] via Eq. (4).

$$T = \text{concat}(X, [\text{EVENT}], E) \tag{3}$$

$$H = h_\theta(T) \in \mathbb{R}^{L \times d} \tag{4}$$

To ground this textual representation in real-world functionality, ESG aligns H with the embedding of a target consequence-level affordance map G^*, derived[4] from PIQA [1], NEWTON [5], and LIBERO-based affordance text [6]. Then, let $g_\phi(G^*)$ denote the embedding of this map. ESG then learns a consequence-grounded space by encouraging H to be closer to $g_\phi(G^*)$ than to embeddings of incorrect (physically inconsistent) affordance maps via Eq. (5).

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\text{sim}(H, g_\phi(G^*))/\tau)}{\sum_{G^-} \exp(\text{sim}(H, g_\phi(G^-))/\tau)} \tag{5}$$

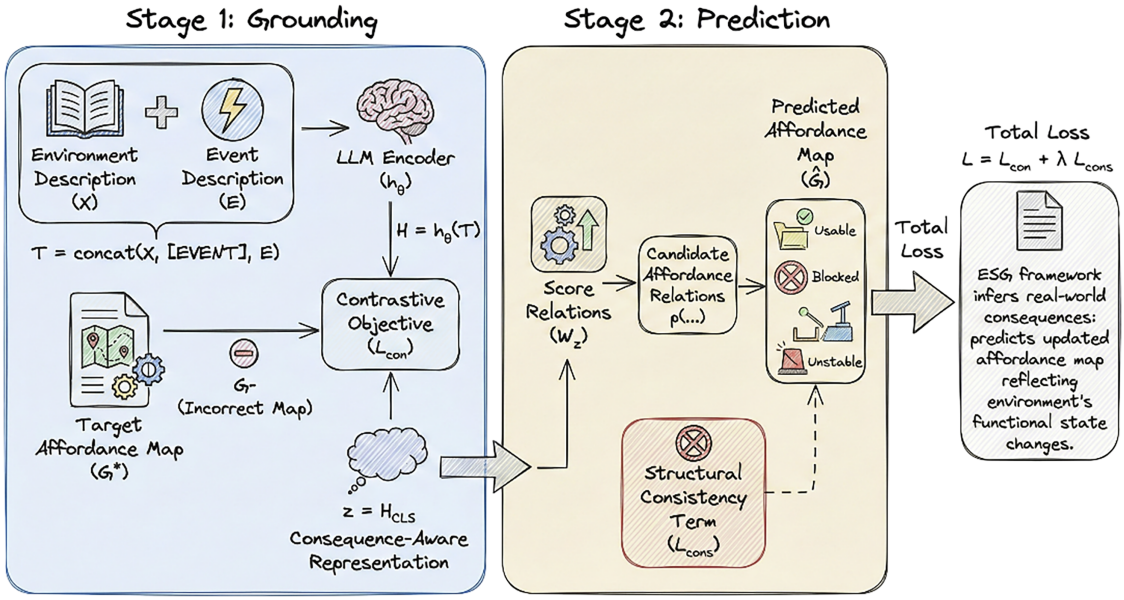

Figure 2: Overview of the ESG framework. Stage 1 (grounding) aligns environment and event descriptions with consequence-level affordance maps, placing textual inputs into a consequence-grounded representation space. Stage 2 (prediction) uses this representation to construct the predicted affordance map $\hat{G}$, capturing how the event changes what in the environment becomes usable, traversable, blocked, or unstable.

This contrastive objective assures that the learned embedding captures how the event modifies the functional state of the environment. Negative affordance maps are constructed by pairing an event description with consequence maps originating from unrelated or physically incompatible events within the same

[3] ESG does not rely on task-specific prompts or template engineering. The LLM encoder processes the raw concatenation of the environment description and event description, and the consequence-grounded structure is learned entirely through the contrastive alignment objective.

[4] Although PIQA, NEWTON, and LIBERO do not provide post-event affordance maps, their annotations contain implicit cues about object usability, traversability, and structural change. ESG converts these annotations into structured consequence-level affordance maps by extracting entities and inferring their affordance transitions (e.g., TRAVERSABLE→BLOCKED, SUPPORTS→UNSTABLE). This construction is necessary because no existing work supplies structured supervision for how events alter the functional state of an environment. Entity extraction is performed using dependency-based parsing of object mentions and environment regions, after which rule-guided transition assignment maps event descriptions into valid affordance-state updates. To reduce noisy supervision, generated affordance maps are filtered using consistency constraints and manually inspected on sampled instances to verify alignment between the textual event and the inferred post-event functionality.

mini-batch, forcing the model to distinguish valid functional transitions from inconsistent ones. We extract the resulting consequence-aware representation as $z = H_{\text{CLS}}$, which is passed to Stage 2 for affordance map prediction.

4.2 Stage 2: Prediction

Stage 2 uses the consequence-grounded representation z to construct the predicted consequence-level affordance map $\hat{G}$, corresponding to the mapping defined in Section 3. Unlike Stage 1, which relies on G^* during training, Stage 2 must infer the updated functional state without any reference graph at inference time. Using z, the LLM[5] assigns a score to each possible consequence relation-i.e., each candidate change in usability, traversability, blockage, or stability. For an entity o_i, a functional relation r_k, and an affordance state a_j, the LLM defines via Eq. (6), where the softmax runs over all candidate affordance relations.

$$p((o_i, r_k, a_j) \mid z) = \text{softmax}(Wz) \tag{6}$$

The predicted consequence-level affordance map is then obtained by selecting the configuration of relations with the highest overall probability via Eq. (7).

$$\hat{G} = \arg\max_{G} \sum_{(o_i, r_k, a_j) \in G} p((o_i, r_k, a_j) \mid z) \tag{7}$$

To assure that $\hat{G}$ reflects physically plausible functional changes, ESG adds a consistency term that penalizes relations violating known consequence patterns (e.g., marking a collapsed support as traversable). Let C denote the library of valid affordance transitions, where the structural consistency term is denoted via Eq. (8).

$$\mathcal{L}_{\text{cons}} = \frac{1}{|A|} \sum_{(o_i, r_k, a_j) \in \hat{G}} \mathbf{1}[(o_i, r_k, a_j) \notin C] \tag{8}$$

The transition library C is constructed from recurring affordance patterns observed across the converted datasets and synthetic scenarios, encoding valid post-event state changes such as USABLE→BLOCKED and STABLE→UNSTABLE. This constraint prevents logically inconsistent predictions and improves global coherence of the predicted consequence map. The final ESG objective combines the grounding loss from Stage 1 and the consistency regularizer from Stage 2 via Eq. (9), where λ controls the influence of the structural consistency constraint.

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda \mathcal{L}_{\text{cons}} \tag{9}$$

5 Experimental Setup

5.1 Datasets

ESG is trained and evaluated on a unified collection of text-only datasets that are converted into consequence-level affordance maps (see Table 1). We build on three public datasets-**PIQA** [1], **NEWTON** [5], and **LIBERO-derived affordance text** [6]-and augment them with a set of synthetic disaster scenarios to test consequence reasoning. **PIQA** [1] provides short everyday situations and candidate actions. We treat the correct action as implying a change in usability or traversability (e.g., whether a tool or surface can be used in

[5] Stage 2 does not use prompt-based scoring. The LLM computes relation scores through a learned projection layer rather than verbal prompts or manual templates.

a particular way) and convert each instance into affordance relations over the relevant entities. **NEWTON** [5] focuses on inferring object properties such as rigidity, brittleness, or elasticity from text. These properties are mapped to affordance states that capture how objects behave under stress or failure, yielding supervision for how material changes affect functional conditions (e.g., whether an object continues to support weight). **LIBERO-derived affordance text** [6] consists of language associated with embodied manipulation tasks. We extract affordance relations from these descriptions-such as which objects can be grasped, pushed, or used as supports-and encode them as pre-and post-action affordance states. **Synthetic disaster scenarios** consist of 2400 expert-designed textual descriptions of environments and destruction events (e.g., bridge collapses, road blockages, landslides). For each scenario, we annotate a consequence-level affordance map specifying how the event changes what in the environment becomes usable, traversable, blocked, or unstable. These scenarios are not used to pretrain the LLM and are held out for evaluation of ESG's ability to generalize to unseen consequence patterns.

Table 1: Dataset statistics used for training and evaluating ESG. Public datasets (PIQA, NEWTON, LIBERO) are converted into consequence-level affordance maps. Synthetic disaster scenarios provide supervision for event-consequence reasoning.

Dataset	# Samples	Type of Knowledge	Affordance Signals
PIQA [1]	21,293	Action Plausibility	Usability/Traversability
NEWTON [5]	32,000	Material + Attribute Reasoning	Support/Stability Changes
LIBERO-text [6]	130,000	Embodied Manipulation Affordances	Object Interaction Affordances
Synthetic Disaster Scenarios	2400	Event→Consequence Reasoning	Blocked/Unstable/Traversable States
Total	185,693	—	—

For all public datasets, we follow an 80/10/10 of train/validation/test split. Disaster scenarios are split so that no scenario template appears in both train and test, ensuring evaluation on genuinely novel event–consequence configurations.

5.2 Evaluation Metrics

We evaluate ESG along two complementary axes: (1) task-level accuracy on standard datasets (PIQA [1] and NEWTON [5]), and (2) affordance-level correctness on consequence-level affordance maps (LIBERO-derived text [6] and synthetic disaster scenarios). All metrics are grounded in the formulation of Sections 3 and 4.

For PIQA [1] and NEWTON [5], we follow the standard multiple-choice evaluation protocol and report **Accuracy (Acc)** via Eq. (10). This measures whether ESG, when queried in a task-specific format, selects the correct object action (PIQA [1]) or property (NEWTON [5]). These scores quantify how consequence-aware grounding affects performance on established real-world reasoning datasets.

$$\mathrm{Acc}_{\mathrm{task}} = \frac{\#\ \text{correct predictions}}{\#\ \text{total instances}} \tag{10}$$

For LIBERO-derived text [6] and synthetic disaster scenarios, the primary goal is to predict how an event changes the functional state of the environment, represented as a consequence-level affordance map $\hat{G}$. Given a reference map G^* with affordance relations (o_i, r_k, a_j), we measure **Affordance Relation Accuracy (ARA)** via Eq. (11). This metric captures how well ESG recovers the set of affordance changes-what becomes usable, traversable, blocked, or unstable-after an event.

$$\mathrm{Acc}_{\mathrm{aff}} = \frac{1}{|G^*|} \sum_{(o_i, r_k, a_j) \in G^*} \mathbf{1}\left[(o_i, r_k, a_j) \in \hat{G}\right] \tag{11}$$

In addition to relation-level performance, we also report a stricter **Graph-Level Consequence Accuracy (GLCA)**, which measures whether the entire predicted map matches the ground truth as shown in Eq. (12), where N is the number of evaluated environments. This metric is particularly informative for synthetic disaster scenarios, where each instance describes a coherent environment-event pair with a well-defined consequence map. High graph-level accuracy indicates that ESG not only predicts individual affordance relations correctly, but also reconstructs a globally consistent picture of the post-event environment.

$$\mathrm{Acc}_{\mathrm{graph}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\left[\hat{G}_n = G_n^*\right] \tag{12}$$

All metrics are reported as percentages. Higher scores indicate better performance and are denoted using ↑. In all result tables, the best performance for each setting is highlighted in green, while lower values are indicated with ↓ where appropriate.

5.3 Hyperparameters

We instantiate ESG with a FLAN-T5-XL encoder-decoder backbone and fine-tune all parameters jointly on the unified corpus described in Section 5.1. Unless otherwise noted, the same training configuration is used across all experiments to assure a fair comparison with baseline models. We optimize all models with AdamW, using a learning rate of 2×10^{-5}, linear warmup over the first 5% of training steps, and linear decay thereafter. We use a global batch size of 32 and train for up to 5 epochs, selecting the checkpoint with the highest validation ARA (see Eq. (11)) on the held-out split of the synthetic disaster scenarios. Gradient norm is clipped at 1.0 to stabilize training. For the contrastive grounding loss $\mathcal{L}_{\mathrm{con}}$ (see Eq. (5)), we set the temperature parameter to $\tau = 0.07$. At each update, negative graphs G^- are sampled from other instances within the same mini-batch, which encourages the model to separate event descriptions that lead to different consequence-level affordance maps. For the structural consistency term $\mathcal{L}_{\mathrm{cons}}$ (see Eq. (8)), we set the weighting coefficient to $\lambda = 0.5$, chosen via a coarse grid search over $\lambda \in \{0.1, 0.3, 0.5, 0.7\}$ on the validation set (see Section 7.1). Furthermore, hyperparameters are tuned using a small grid over learning rates $\{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}\}$ and batch sizes $\{16, 32\}$ (see Section 7.1).

For consequence-level affordance prediction, ESG uses the encoder representation z extracted from the final hidden state of the FLAN-T5-XL encoder. Rather than autoregressively generating graph triples through the decoder, ESG formulates affordance prediction as a structured multi-class relation classification problem over a predefined vocabulary of affordance relations and states. Specifically, each candidate triple (o_i, r_k, a_j) is represented through learned embeddings of entities, relations, and affordance states, and a lightweight projection layer computes compatibility scores conditioned on z. The softmax in Eq. (6) is therefore applied over candidate affordance transitions rather than over free-form generated text. This design was chosen to improve structural consistency and reduce invalid or physically implausible graph generations that can arise from unconstrained autoregressive decoding. The decoder component of FLAN-T5-XL is retained during training for parameter consistency with the underlying encoder-decoder architecture, but ESG primarily relies on encoder-side representations for consequence-level graph prediction. Therefore, the framework should be interpreted as a structured classification-based grounding model rather than a fully generative graph-construction system. This formulation simplifies optimization, improves stability across training runs for direct incorporation of structural consistency constraints through Eq. (8). All metrics are computed on the corresponding validation splits and then fixed for final test-time evaluation.

5.4 Baselines

We evaluate ESG against six State-of-the-Art (SOTA) LLMs, grouped into two complementary regimes: zero-shot and few-shot. In the zero-shot setting, we compare ESG to three frontier general-purpose LLMs-GPT-4.1[6], Claude 3 Sonnet[7], and Llama-3-8B-Instruct[8]-each queried without examples to assess their inherent ability to infer post-event consequences from text alone. In the few-shot setting, we benchmark against three additional models that receive limited supervision but do not use ESG's affordance-map grounding-a standard fine-tuned T5-XL model (identical backbone without grounding), FLAN-T5-XL adapted with $k = 8$ in-context examples per dataset, and Llama-3-8B-Instruct provided with 5–10 demonstration examples. To assure a fair comparison, we evaluate ESG under both zero-shot and few-shot conditions using the same inference protocol applied to competing models, while additionally including a fine-tuned T5-XL baseline with the identical backbone but without consequence-grounded affordance supervision. This comparison isolates the contribution of ESG's grounding mechanism from gains attributable solely to model scale, parameter count, or supervised adaptation. Furthermore, few-shot baselines are provided with the same task inputs, candidate outputs, and dataset splits used by ESG, for that performance differences primarily reflect the ability to model structured consequence-level affordance transitions rather than differences in data access or evaluation conditions.

6 Results and Analysis

6.1 Comparison with State-of-the-Arts

Tables 2 and 3 present a comparison between ESG and six SOTA LLMs under both zero-shot and few-shot evaluation settings. Across the reported datasets and metrics, ESG attains the highest scores among the evaluated models, including frontier commercial models (GPT-4.1, Claude 3 Sonnet) and leading open-source models (Llama-3-8B-Instruct). These results suggest that consequence-aware grounding provides complementary information beyond that available through general-purpose pretraining or instruction tuning alone.

Table 2: Comparison with SOTA LLMs in zero-shot and few-shot settings on PIQA and NEWTON. Accuracy (↑) (%) is reported. Best results are highlighted in green.

Model	Setting	PIQA	NEWTON
Zero-Shot Baselines			
GPT-4.1	Zero-shot	77.5	68.2
Claude 3 Sonnet	Zero-shot	75.9	66.8
Llama-3-8B-Instruct	Zero-shot	72.3	63.1
ESG (Ours)	Zero-shot	84.3	79.6
Few-Shot Baselines			
T5-XL (No Grounding)	Few-shot	78.1	70.4
FLAN-T5-XL (8-shot)	Few-shot	80.2	72.1
Llama-3-8B-Instruct (5–10 shot)	Few-shot	81.5	73.6
ESG (Ours)	Few-shot	86.1	81.0

[6] https://openai.com/index/gpt-4-1/
[7] https://www.anthropic.com/news/claude-3-5-sonnet
[8] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Table 3: Comparison with SOTA LLMs in zero-shot and few-shot settings for affordance-level evaluation on LIBERO-text and synthetic disaster scenarios. Unlike PIQA and NEWTON, these benchmarks require prediction of structured consequence graphs rather than discrete task labels. Therefore, evaluation is performed using ARA (↑) and GLCA (↑), which measure local affordance-relation correctness and graph-level consequence consistency, respectively. All metrics are reported as percentages. Best scores in each column are shown in green.

Model	Setting	ARA	GLCA
Zero-Shot Baselines			
GPT-4.1	Zero-shot	54.1	31.4
Claude 3 Sonnet	Zero-shot	52.7	29.6
Llama-3-8B-Instruct	Zero-shot	48.0	22.4
ESG (Ours)	Zero-shot	66.8	44.1
Few-Shot Baselines			
T5-XL (No Grounding)	Few-shot	57.3	33.5
FLAN-T5-XL (8-shot)	Few-shot	59.2	35.0
Llama-3-8B-Instruct (5–10 shot)	Few-shot	60.7	36.2
ESG (Ours)	Few-shot	69.4	47.5

In the *task-level evaluations* on PIQA and NEWTON, ESG improves zero-shot accuracy by +6.8% to +12.0% on PIQA and +11.4% to +16.5% on NEWTON relative to the strongest competing models. Even when few-shot demonstrations are provided, ESG maintains higher performance than the best few-shot baseline, with improvements of +4.6% on PIQA and +7.4% on NEWTON. These results are consistent with the hypothesis that grounding event descriptions in affordance-oriented representations can support reasoning about physical plausibility, object properties, and event outcomes. The benefits of ESG are also observed in the *affordance-level evaluations* on LIBERO-derived text and synthetic disaster scenarios, which require predicting fine-grained functional changes in the environment. ESG exceeds the strongest zero-shot baseline by +12.7% ARA and +12.7% GLCA, and the strongest few-shot model by +8.7% ARA and +11.3% GLCA. Given that the disaster benchmark is synthetically constructed, these results should be interpreted as evidence of improved consequence reasoning within the evaluated settings rather than as a direct measure of real-world deployment performance.

6.2 Cross-Domain Analysis

Across all train-test configurations shown in Table 4, ESG generally shows stronger cross-domain transfer than T5-XL (No Grounding), FLAN-T5-XL (8-shot), and Llama-3-8B-Instruct (8-shot). When trained on PIQA, ESG obtains 63.4%/27.8% on PIQA→PIQA and transfers to NEWTON (56.1%/20.7%), LIBERO (58.8%/23.1%), and Disaster (43.5%/18.9%). Similarly, when trained on NEWTON, ESG reaches 71.4%/32.8% on NEWTON while transferring to PIQA (61.0%/25.9%), LIBERO (61.5%/24.8%), and Disaster (46.0%/20.3%). Training ESG on LIBERO yields 74.2%/31.6% on LIBERO→LIBERO and competitive performance on the remaining domains. Likewise, ESG trained on Disaster scenarios achieves 74.9%/36.7% on Disaster→Disaster while maintaining performance on PIQA (58.3%/24.5%) and LIBERO (62.7%/25.0%). As expected, performance generally decreases under cross-domain evaluation relative to in-domain testing, indicating that consequence representations are transferable but remain sensitive to domain shifts in language, object distributions, and environmental conditions.

Table 4: Cross-domain generalization using ARA (↑)/GLCA (↑). For cross-domain evaluation, PIQA and NEWTON are converted into consequence-oriented affordance maps and are therefore evaluated using ARA and GLCA rather than their original task-accuracy metrics. All datasets are evaluated in their affordance-map converted form (PIQA, NEWTON, LIBERO-text) or native form (Disaster). Each model is trained on one domain (rows) and tested on another (columns). Only trainable models are included because cross-domain evaluation requires training on a source domain and testing on a target domain. Closed-source API models (GPT-4.1 and Claude 3 Sonnet) do not support domain-specific training and are therefore excluded. Best scores in each column are shown in green.

Training Domain	Test: PIQA	Test: NEWTON	Test: LIBERO	Test: Disaster
T5-XL (No Grounding)				
Train on PIQA	52.1/18.4	41.2/12.1	44.0/15.7	28.9/10.3
Train on NEWTON	49.5/16.2	58.0/22.4	46.3/16.9	31.7/11.8
Train on LIBERO	47.8/15.5	45.1/14.2	61.0/24.3	34.5/13.1
Train on Disaster	46.0/15.0	44.3/13.0	47.9/18.0	57.3/26.1
FLAN-T5-XL (8-shot)				
Train on PIQA	58.6/22.5	48.3/15.8	51.7/18.8	36.2/14.7
Train on NEWTON	55.4/21.1	65.3/27.6	53.8/20.3	39.7/16.0
Train on LIBERO	53.2/20.4	51.1/17.4	69.4/28.5	42.8/17.5
Train on Disaster	52.0/19.8	50.4/17.0	55.6/21.3	66.1/30.8
Llama-3-8B-Instruct (8-shot)				
Train on PIQA	55.7/20.3	46.8/14.6	48.9/17.4	33.9/12.8
Train on NEWTON	53.0/19.4	62.1/24.1	50.4/18.2	36.5/14.2
Train on LIBERO	51.3/18.4	49.9/16.0	66.2/26.8	39.4/15.0
Train on Disaster	50.2/18.7	48.9/16.1	52.7/19.0	63.8/27.9
ESG (Ours)				
Train on PIQA	63.4/27.8	56.1/20.7	58.8/23.1	43.5/18.9
Train on NEWTON	61.0/25.9	71.4/32.8	61.5/24.8	46.0/20.3
Train on LIBERO	59.3/24.1	57.0/19.9	74.2/31.6	48.8/21.4
Train on Disaster	58.3/24.5	54.0/19.6	62.7/25.0	74.9/36.7

FLAN-T5-XL (8-shot) consistently outperforms T5-XL (No Grounding) across most training and evaluation domains, indicating improved transfer after limited demonstration-based adaptation. Llama-3-8B-Instruct (8-shot) further improves upon the non-grounded baseline in several settings, particularly on in-domain evaluations. Across the evaluated train-test configurations, ESG attains the highest ARA and GLCA scores in the reported experiments. However, the magnitude of improvement varies across domains, with smaller gains observed in some cross-domain settings than in corresponding in-domain evaluations.

It is important to note that closed-source models (GPT-4.1, Claude 3 Sonnet) are *not* included in this cross-domain table. Cross-domain evaluation requires training the model on one dataset and then testing on another. However, closed-source models cannot be fine-tuned or trained within our experimental pipeline-they can only be queried at inference time. Because they do not support parameter updates or dataset-conditioned training, they cannot participate in a "Train on A/Test on B" protocol. Therefore, only models capable of fine-tuning or adaptation through demonstration-based training are included in this cross-domain generalization setting.

6.3 Computational Analysis

Table 5 summarizes the computational requirements of ESG across different hardware platforms and evaluation settings. Despite being built on a FLAN–T5-XL backbone, ESG remains computationally feasible even on edge devices. On a Jetson Xavier NX, INT8-quantized ESG performs zero-shot inference in 5.9 s and few-shot inference in 8.4 s, requiring only 5.3 GB of memory, indicating that consequence-aware reasoning can be executed on low-power embedded hardware under the evaluated settings. On an RTX 4090, ESG performs real-time inference, producing zero-shot predictions in under a second (0.94 s) and few-shot predictions in 1.63 s, with moderate memory use (13–14 GB). Among the evaluated platforms, the A100 80 GB provides the fastest throughput, achieving 51 and 78 ms for zero-shot and few-shot inference, respectively. Full ESG training is feasible on all three devices: while the Jetson Xavier NX requires 94.2 h for 5 epochs due to limited compute bandwidth, the RTX 4090 completes training in 41.7 h, and the A100 reduces this to 15.3 h. Across all settings, the FLOPs scale consistently with sequence length-1.36×10^{12} FLOPs per zero-shot query and 2.41×10^{12} per few-shot query-highlighting that ESG's computational footprint is dominated by transformer forward passes rather than the affordance-mapping components.

Table 5: Computational cost of ESG across hardware and evaluation settings. All estimates are computed using representative input lengths and consequence graph sizes derived from the synthetic Disaster benchmark, which contains the largest environmental state graphs evaluated in this work. FLOPs denote per-query inference FLOPs (in 10^{12}). The reported estimates are intended to characterize the computational requirements of ESG under comparable evaluation conditions across hardware platforms. All metrics are lower-is-better (↓). Best scores in each column are shown in green.

Setting	Jetson Xavier NX (8 GB, INT8)			RTX 4090 (24 GB, FP16)			A100 (80 GB, FP16)		
	Mem (GB)	FLOPs (10^{12})	Time	Mem (GB)	FLOPs (10^{12})	Time	Mem (GB)	FLOPs (10^{12})	Time
Zero-shot ESG	5.2	1.36	5.9 s	13.4	1.36	0.94 s	13.2	1.36	0.051 s
Few-shot ESG (8-shot)	5.3	2.41	8.4 s	13.5	2.41	1.63 s	13.3	2.41	0.078 s
Full ESG Training (5 epochs)	7.1	4.6×10^4	94.2 h	21.0	4.6×10^4	41.7 h	45.3	4.6×10^4	15.3 h

7 Ablation Study

To understand the contribution of each component in ESG, we conduct a systematic ablation across four key mechanisms: (1) contrastive grounding ($\mathcal{L}_{\text{con}}$), which aligns text with consequence-level affordance maps; (2) structural consistency regularization ($\mathcal{L}_{\text{cons}}$); (3) affordance-map supervision, which supplies post-event affordance states; and (4) in-batch negative sampling, which encourages separation of physically inconsistent event outcomes. Each ablation removes exactly one component while keeping all others fixed. We report two representative metrics-ARA and GLCA-on the synthetic disaster scenario evaluation set. Table 6 shows that each component contributes meaningfully, with contrastive grounding providing the largest improvement (+7.9% ARA, +6.1% GLCA over the model without $\mathcal{L}_{\text{con}}$). Removing structural consistency reduces global coherence significantly, lowering GLCA from 47.5% to 40.8%. Eliminating affordance-map supervision greatly harms fine-grained relation prediction, indicating that ESG benefits from supervision of post-event functionality. Finally, removing negative sampling reduces ESG's ability to discriminate subtle consequence differences, particularly for events that share surface-level descriptions but diverge in functional outcomes.

Table 6: Ablation and diagnostic analysis of ESG conducted on the synthetic Disaster benchmark using ARA (↑) and GLCA (↑). Additional variants evaluate supervision quality, graph prediction architecture, negative sampling robustness, and reasoning strategies. Best scores in each column are shown in green.

Model Variant	ARA	GLCA
Full ESG (Ours)	69.4	47.5
Core ESG Components		
– Contrastive Grounding ($\mathcal{L}_{\text{con}}$ removed)	61.5	41.4
– Structural Consistency ($\mathcal{L}_{\text{cons}}$ removed)	63.2	40.8
– Affordance-Map Supervision Removed	58.9	38.2
– No Negative Sampling	65.1	42.7
Architecture Variants		
– Generative Decoder-based Graph Prediction	64.8	41.9
– Encoder-only Linear Triple Classifier	67.3	45.1
– Open-Vocabulary Relation Generation	62.7	39.6
Supervision Quality		
– Noisy Affordance Maps (10% corrupted transitions)	64.2	42.0
– Noisy Affordance Maps (20% corrupted transitions)	60.8	38.7
– Clean Human-Verified Graphs Only	68.7	46.9
Negative Sampling Analysis		
– Random In-Batch Negatives	65.1	42.7
– Hard Negatives (similar events only)	66.0	43.5
– Diversity-Constrained Negatives	68.2	46.1
Reasoning and Baseline Variants		
– GPT-4.1 + Optimized CoT Prompting	57.8	34.5
– Claude 3 Sonnet + Optimized CoT Prompting	55.9	32.8
– Retrieval-Augmented FLAN-T5-XL	63.7	40.9
– Retrieval + Chain-of-Thought	65.0	42.2

Beyond the core component analysis, Table 6 additionally evaluates architectural variants, supervision quality, negative sampling strategies, and stronger reasoning baselines. The results show that encoder-grounded structured triple classification consistently outperforms decoder-based generative graph prediction and open-vocabulary relation generation, supporting ESG's use of a constrained relation-state prediction framework for stable consequence reasoning. Moreover, introducing noisy affordance supervision causes progressive degradation in both ARA and GLCA, confirming that graph quality directly influences consequence-level coherence. However, the relatively moderate drop under 10% corruption also suggests partial robustness to annotation imperfections. The negative-sampling experiments further suggest that diversity-constrained negatives outperform purely random in-batch sampling, reducing the risk of leakage from semantically similar synthetic scenarios while improving representation separation between distinct environmental outcomes.

To complement the tabular results, Fig. 3 presents a six-panel heatmap analysis that visualizes the ablation findings across three complementary perspectives. Panels (a) and (b) encode the absolute ARA and GLCA scores attained when each component is individually removed, where darker shading indicates weaker performance and confirms that affordance-map supervision produces the most severe degradation across both metrics. Panel (c) directly contrasts Full ESG against the best-performing ablated variant, making that no single-component removal can approximate the complete model, and that the performance gap persists across both evaluation dimensions. Panels (d) and (e) isolate the point-wise performance drop incurred by each removal, providing a cleaner signal of marginal contribution independent of baseline offsets. The deeper shading for $\mathcal{L}_{\text{con}}$ in panel (d) visually corroborates its role as the dominant driver of ARA, while the corresponding intensity for $\mathcal{L}_{\text{cons}}$ in panel (e) reflects its outsized influence on global coherence. Panel (f) decomposes the cumulative performance gain across all four components as proportional contribution shares, revealing that contrastive grounding and affordance-map supervision jointly account for the majority of ARA improvement, whereas structural consistency contributes the highest individual share to GLCA recovery. The extended diagnostic analysis additionally shows that stronger prompting strategies, retrieval augmentation, and chain-of-thought reasoning improve baseline LLM performance but remain substantially below ESG on both affordance-level metrics. These results suggest that one limitation of pure LLM reasoning may be not only prompt sensitivity, but also the absence of structured representations for modeling causal environmental transitions. Furthermore, the remaining ESG failure cases primarily involve long-range cascading consequences, ambiguous structural dependencies, and rare multi-event interactions, indicating that consequence-level reasoning remains challenging even with structured grounding supervision. These findings are consistent with interpreting ESG as a consequence-grounded structured reasoning framework rather than a purely distributional language modeling approach.

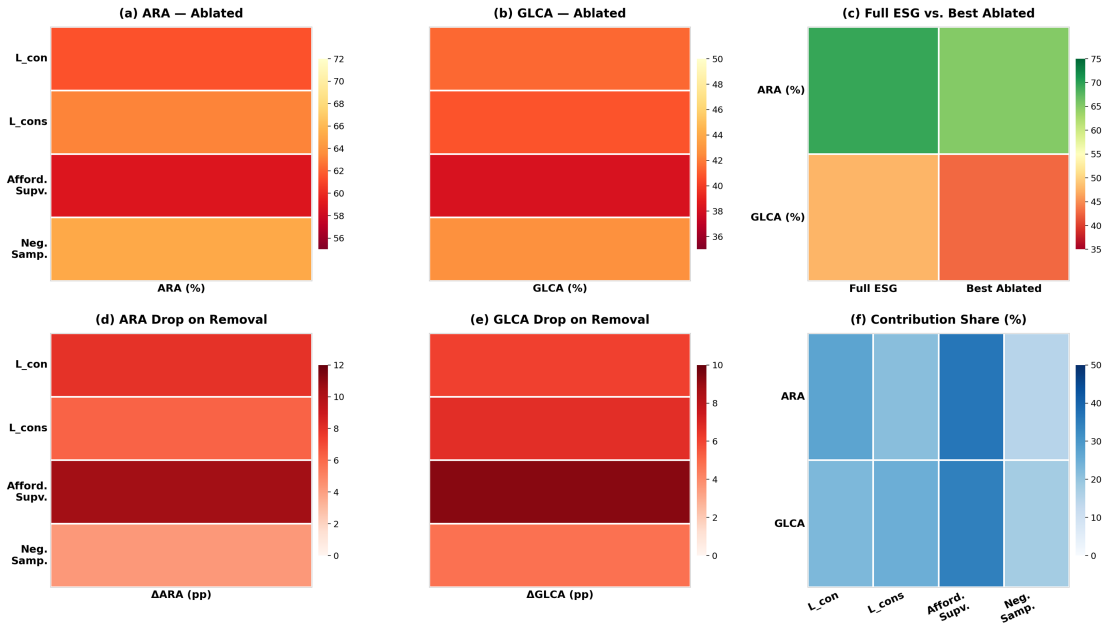

Figure 3: Heatmap analysis of ESG ablation study across four key components. Each panel visualizes a distinct perspective of component contribution: absolute ARA and GLCA scores under individual removal (**a**,**b**), full model vs. best ablated variant (**c**), per-component performance drop in points (**d**,**e**), and proportional contribution share to total gain (**f**). Darker shading consistently indicates greater performance impact, confirming that all four mechanisms-contrastive grounding ($\mathcal{L}_{\text{con}}$), structural consistency ($\mathcal{L}_{\text{cons}}$), affordance-map supervision, and in-batch negative sampling-are individually necessary for robust consequence-level affordance prediction.

7.1 Hyperparameter Sensitivity Analysis

We conduct a comprehensive hyperparameter sensitivity study to understand how key components of ESG influence its ability to learn consequence-grounded representations. Figs. 4–6 summarize the effects of three critical hyperparameter groups: the contrastive temperature τ in the grounding loss, the structural-consistency weight λ in the affordance-regularization term, and core training parameters such as learning rate and batch size. Fig. 4 examines the impact of the contrastive temperature τ in $\mathcal{L}_{\text{con}}$. We observe a clear unimodal trend: very small temperatures (e.g., $\tau = 0.03$) produce overly sharp similarity distributions and hinder generalizability, while larger values (e.g., $\tau = 0.10$) weaken contrastive discrimination between correct and incorrect affordance maps. The optimal performance occurs at $\tau = 0.07$, which is associated with the highest observed ARA and GLCA values, indicating that moderate softness in the contrastive comparisons provides the strongest observed alignment between textual descriptions and consequence-level affordances.

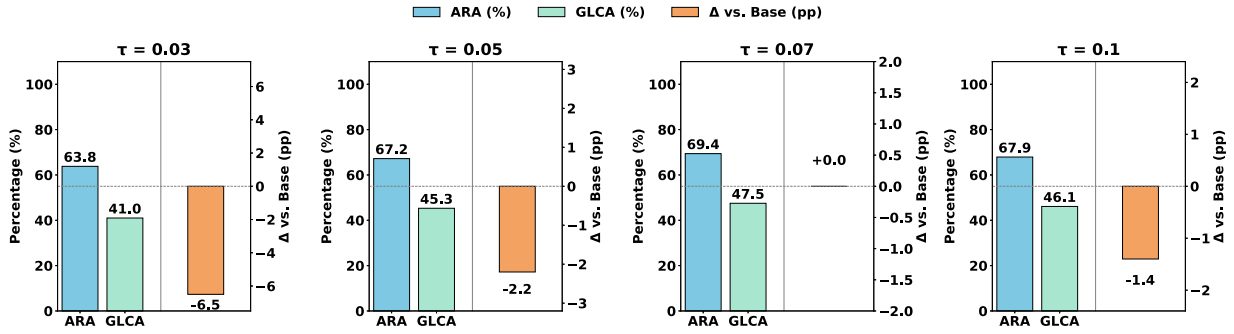

Figure 4: Sensitivity of ESG to the contrastive temperature τ used in the grounding loss $\mathcal{L}_{\text{con}}$. Each panel reports ARA (↑) and GLCA (↑) as percentages, along with the change in GLCA (Delta vs. the default $\tau = 0.07$). ESG achieves peak performance at $\tau = 0.07$, while both lower and higher temperatures reduce contrastive separation, leading to degraded consequence-map prediction accuracy.

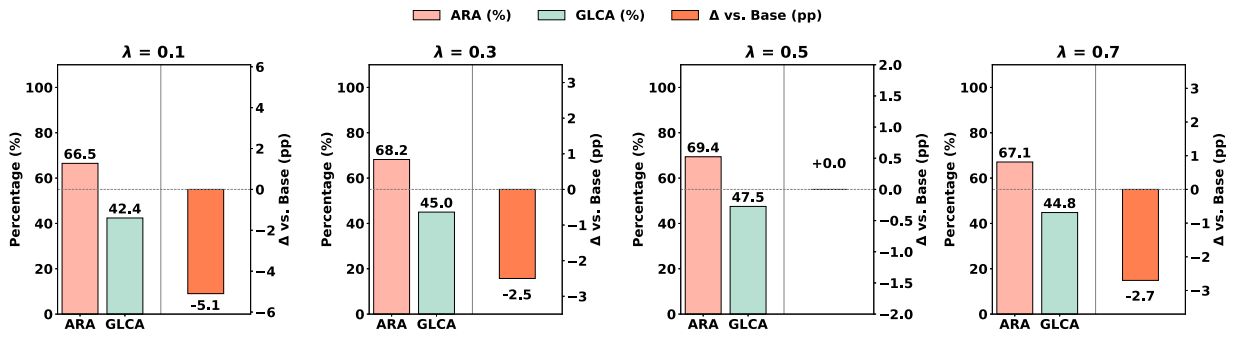

Figure 5: Sensitivity of ESG to the structural-consistency weight λ in the regularizer $\mathcal{L}_{\text{cons}}$. Each panel reports ARA (↑), GLCA (↑), and the change in GLCA relative to the default $\lambda = 0.5$. Moderate regularization ($\lambda = 0.5$) yields the strongest affordance-map consistency, whereas both weaker and stronger penalties degrade global consequence coherence.

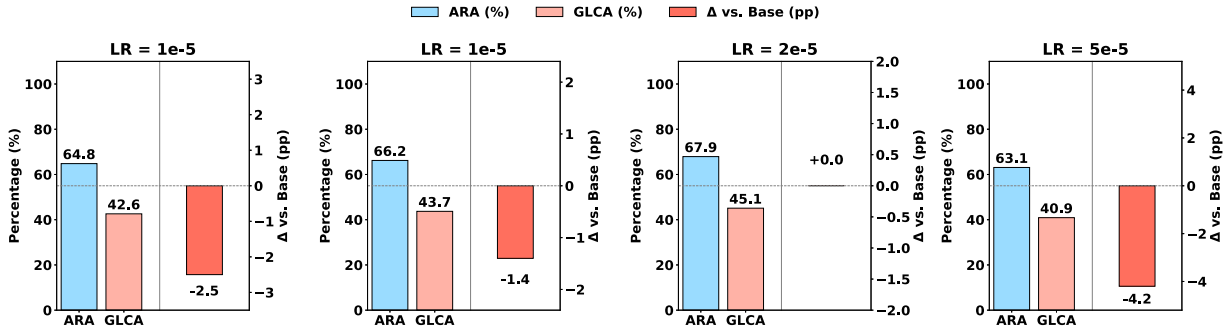

Figure 6: Sensitivity of ESG to core training hyperparameters. Each panel shows the effect of learning rate and batch size on ARA (↑) and GLCA (↑). The default configuration (2×10^{-5}, batch size 32) yields the highest overall accuracy, while both overly small and excessively large learning rates lead to reduced stability in consequence-map prediction.

Fig. 5 analyzes the structural-consistency weight λ that penalizes invalid or physically implausible affordance transitions, while Fig. 6 evaluates sensitivity to core training hyperparameters. Performance improves as λ increases from 0.1 to 0.5, suggesting that mild regularization is insufficient for coherent environmental state transitions. However, overly strong regularization ($\lambda = 0.7$) begins to suppress legitimate variations in affordance structures, leading to a moderate drop in both ARA and GLCA. The highest scores in our experiments are observed at $\lambda = 0.5$, validating our default setting. Similarly, varying the learning rate and batch size shows that the default configuration (2×10^{-5}, batch size 32) is associated with the highest consequence-level reasoning performance among the evaluated configurations. Lower learning rates slow optimization and cause underfitting, while higher learning rates introduce instability and degrade global affordance coherence. Batch size interacts with the learning rate in predictable ways, with larger batches stabilizing gradients and improving GLCA at moderate learning rates.

7.2 Comparison with Stronger Supervised Baselines

Table 7 shows that ESG attains the highest performance among the evaluated supervised baselines across both task-level and affordance-level evaluations, even when competing models are trained under identical data splits, optimization settings, and supervision conditions. While standard supervised fine-tuning and contrastive alignment improve performance relative to weaker baselines, they remain below ESG on PIQA, NEWTON, ARA, and GLCA. In particular, the strongest supervised baseline-Llama-3-8B-Instruct with supervised fine-tuning-reaches 84.0% on PIQA and 77.8% on NEWTON, whereas ESG obtains 86.1% and 81.0%, respectively. Similar trends are observed on affordance-level metrics, where ESG improves ARA from 65.5% to 69.4% and increases global consequence coherence from 42.6% to 47.5% GLCA. These results are consistent with the view that supervised adaptation alone may not fully capture post-event functional transitions. Instead, ESG combines consequence-grounded affordance supervision, structured transition constraints, and contrastive alignment between event descriptions and consequence-level affordance maps. Furthermore, the improvements observed across both task-level and graph-level evaluations suggest that ESG captures information that transfers across the evaluated consequence-reasoning tasks, although further evaluation on additional real-world domains would be needed to fully assess generalization beyond the current benchmarks.

Table 7: Comparison with stronger supervised baselines under identical training conditions. All models are trained on the same datasets and evaluated on the same held-out splits. Metrics are Accuracy (↑) for PIQA/NEWTON and ARA (↑)/GLCA (↑) for affordance-level evaluation. Best scores in each column are shown in green.

Model	PIQA	NEWTON	ARA	GLCA
FLAN-T5-XL (Full Fine-tuning)	82.4	75.8	62.1	39.4
FLAN-T5-XL + Contrastive Alignment	83.1	77.0	64.3	41.2
Llama-3-8B-Instruct (Supervised Fine-tuning)	84.0	77.8	65.5	42.6
ESG (Ours)	86.1	81.0	69.4	47.5

7.3 Statistical Significance Analysis

To assess the reliability and robustness of ESG's performance gains, we additionally report mean performance, standard deviation, 95% Confidence Intervals (CI), paired significance testing, and error-category analysis across three independent runs with different random seeds. For each metric, confidence intervals are computed using the standard normal approximation, and statistical significance is evaluated against the strongest supervised baseline using paired two-tailed t-tests. Table 8 shows that ESG shows lower variance and higher mean performance than the strongest supervised baseline across the reported evaluation metrics. Specifically, ESG maintains standard deviations between 0.4 and 0.6 across PIQA, NEWTON, ARA, and GLCA, indicating stable optimization behavior despite changes in random initialization and mini-batch ordering. All reported p-values are below 0.01, indicating that the observed differences between ESG and the strongest supervised baseline are statistically significant under the adopted testing procedure. The largest absolute improvements are observed on affordance-level metrics, where ESG achieves higher ARA and GLCA scores while maintaining narrow confidence intervals. These results are consistent with the view that consequence-grounded representations contribute to improved performance on the evaluated post-event reasoning tasks relative to the considered baselines. Moreover, the narrow confidence intervals suggest that ESG shows relatively stable behavior across repeated runs and random initializations within the evaluated experimental setting.

Table 8: Statistical significance and error analysis across three independent runs. Results are reported as mean ± standard deviation. CI denotes the 95% confidence interval. p-values are computed against the strongest supervised baseline (Llama-3-8B-Instruct supervised fine-tuning). Best scores are shown in green.

Metric/Failure Type	ESG (Mean ± Std)	95% CI	Baseline	p-Value	Common Failure Pattern
PIQA	86.1 ± 0.4	±0.45	84.0 ± 0.5	0.008	Incorrect physical affordance selection
NEWTON	81.0 ± 0.5	±0.57	77.8 ± 0.6	0.004	Failure to infer material state changes
ARA	69.4 ± 0.6	±0.68	65.5 ± 0.7	0.003	Missing local affordance transitions
GLCA	47.5 ± 0.5	±0.56	42.6 ± 0.8	0.002	Cascading consequence inconsistency
Long-range cascading failures	38.9 ± 0.7	±0.79	31.4 ± 1.0	0.006	Incomplete downstream propagation
Multi-event interaction reasoning	41.7 ± 0.6	±0.68	34.8 ± 0.9	0.005	Conflicting affordance transitions
Ambiguous structural dependencies	44.2 ± 0.5	±0.57	37.1 ± 0.8	0.004	Incorrect support/traversal inference
Rare environmental configurations	40.5 ± 0.8	±0.91	33.0 ± 1.1	0.007	Weak generalization to unseen layouts

Beyond aggregate performance, Table 8 also provides a detailed error analysis of the remaining failure categories. Although ESG substantially reduces physically inconsistent reasoning errors relative to the strongest supervised baseline, the model still struggles with scenarios involving long-range cascading failures, ambiguous structural dependencies, and rare environmental configurations. In particular, the lowest

performance is observed on long-range cascading consequences, where downstream environmental effects propagate across multiple entities and traversal paths. Similarly, multi-event interaction scenarios remain challenging because conflicting affordance transitions can produce globally inconsistent environmental states. These findings suggest that while ESG improves structured consequence reasoning considerably, certain forms of temporally extended causal propagation and complex environmental interaction remain difficult for current text-grounded reasoning architectures.

7.4 Affordance Map Construction and Annotation Protocol

We provide additional details regarding graph construction, annotation consistency, and quality control. ESG converts textual descriptions from PIQA, NEWTON, LIBERO-text, and synthetic disaster scenarios into structured consequence-level affordance graphs through a semi-automatic pipeline consisting of entity extraction, relation assignment, transition inference, and consistency validation. The overall construction workflow is summarized in Table 9. For PIQA and NEWTON, affordance transitions are derived from implicit cues describing usability, material failure, support behavior, or traversal constraints. For example, descriptions involving brittle or damaged objects are converted into transitions such as SUPPORTS→UNSTABLE, while obstruction-related actions are mapped into TRAVERSABLE→BLOCKED. LIBERO-derived text contributes action-centric affordance priors, which are extended into post-event functional transitions using the same rule-guided conversion framework. Synthetic disaster scenarios are annotated separately using manually designed event-consequence templates describing bridge collapses, landslides, debris obstruction, flooding, and infrastructure failure, providing a controlled evaluation setting for consequence-level reasoning under diverse environmental disruptions.

Table 9: Overview of the consequence-level affordance map construction pipeline used in ESG.

Stage	Method	Description
Entity Extraction	Dependency-based parsing + noun phrase extraction	Environment entities, objects, supports, regions, and traversal paths are extracted from textual descriptions using dependency parsing and rule-based noun phrase identification.
Relation Assignment	Rule-guided affordance mapping	Extracted entities are assigned candidate affordance relations such as SUPPORTS, BLOCKS, TRAVERSABLE, and USABLE based on event semantics and contextual object interactions.
Transition Inference	Event-to-state conversion heuristics	Structural events (e.g., collapse, blockage, deformation, obstruction) are mapped into post-event affordance transitions such as TRAVERSABLE→BLOCKED and STABLE→UNSTABLE.
Consistency Filtering	Constraint-based validation	Generated graphs are filtered using a predefined library of valid affordance transitions to remove logically inconsistent or physically implausible state updates.

(Continued)

Table 9 (continued)

Stage	Method	Description
Manual Verification	Human inspection on sampled instances	Randomly sampled graphs from each dataset are manually reviewed to verify alignment between the textual event description and the inferred consequence-level affordance map.

To reduce annotation noise and improve reproducibility, all synthetic disaster scenarios were independently reviewed by three annotators with prior experience in structured reasoning and embodied-environment representations (see Table 10). Annotators were provided with fixed affordance categories and transition definitions during labeling. Inter-annotator agreement measured using Cohen's κ achieved 0.84 for entity identification, 0.81 for affordance relation assignment, and 0.79 for post-event transition labeling, indicating strong agreement across annotation stages. Disagreements were resolved through majority voting followed by final consistency verification. To further avoid leakage or dataset memorization effects, scenario templates are split such that no template structure appears in both training and testing sets. Additionally, affordance-map construction is performed independently from evaluation labels for downstream prediction targets are not directly exposed during graph generation. These safeguards reduce the possibility that ESG benefits from annotation leakage or template overlap, and improve the reproducibility and robustness of the proposed consequence-grounding framework.

Table 10: Inter-annotator agreement for synthetic disaster scenario annotation.

Annotation Task	Cohen's κ
Entity Identification	0.84
Affordance Relation Assignment	0.81
Post-Event Transition Labeling	0.79

7.5 Quantifying Physical Inconsistency in Standard LLM Reasoning

To complement the illustrative bridge-collapse example shown in Fig. 1, we additionally perform a targeted evaluation measuring how frequently standard LLMs produce physically inconsistent consequence predictions in event-based reasoning scenarios. Specifically, we construct a diagnostic benchmark of 300 held-out event descriptions covering bridge collapse, flooding, blockage, landslides, infrastructure failure, and structural deformation. For each scenario, models are required to infer post-event environmental functionality, including whether entities remain usable, traversable, stable, or blocked. Predictions are manually evaluated against reference consequence-level affordance maps to determine whether generated reasoning is physically consistent with the described event. Table 11 shows that standard LLMs often generate linguistically plausible but functionally inconsistent interpretations of environmental changes. Common failure patterns include incorrectly marking collapsed structures as traversable, treating obstructed regions as accessible, or failing to propagate downstream functional consequences after structural failure. Although supervised fine-tuning reduces some of these inconsistencies, baseline models still show error rates ranging from 29.6% to 41.2%. In contrast, ESG reduces physically inconsistent predictions to 16.7%, indicating that consequence-grounded affordance alignment is associated with improved consistency between predicted consequences and reference post-event affordance states within the evaluated benchmark.

Table 11: Frequency of physically inconsistent predictions on the diagnostic consequence-reasoning benchmark. Lower values are better (↓). Best scores are shown in green.

Model	Physically Inconsistent Predictions (%)
GPT-4.1	31.8
Claude 3 Sonnet	34.5
Llama-3-8B-Instruct	41.2
FLAN-T5-XL (Full Fine-tuning)	29.6
ESG (Ours)	16.7

7.6 Prompt Sensitivity and Synthetic Scenario Difficulty Analysis

To further examine evaluation fairness, we additionally evaluate whether prompt optimization narrows the gap between ESG and frontier zero-shot LLMs. Specifically, GPT-4.1 and Claude 3 Sonnet are re-evaluated using manually optimized prompts containing consequence-reasoning instructions, chain-of-thought guidance, and structured output formatting constraints. We also analyze the relative difficulty of the synthetic disaster scenarios compared with public datasets such as PIQA and NEWTON by measuring graph complexity, average number of entities, affordance transitions, and multi-step consequence dependencies. Table 12 shows that prompt optimization moderately improves the performance of frontier zero-shot LLMs, particularly on consequence-level reasoning metrics. However, even with optimized prompting, GPT-4.1 and Claude 3 Sonnet remain below ESG on the reported ARA and GLCA metrics. These results indicate that differences between ESG and the evaluated frontier models are not fully eliminated through prompt optimization alone. Within the evaluated setting, this observation is consistent with the view that consequence-grounded supervision provides useful information for modeling structured post-event environmental transitions beyond what can be obtained through prompting strategies alone.

Table 12: Effect of prompt optimization on frontier LLMs and comparison of dataset difficulty characteristics. Metrics are Accuracy (↑), ARA (↑), and GLCA (↑). Best scores in each column are shown in green.

Model/Dataset	PIQA	NEWTON	ARA	GLCA	Avg. Entities	Avg. Transitions
GPT-4.1 (Standard Prompting)	77.5	68.2	54.1	31.4	–	–
GPT-4.1 (Optimized Prompting)	79.2	70.5	57.0	34.2	–	–
Claude 3 Sonnet (Standard Prompting)	75.9	66.8	52.7	29.6	–	–
Claude 3 Sonnet (Optimized Prompting)	77.4	68.9	55.1	32.0	–	–
ESG (Ours)	86.1	81.0	69.4	47.5	–	–
PIQA	–	–	–	–	2.1	1.3
NEWTON	–	–	–	–	2.8	1.7
Synthetic Disaster Scenarios	–	–	–	–	6.4	5.9

The lower portion of Table 12 further demonstrates that the synthetic disaster scenarios are substantially more structurally complex than PIQA or NEWTON. While PIQA and NEWTON typically involve fewer than three entities and one to two affordance transitions per instance, synthetic disaster scenarios contain on average 6.4 entities and 5.9 interdependent affordance transitions. These scenarios additionally require multi-step reasoning about cascading environmental effects such as obstruction propagation, structural instability, and downstream traversal failure. Therefore, the synthetic benchmark evaluates broader consequence-level reasoning capabilities beyond the localized action plausibility or attribute inference tested by existing public

datasets. This analysis provides additional transparency regarding the difficulty and role of the synthetic evaluation setting in ESG.

7.7 Scaling Analysis and Cross-Domain Transfer Limitations

To better understand the scalability of ESG and the remaining limitations of cross-domain transfer, we additionally evaluate ESG across multiple backbone sizes and analyze transfer degradation between structurally dissimilar domains. In particular, we investigate whether increasing model scale continues to provide consistent gains when combined with consequence-grounded affordance supervision, or whether performance begins to saturate for larger LLMs. Table 13 shows that ESG benefits consistently from increased backbone capacity, with performance improving from FLAN-T5-Base to FLAN-T5-XL across all evaluation metrics. The inclusion of parameter counts further highlights a consistent relationship between model capacity and consequence-level reasoning performance, with gains observed as the backbone scales from 250M to 3B parameters. However, the magnitude of improvement gradually decreases as model scale increases, suggesting partial diminishing returns at larger parameter sizes. For example, the improvement from FLAN-T5-Base to FLAN-T5-Large is larger than the improvement from FLAN-T5-Large to FLAN-T5-XL, particularly on PIQA and NEWTON. These findings indicate that consequence-grounded affordance supervision remains beneficial even for stronger encoder-decoder backbones, although scaling alone does not fully resolve consequence-level reasoning challenges.

Table 13: Backbone size analysis of ESG across different FLAN-T5 variants and representative cross-domain transfer evaluation using ARA (↑)/GLCA (↑). Best scores in each column are shown in green.

Model Backbone	# Params	PIQA	NEWTON	ARA	GLCA
FLAN-T5-Base + ESG	250M	79.5	73.2	61.0	38.8
FLAN-T5-Large + ESG	780M	82.0	76.4	65.3	43.0
FLAN-T5-XL + ESG	3B	86.1	81.0	69.4	47.5
Representative Cross-Domain Transfer (Train on Disaster)					
Transfer Setting			**ARA/GLCA**		
Disaster→PIQA			58.3/24.5		
Disaster→NEWTON			54.0/19.6		
Disaster→LIBERO			62.7/25.0		
Disaster→Disaster			74.9/36.7		

The lower portion of Table 13 reports representative cross-domain transfer results obtained by training ESG on Disaster scenarios and evaluating on PIQA, NEWTON, LIBERO, and Disaster benchmarks. Although ESG substantially outperforms competing baselines across all train-test configurations, transfer from the Disaster domain to PIQA and NEWTON still produces noticeable performance degradation relative to in-domain evaluation. This drop primarily arises because synthetic disaster scenarios contain denser entity interactions, larger consequence graphs, and more cascading environmental transitions than the comparatively localized reasoning patterns present in PIQA or NEWTON. Transfer to LIBERO remains comparatively stronger, suggesting partial overlap between affordance-oriented reasoning patterns across the two domains. Therefore, while ESG learns partially domain-invariant representations of functional environmental change, the results suggest that fully generalized consequence reasoning across structurally different domains remains an open challenge.

8 Conclusion and Future Work

This work introduced ESG, a framework that allows LLMs to infer the *real-world consequences* of events from text alone. Through aligning event descriptions with consequence-level affordance maps, ESG learns a consequence-grounded representation for modeling functional environmental changes. Across the evaluated task-level datasets (PIQA and NEWTON), affordance-level evaluations (LIBERO-derived text and synthetic scenarios), and cross-domain transfer settings, ESG attains higher performance than the considered zero-shot and few-shot baselines. The reported results include improvements on PIQA, NEWTON, and consequence-prediction benchmarks, together with higher ARA and GLCA scores under both in-domain and cross-domain evaluation settings. At the same time, the cross-domain experiments indicate that performance remains sensitive to domain shifts, particularly when transferring across datasets with different linguistic characteristics and affordance structures. ESG remains computationally efficient, supporting real-time inference on GPUs (51 ms on A100, 0.94 s on RTX 4090) and maintaining deployability on edge devices (5.9 s on Jetson Xavier NX).

While ESG provides an initial step toward consequence-oriented reasoning from text, several directions remain open for future exploration. First, extending ESG to multimodal settings-integrating imagery, maps, or 3D spatial data-may further improve grounding in visually complex environments such as disaster zones or robotic manipulation scenes. Second, expanding the consequence library to capture longer causal chains and temporal dynamics could allow reasoning about multi-step or cascading events. Third, ESG could be combined with planning or simulation modules to support decision-making in safety-critical domains such as autonomous navigation, infrastructure monitoring, or emergency response. Finally, scaling consequence-level affordance maps to broader datasets and real-world scenarios would help evaluate ESG's robustness and better characterize its generalization beyond the current benchmarks, including the synthetic consequence-reasoning settings considered in this work.

Acknowledgement: During the preparation of this manuscript, the author utilized ChapGPT-5.5 to refine the academic language. The author has carefully reviewed and revised the output and accepted full responsibility for all content.

Funding Statement: The researcher would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2026).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Manaswi Kulahara and Khadija Parwez; methodology, Manaswi Kulahara and Khadija Parwez; software, Khadija Parwez; validation, Manaswi Kulahara, Khadija Parwez and Faisal Alhwikem; formal analysis, Manaswi Kulahara and Khadija Parwez; investigation, Khadija Parwez; resources, Faisal Alhwikem and Fawwad Hassan Jaskani; data curation, Khadija Parwez; writing—original draft preparation, Khadija Parwez; writing—review and editing, Manaswi Kulahara, Faisal Alhwikem and Fawwad Hassan Jaskani; visualization, Khadija Parwez; supervision, Manaswi Kulahara and Faisal Alhwikem; project administration, Manaswi Kulahara. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are openly available in public repositories. The resources are available at: https://doi.org/10.1609/aaai.v34i05.6239, https://doi.org/10.18653/v1/2023.findings-emnlp.652, and https://proceedings.neurips.cc/paper_files/paper/2023/file/8c3c666820ea055a77726d66fc7d447f-Paper-Datasets_and_Benchmarks.pdf.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bisk Y, Zellers R, Bras RL, Gao J, Piqa CY. Reasoning about physical commonsense in natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York, NY, USA. p. 7432–9.
2. Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, Fedorenko E. Dissociating language and thought in large language models. Trends Cogn Sci. 2024;28(6):517–40. doi:10.1016/j.tics.2024.01.011.
3. Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, et al. Language models as knowledge bases?. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 Nov 3–7; Hong Kong, China. p. 2463–73.
4. AlKhamissi B, Li M, Celikyilmaz A, Diab M, Ghazvininejad M. A review on language models as knowledge bases. arXiv:2204.06031. 2022.
5. Wang Y, Duan J, Fox D, Srinivasa S. NEWTON: are large language models capable of physical reasoning?. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023 Dec 6–10; Singapore. p. 9743–58.
6. Liu B, Zhu Y, Gao C, Feng Y, Liu Q, Zhu Y, et al. LIBERO: benchmarking knowledge transfer for lifelong robot learning. Adv Neural Inf Process Syst. 2023;36:44776–91.
7. Li X, He X, Zhang L, Wu M, Li X, Liu Y. A comprehensive survey on world models for embodied AI. arXiv:2510.16732. 2025.
8. Zhao B, Wang Z, Fang J, Gao C, Man F, Cui J, et al. Embodied-R: collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; Dublin, Ireland. p. 11071–80.
9. Kulahara M, Kashyap GS, Joshi N, Soni A. Can we predict the unpredictable? Leveraging DisasterNet-LLM for multimodal disaster classification. arXiv:2506.23462. 2025.
10. Feng J, Zeng J, Long Q, Chen H, Zhao J, Xi Y, et al. A survey of large language model-powered spatial intelligence across scales: advances in embodied agents, smart cities, and earth science. arXiv:2504.09848. 2025.
11. Mirjalili V, Giahi R, Kollipara S, Kekuda A, Yao K, Zhao K, et al. Spatial reasoning in foundation models: benchmarking object-centric spatial understanding. arXiv:2509.21922. 2025.
12. Grover U, Ranjan R, Mao M, Dong TT, Praveen S, Wu Z, et al. Embodied foundation models at the edge: a survey of deployment constraints and mitigation strategies. arXiv:2603.16952. 2026.
13. Li Z, Guo Y, Liu J, Zhan J, Jiang X, Wang C, et al. Structured causal video reasoning via multi-objective alignment. arXiv:2604.04415. 2026.
14. Team H, Yu X, Liu Z, Wang Z, Zhang H, Rao Y, et al. HY-Embodied-0.5: embodied foundation models for real-world agents. arXiv:2604.07430. 2026.
15. Zellers R, Bisk Y, Schwartz R, Choi Y. SWAG: a large-scale adversarial dataset for grounded commonsense inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31–Nov 4; Brussels, Belgium. p. 93–104.
16. Ha D, Schmidhuber J. World models. arXiv:1803.10122. 2018.
17. Yi K, Gan C, Li Y, Kohli P, Wu J, Torralba A, et al. CLEVRER: collision events for video representation and reasoning. arXiv:1910.01442. 2019.
18. Adak S, Agrawal D, Mukherjee A, Aditya S. Text2Afford: probing object affordance prediction abilities of language models solely from text. In: Proceedings of the 28th Conference on Computational Natural Language Learning; 2024 Nov 15–16; Miami, FL, USA, p. 342–64.
19. Liu X, Yu H, Zhang H, Xu Y, Lei X, Lai H, et al. AgentBench: evaluating LLMs as agents. In: Proceedings of the International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria. p. 52989–3046.
20. Shah D, Osiński B, Ichter B, Levine S. LM-Nav: robotic navigation with large pre-trained models of language, vision, and action. In: Proceedings of the Conference on Robot Learning; 2022 Dec 14–18; Auckland, New Zealand. p. 492–504.
21. Kjellström H, Romero J, Kragić D. Visual object-action recognition: inferring object affordances from human demonstration. Comput Vis Image Underst. 2011;115(1):81–90.

22. Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, et al. Do as I can, not as I say: grounding language in robotic affordances. arXiv:2204.01691. 2022.
23. Fang Z, Huang Z, Wei J, Hua Y. A survey for scene graph generation based on pre-trained model. In: Proceedings of the 2025 5th International Conference on Robotics, Automation, and Artificial Intelligence (RAAI); 2025 Dec 18–20; Singapore. p. 257–66.
24. Hosseini M, Sevtsuk A, Miranda F, Cesar RM Jr, Silva CT. Mapping the walk: a scalable computer vision approach for generating sidewalk network datasets from aerial imagery. Comput Environ Urban Syst. 2023;101:101950.
25. Persiani M, Hellström T. Unsupervised inference of object affordance from text corpora. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics; 2019 Sep 30–Oct 2; Turku, Finland. p. 115–20.
26. Nguyen T, Vu MN, Huang B, Van Vo T, Truong V, Le N, et al. Language-conditioned affordance-pose detection in 3D point clouds. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 3071–8.