LANGUAGE2UTILITY: TRANSFERABLE NATURAL-LANGUAGE PREFERENCE GROUNDING FOR MULTI-OBJECTIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective reinforcement learning (MORL) addresses sequential decision making with competing objectives, but preference-conditioned MORL typically takes numerical preference specifications as given. In practice, however, users express desired trade-offs in natural language, creating a semantic gap between linguistic preferences and the numerical preference specifications underlying multi-objective utility. Motivated by a relational view of preference semantics over objective sets, we study transferable grounding of natural-language preferences across heterogeneous objective spaces and introduce LANGUAGE2UTILITY. LANGUAGE2UTILITY encodes preference-objective relations and contextualizes them with a permutation-equivariant set model over variable-cardinality objective sets. To accommodate practical preference expressions that may not uniquely specify a numerical trade-off, learning combines ordinal constraints with supervision over sets of semantically compatible numerical preferences. We further decouple shared semantic grounding from downstream domain-specific behavioral realization. A semantic-to-behavior bound formalizes the separation by relating language-conditioned utility regret to semantic misalignment and realization error. Across four MO-Gymnasium benchmarks, LANGUAGE2UTILITY improves both semantic grounding and behavioral relevance in-domain. Under leave-one-domain-out transfer, where both the target task and its objective cardinality are unseen during training, LANGUAGE2UTILITY achieves better zero-shot transfer than the compared baselines. Human-authored evaluations across three practical application scenarios further support semantic transfer to new task contexts and objective sets.

1 INTRODUCTION

Many sequential decision-making problems require balancing multiple, often conflicting objectives (Liu et al., 2025c). In autonomous driving, for example, a vehicle may trade off travel time, passenger comfort, and energy consumption. Multi-objective reinforcement learning (MORL) (Hayes et al., 2022) provides a principled framework for such problems. Preference-conditioned MORL enables a single controller to realize diverse trade-offs according to numerical preference specifications (Yang et al., 2019; Abels et al., 2019; Basaklar et al., 2023). In practice, however, users often express desired trade-offs in natural language rather than providing an explicit numerical specification. This creates a semantic gap between natural-language preferences and the numerical preference specifications underlying multi-objective utility. Informed by prior uses of *preference grounding* to connect user intent with downstream decision-making (Huang et al., 2023; Srivastava et al., 2026), we employ this term in MORL to denote interpreting a natural-language preference as a numerical preference specification. Yet in standard preference-conditioned MORL formulations, such specifications are typically taken as given, leaving their grounding from user language unspecified.

A straightforward alternative is to directly prompt a general-purpose large language model (LLM) to infer a numerical preference specification from natural-language input (Hwang et al., 2024; Sethuraman et al., 2026). However, direct preference inference can be sensitive to the language model and prompting procedure, with structured output reliability not always guaranteed. A complementary strategy is to learn a dedicated adapter from language to numerical preference weights (Liu et al., 2025a). Such adaptations are typically tailored for a predefined, task-specific objective space, thereby

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

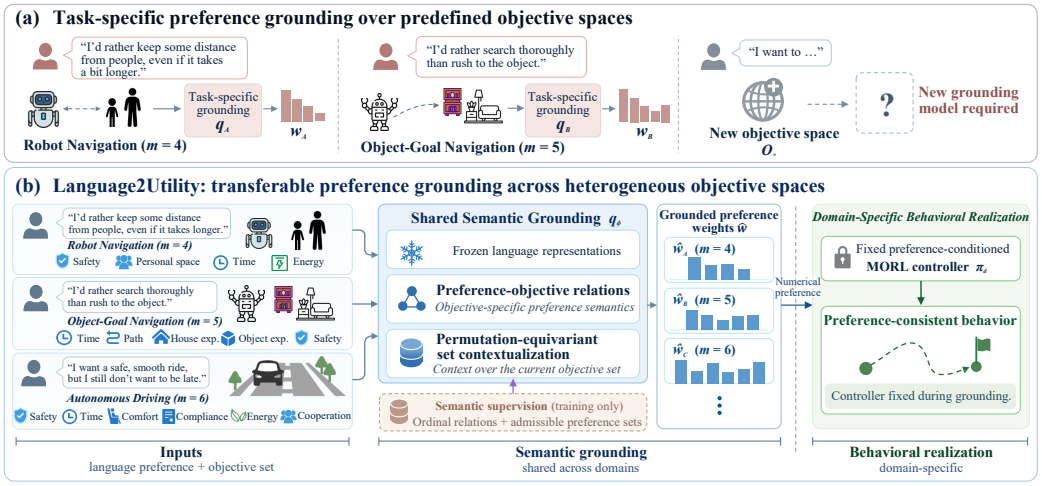


Figure 1: From task-specific to transferable preference grounding for MORL. (a) Task-specific grounding is tied to predefined objective spaces and requires a new grounding function when the objective space changes. (b) LANGUAGE2UTILITY shares semantic grounding across heterogeneous, variable-cardinality objective spaces through preference-objective relational semantics and permutation-equivariant set contextualization. The application scenarios illustrate heterogeneous objective spaces for semantic grounding, whereas behavioral realization is evaluated separately on MO-Gymnasium using fixed domain-specific MORL controllers.

binding the grounding model to fixed objective identifiers and output dimensions. This dependency impedes cross-domain generalization in MORL, as the composition and cardinality of the objective space can vary. Therefore, the key question is: can a single grounding model effectively capture preference semantics that transfer across heterogeneous objective spaces?

In this work, we bridge this gap with LANGUAGE2UTILITY, a transferable natural-language preference grounding framework for MORL. LANGUAGE2UTILITY is motivated by a relational view of preference semantics over objective sets. In particular, objectives may differ across domains, while their preference-relative roles, such as being dominant, secondary, or negligible, remain comparable. Building on this structure, LANGUAGE2UTILITY employs a frozen language model to encode preference-objective relational semantics. A permutation-equivariant set model then contextualizes the resulting semantics over the current objectives, supporting variable-cardinality objective sets. To accommodate practical preference expressions that may not uniquely specify a numerical trade-off, learning combines ordinal preference constraints with supervision over admissible preference sets rather than enforcing a unique preference specification. Figure 1 contrasts task-specific grounding over predefined objective spaces with the shared grounding model of LANGUAGE2UTILITY across heterogeneous objective spaces.

The shared semantic grounding in LANGUAGE2UTILITY is further decoupled from domain-specific behavioral realization. Semantic grounding infers a numerical preference specification from language and objective semantics, whereas behavioral realization maps the specification to domain-specific policy behavior. Accordingly, the grounding model is shared across domains, while domain-specific preference-conditioned MORL controllers are independently trained and held fixed. A *semantic-to-behavior bound* formalizes the decomposition by upper-bounding language-conditioned utility regret with semantic misalignment and realization error.

Our main contributions are threefold: (1) To our knowledge, we are the first to formulate and evaluate transferable natural-language preference grounding for MORL across heterogeneous, variable-cardinality objective spaces. We introduce LANGUAGE2UTILITY, which combines preference-objective relational semantics with permutation-equivariant set contextualization. (2) We decouple semantic grounding from domain-specific behavioral realization and formalize the separation with a semantic-to-behavior bound relating utility regret to semantic misalignment and realization error. (3) Across four MO-Gymnasium tasks (Felten et al., 2023), LANGUAGE2UTILITY achieves consistent gains in semantic grounding and behavioral relevance, both in-domain and under leave-one-domain-

out (LODO) transfer with simultaneous domain and cardinality shift. Ablations show complementary contributions from preference-objective representations and set-level contextualization, while human-authored evaluations across three practical application scenarios further support semantic transfer to new task contexts and objective sets.

2 Related Work

LLMs for Preference Specification. LLMs have been used to translate task descriptions or feedback into parameterized or executable reward functions (Xie et al., 2024; Ma et al., 2024). Complementary work studies LM-assisted preference elicitation through uncertainty-aware queries or diagnostic cases (Handa et al., 2024; Li et al., 2025). More closely related to our setting, language instructions or context-dependent preference rules have been translated into numerical weights for preference-conditioned MORL policies (Hwang et al., 2024; Sethuraman et al., 2026). Prompt-aware prediction of weights over predefined LLM alignment objectives has also been studied (Liu et al., 2025a). These numerical weight-inference approaches are defined over predefined objective schemas, leaving cross-domain grounding under changing objective identities and cardinalities unaddressed.

Preference-Conditioned MORL. Multi-objective reinforcement learning primarily studies how specified trade-offs are realized through policy behavior. Multi-policy approaches construct policy coverage sets or approximate Pareto trade-offs (Van Moffaert & Nowé, 2014; Liu et al., 2025b;c; Alegre et al., 2026). Preference-conditioned methods instead amortize control across trade-offs by conditioning a single policy on numerical preference weights (Abels et al., 2019; Yang et al., 2019; Basaklar et al., 2023; Yang et al., 2025b; Ambadkar et al., 2026; Shianifar et al., 2026). Related work also considers variations in numerical preferences (Yuan et al., 2025) and environment configurations (Teoh et al., 2025), as well as objective expansion within a fixed task family (Chen et al., 2026). These settings take numerical preference specifications as given and primarily concern control-side adaptation. In contrast, we study transfer of semantic grounding across heterogeneous objective sets, while behavioral realization remains domain specific.

3 Language2Utility: Transferable Preference Grounding over Variable Objective Sets

Given a natural-language preference and a variable-cardinality objective set, Language2Utility infers simplex-valued numerical preferences over the supplied objectives. The grounding model is shared across heterogeneous objective spaces, while behavioral realization remains domain specific.

3.1 Problem Formulation

Variable-cardinality objective sets. A domain d is associated with an objective set $\mathcal{O}_d = \{o_{d,i}\}_{i=1}^{m_d}$, where the cardinality m_d may vary across domains. Given a natural-language preference x, we define the shared grounding function

$$q_\phi : (x, \mathcal{O}_d) \longmapsto \widehat{\mathbf{w}}_x \in \Delta^{m_d-1}, \qquad \widehat{\mathbf{w}}_{x,i} \geq 0, \quad \mathbf{1}^\mathsf{T}\widehat{\mathbf{w}}_x = 1. \tag{1}$$

Since the ordering of objectives is arbitrary, q_ϕ is permutation equivariant: $q_\phi(x, \Pi\mathcal{O}_d) = \Pi q_\phi(x, \mathcal{O}_d)$ for any permutation matrix Π. The parameters ϕ are shared across domains and cardinalities, with output dimensionality determined by $|\mathcal{O}_d|$.

Numerical preference specification. Preference-conditioned MORL represents objective trade-offs by a simplex-valued preference vector $\mathbf{w} \in \Delta^{m_d-1}$. For an expected discounted vector return $\mathbf{R} \in \mathbb{R}^{m_d}$, we define the corresponding linear *utility* as $U_\mathbf{w}(\mathbf{R}) = \mathbf{w}^\mathsf{T}\mathbf{R}$. Accordingly, q_ϕ infers $\widehat{\mathbf{w}}_x = q_\phi(x, \mathcal{O}_d)$, the numerical preference specification that parameterizes this utility.

Set-valued preference semantics. Natural language may underdetermine the numerical trade-off, leaving a single preference expression compatible with multiple points in the simplex. Let $c = c(x)$ denote the semantic specification associated with x. We represent the numerical preferences compatible with c by a finite, nonempty admissible set

$$\mathcal{W}_c = \left\{\mathbf{w}_c^{(1)}, \ldots, \mathbf{w}_c^{(K_c)}\right\} \subset \Delta^{m_d-1}. \tag{2}$$

We further denote by $\mathcal{C}_c$ the ordinal relations entailed by c, including $o_i \succ o_j$ and $o_i \approx o_j$. Grounding therefore seeks $\widehat{\mathbf{w}}_x$ consistent with both the admissible set $\mathcal{W}_c$ and the ordinal structure $\mathcal{C}_c$, rather than a unique numerical target.

3.2 Relational Grounding over Objective Sets

3.2.1 Preference-Objective Relational Representations

From a relational view of preference semantics over objective sets, we characterize each objective by how the expressed preference prioritizes it (e.g., as dominant, secondary, or negligible). For each preference x and objective $o_{d,i}$, a frozen language encoder provides three complementary representations: $\mathbf{e}^P(x)$ captures global preference semantics, $\mathbf{e}_i^O(o_{d,i})$ encodes task-conditioned objective semantics, and $\mathbf{e}_i^J(x, o_{d,i})$ characterizes how the expressed preference prioritizes objective $o_{d,i}$ within the task context.

We combine the three representations by independently normalizing and concatenating them as $\mathbf{r}_i = [\bar{\mathbf{e}}^P; \bar{\mathbf{e}}_i^O; \bar{\mathbf{e}}_i^J]$, and obtain the objective-wise relational representation through the shared transformation

$$\mathbf{h}_i^0 = E_\phi(\mathbf{r}_i). \tag{3}$$

The same E_ϕ is shared across domains and cardinalities, ensuring identical application to each objective. We use the frozen Qwen3-Embedding-8B model (Zhang et al., 2025) to obtain the language representations. Exact input templates, dimensions, and normalization details are provided in Appendix A.4.

3.2.2 Permutation-Equivariant Set Contextualization

The weight assigned to an objective depends on its preference relation relative to the other objectives in the current set. We therefore contextualize the objective-wise relational representations jointly before forming the numerical preference specification. Let $\mathbf{H}^0 = [\mathbf{h}_1^0; \ldots; \mathbf{h}_{m_d}^0]$. We apply permutation-equivariant self-attention (Lee et al., 2019):

$$\begin{aligned} \mathbf{A} &= \mathrm{MHA}(\mathbf{H}^0, \mathbf{H}^0, \mathbf{H}^0), \\ \mathbf{H}^1 &= \mathrm{LN}_1(\mathbf{H}^0 + \mathbf{A}), \\ \mathbf{H}^2 &= \mathrm{LN}_2\big(\mathbf{H}^1 + \mathrm{FFN}(\mathbf{H}^1)\big). \end{aligned} \tag{4}$$

This contextualization allows each preference-objective relation to be interpreted relative to the complete objective set. A shared scalar map produces $u_i = S_\phi(\mathbf{h}_i^2)$, followed by simplex normalization,

$$\widehat{\mathbf{w}}_{x,i} = \frac{\exp(u_i/\tau_w)}{\sum_{j=1}^{m_d} \exp(u_j/\tau_w)}, \qquad \tau_w > 0. \tag{5}$$

Together, shared objective-wise processing and permutation-equivariant contextualization preserve the permutation equivariance of q_ϕ while supporting variable-cardinality objective sets. Formal equivariance and architecture details are provided in Appendix A.5.

3.3 Learning under Semantic Underdetermination

To accommodate semantic underdetermination, we supervise the grounding model with ordinal and set-valued constraints rather than a unique numerical target. For a specification $c = c(x)$, ordinal relations $\mathcal{C}_c$ constrain relative objective importance, while the admissible set $\mathcal{W}_c$ constrains the compatible region of the preference simplex.

Ordinal consistency. For $(i, j, y) \in \mathcal{C}_c$, let $\Delta u_{ij} = u_i - u_j$, where $y = +1, 0, -1$ denotes $o_i \succ o_j$, $o_i \approx o_j$, and $o_j \succ o_i$, respectively. We impose

$$L_{\mathrm{ord}}(x) = \frac{1}{|\mathcal{C}_c|} \sum_{(i,j,y)\in\mathcal{C}_c} \ell_{\mathrm{ord}}(\Delta u_{ij}, y). \tag{6}$$

Since simplex normalization preserves logit ordering, this objective enforces the ordinal semantics of c without committing to a particular point in the simplex.

Admissible-set consistency. To constrain the full preference geometry without privileging a single admissible weight, we minimize the distance from the prediction to the admissible set,

$$L_{\text{set}}(x) = \min_{\mathbf{w} \in \mathcal{W}_c} \|\widehat{\mathbf{w}}_x - \mathbf{w}\|_1. \tag{7}$$

Thus, any numerical preference compatible with the semantic specification remains a valid realization.

Semantic learning objective. We jointly enforce the ordinal and set-valued constraints through

$$L_{\text{sem}} = \lambda_{\text{ord}} L_{\text{ord}} + \lambda_{\text{set}} L_{\text{set}}. \tag{8}$$

The exact ordinal loss and training details are provided in Appendix A.6.

3.4 Decoupling Semantic Grounding from Behavioral Realization

Grounded utility and behavioral realization. Given x and $\mathcal{O}_d$, the grounding model produces $\widehat{\mathbf{w}}_x = q_\phi(x, \mathcal{O}_d)$, which parameterizes the grounded utility $U_{\widehat{\mathbf{w}}_x}(\widetilde{\mathbf{R}}) = \widehat{\mathbf{w}}_x^\top \widetilde{\mathbf{R}}$. For each domain d, let $\mathcal{R}_d \subset \mathbb{R}^{m_d}$ be the bounded set of feasible normalized expected returns, and let $\mathcal{G}_{\theta,d} : \Delta^{m_d-1} \to \mathcal{R}_d$ denote the return realized by the fixed preference-conditioned controller when conditioned on $\mathbf{w}$. The language-induced behavior is therefore $\widetilde{\mathbf{R}}_x = \mathcal{G}_{\theta,d}(q_\phi(x, \mathcal{O}_d))$. The grounding function q_ϕ is shared across domains, whereas $\mathcal{G}_{\theta,d}$ captures domain-specific behavioral realization; its exact form under affine return normalization is given in Appendix A.1.

Semantic-to-behavior decomposition. Let $D_d = \sup_{\mathbf{R},\mathbf{R}' \in \mathcal{R}_d} \|\mathbf{R} - \mathbf{R}'\|_\infty$ denote the ℓ_∞ diameter of $\mathcal{R}_d$. For a semantic specification c, let $\overline{\mathbf{w}}_c = |\mathcal{W}_c|^{-1} \sum_{\mathbf{w} \in \mathcal{W}_c} \mathbf{w}$ be the centroid of its admissible set; by linearity, $\overline{\mathbf{w}}_c^\top \mathbf{R}$ equals the average utility over $\mathcal{W}_c$. For a query $\mathbf{q} \in \Delta^{m_d-1}$, define the realization gap $\varepsilon_{\theta,d}(\mathbf{q}) = \sup_{\mathbf{R} \in \mathcal{R}_d} \mathbf{q}^\top \mathbf{R} - \mathbf{q}^\top \mathcal{G}_{\theta,d}(\mathbf{q})$; for x associated with c, define $\text{Reg}^\star_{\text{lang}}(x) = \sup_{\mathbf{R} \in \mathcal{R}_d} \overline{\mathbf{w}}_c^\top \mathbf{R} - \overline{\mathbf{w}}_c^\top \mathcal{G}_{\theta,d}(\widehat{\mathbf{w}}_x)$.

Proposition 1 (Semantic-to-behavior bound)**.** *For every language preference x in domain d,*

$$\begin{aligned} \text{Reg}^\star_{\text{lang}}(x) &\le D_d \|\widehat{\mathbf{w}}_x - \overline{\mathbf{w}}_c\|_1 + \varepsilon_{\theta,d}(\widehat{\mathbf{w}}_x) \\ &\le D_d[\text{Set-L1}(\widehat{\mathbf{w}}_x, \mathcal{W}_c) + \rho_c] + \varepsilon_{\theta,d}(\widehat{\mathbf{w}}_x), \end{aligned} \tag{9}$$

where $\text{Set-L1}(\widehat{\mathbf{w}}_x, \mathcal{W}_c) = \min_{\mathbf{w} \in \mathcal{W}_c} \|\widehat{\mathbf{w}}_x - \mathbf{w}\|_1$ *and* $\rho_c = \max_{\mathbf{w} \in \mathcal{W}_c} \|\mathbf{w} - \overline{\mathbf{w}}_c\|_1$.

The bound separates semantic grounding error from domain-specific realization error. In its set-valued form, $\text{Set-L1}(\widehat{\mathbf{w}}_x, \mathcal{W}_c)$ measures distance to the admissible preference set, while ρ_c captures its intrinsic dispersion around the centroid. A proof is provided in Appendix A.2.

4 Experiments

We evaluate whether LANGUAGE2UTILITY (i) achieves accurate semantic grounding and behavioral relevance in-domain, (ii) transfers zero-shot to held-out tasks and unseen objective cardinalities, and (iii) generalizes semantically to human-authored preferences in new application scenarios.

4.1 Experimental Setup

Domains and fixed MORL controllers. We evaluate on four MO-Gymnasium domains with distinct objective cardinalities: DST ($m = 2$), Minecart ($m = 3$), Lunar Lander ($m = 4$), and Fruit Tree ($m = 6$) (Felten et al., 2023). To isolate semantic grounding from behavioral realization, each domain uses an independently trained preference-conditioned MORL controller based on GPI-LS or GPI-PD (Alegre et al., 2023; Felten et al., 2023). The controllers are trained without language supervision and remain fixed throughout grounding experiments. Training configurations and empirical return coverage are provided in Appendix B.1.

Controlled semantic benchmark. For each domain, we discretize the preference simplex and group numerical preferences based on their qualitative role characteristics across the objectives. Each group defines an admissible preference set $\mathcal{W}_c$ and the corresponding semantic specification c. A frozen LLM then produces semantically preserving language realizations for each specification, following controlled language-generation protocols (Zhou et al., 2025; Dey et al., 2026). The train-validation-test splits are standardized at the specification level, ensuring that all language realizations retain the same c within a single split. The test set contains 38 semantic specifications and 174 language realizations. Full benchmark construction and split statistics are given in Appendix B.

Evaluation metrics. Semantic grounding is evaluated using Set-L1, Centroid-L1, and Ord. Acc., which measure distance to the admissible preference set, distance to its centroid, and ordinal consistency, respectively. Behavioral relevance is evaluated using Pair Acc., J_{lang}, and Reg_{lang}, which measure preference-consistent return ordering, realized semantic utility, and regret relative to the empirical reference utility. Learned methods report mean±standard deviation over five seeds. Metric definitions and aggregation are provided in Appendix C.1.

Comparison methods. We compare with Frozen Cosine and Regularized Linear references; objective-wise neural predictors based on Pointwise MLPs and ArmoRM-style gating (Wang et al., 2024); and the permutation-invariant DeepSets architecture (Zaheer et al., 2017). Each neural predictor is evaluated with and without the Joint preference-objective representation, with DeepSets w/ Joint providing the closest set-based alternative to LANGUAGE2UTILITY. We additionally evaluate direct LLM predictors that infer numerical preferences from the language preference and objective set without task-specific training (Hwang et al., 2024). For direct LLM baselines, we evaluate Qwen3-8B (Yang et al., 2025a) and Ministral 3 8B (Liu et al., 2026). Broader comparisons with frontier LLMs under the same prompting protocol are provided in Appendix D.1.

4.2 IN-DOMAIN GROUNDING AND BEHAVIORAL RELEVANCE

Overall grounding performance. Table 1 compares learned preference predictors across the four MO-Gymnasium domains under in-domain evaluation. LANGUAGE2UTILITY achieves the best semantic grounding and behavioral relevance, reducing Set-L1 by 10.8% and Reg_{lang} by 11.6% relative to the best competing learned predictor on each metric. Across the three neural architectures, Joint preference-objective representations reduce Set-L1 by 55.4–61.6% and Reg_{lang} by 55.4–72.1%, supporting the value of relational preference semantics. Appendix D.1 extends the comparison to frontier direct LLMs, where LANGUAGE2UTILITY improves preference-geometry recovery and behavioral relevance while reducing median inference latency from seconds to 116.84 ms.

Table 1: In-domain comparison of preference-grounding methods. Results report mean±standard deviation over five seeds. Arrows indicate metric direction; best and second-best values are bolded and underlined, respectively.

Method	Set-L1↓	Centroid-L1↓	Ord. Acc.↑	Pair Acc.↑	J_{lang}↑	Reg_{lang}↓
Frozen Cosine	.5422±.0000	.5770±.0000	.5378±.0000	.7554±.0000	.5787±.0001	.0936±.0001
Regularized Linear	.2716±.0004	.3006±.0003	.8497±.0011	.9544±.0003	.6552±.0003	.0171±.0003
ArmoRM	.4215±.0035	.4486±.0046	.7147±.0129	.8666±.0044	.6289±.0018	.0435±.0018
Pointwise MLP	.2853±.0038	.3262±.0036	.7923±.0113	.9074±.0047	.6401±.0018	.0323±.0018
DeepSets	.3692±.0348	.4092±.0317	.7246±.0181	.8598±.0177	.6210±.0054	.0513±.0054
ArmoRM w/ Joint	.1617±.0021	.1970±.0025	.8811±.0015	.9685±.0017	<u>.6586±.0010</u>	<u>.0138±.0010</u>
Pointwise MLP w/ Joint	<u>.1273±.0027</u>	<u>.1565±.0049</u>	<u>.9176±.0093</u>	<u>.9760±.0014</u>	.6580±.0006	.0144±.0006
DeepSets w/ Joint	.1475±.0094	.1790±.0128	.8857±.0115	.9724±.0024	.6581±.0009	.0143±.0009
Ours	**.1136±.0050**	**.1418±.0049**	**.9211±.0072**	**.9796±.0008**	**.6602±.0008**	**.0122±.0008**

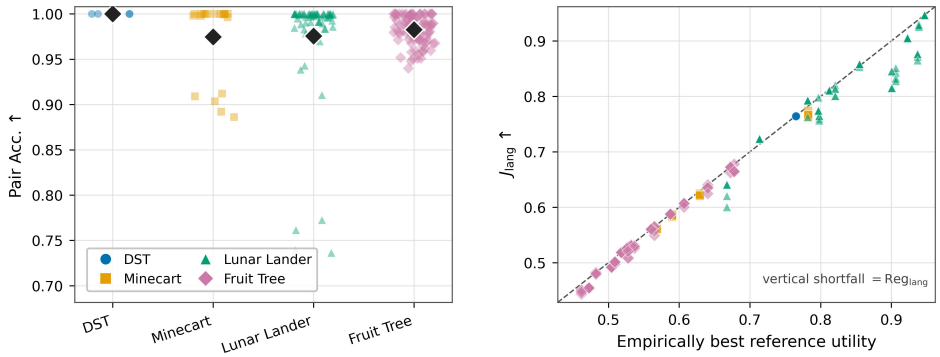
What Enables Accurate Grounding? The ablations in Table 2 examine relational representations (*w/o Joint*, *Joint-only*), set contextualization (*w/o Objective Interaction*), and semantic supervision (*Ordinal-only*, *Semantic-set-only*). The largest degradation occurs when Joint representations are removed: Set-L1 increases from .1136 to .3037 and Pair Acc. decreases from .9796 to .9000. The Joint-only configuration retains competitive performance, further highlighting the informativeness of preference-objective relational representations. Set contextualization provides a complementary

contribution, with removal of objective interaction increasing Set-L1 to .1291 and degrading both semantic and behavioral metrics. The supervision ablations further show that admissible-set constraints anchor preference geometry, while ordinal constraints contribute complementary ordering information. Intervention-based diagnostics of Joint alignment and cross-objective influence are provided in Appendix D.4.

Table 2: In-domain ablation of relational representations, set contextualization, and semantic supervision. Results report mean±standard deviation over five seeds. Arrows indicate metric direction; best and second-best values are bolded and underlined, respectively.

Method	Set-L1↓	Centroid-L1↓	Ord. Acc.↑	Pair Acc.↑	J_{lang}↑	Reg_{lang}↓
Joint-only	.1333±.0078	.1648±.0094	.9094±.0076	<u>.9783±.0027</u>	.6590±.0010	.0133±.0010
w/o Joint	.3037±.0088	.3451±.0073	.7726±.0053	.9000±.0033	.6402±.0034	.0322±.0034
w/o Objective Interaction	.1291±.0054	.1543±.0064	<u>.9122±.0032</u>	.9749±.0020	.6588±.0008	.0136±.0008
Ordinal-only	.5852±.0040	.6395±.0044	.8615±.0110	.9254±.0020	.6374±.0017	.0349±.0017
Semantic-set-only	<u>.1236±.0070</u>	<u>.1516±.0055</u>	.9001±.0074	.9756±.0011	<u>.6593±.0010</u>	<u>.0131±.0010</u>
Ours (full)	**.1136±.0050**	**.1418±.0049**	**.9211±.0072**	**.9796±.0008**	**.6602±.0008**	**.0122±.0008**

Behavioral realization of grounded preferences. Figure 2 examines the behavior induced by the grounded preferences through the fixed domain-specific controllers. Across domains, LANGUAGE2UTILITY preserves the semantic ordering of robust return pairs and realizes utilities close to the empirical reference, yielding high Pair Acc. and low Reg_{lang}. Thus, the semantic grounding gains remain behaviorally relevant after domain-specific realization, providing an empirical counterpart to the decomposition in Section 3.4.



(a) Return-pair consistency (b) Closed-loop utility

Figure 2: Behavioral realization of LANGUAGE2UTILITY grounded preferences on the in-domain test set. (a) Preference consistency over robust return pairs. Black diamonds denote domain means. (b) Realized semantic utility and the best empirical reference utility. Vertical gaps correspond to Reg_{lang}.

4.3 ZERO-SHOT TRANSFER UNDER SIMULTANEOUS DOMAIN AND CARDINALITY SHIFT

4.3.1 LODO TRANSFER PERFORMANCE

In leave-one-domain-out (LODO) evaluation across the four MO-Gymnasium benchmarks, the grounding model is trained on three source domains and evaluated zero-shot on the unseen target domain. Since the benchmarks have distinct objective cardinalities, withholding the target domain also removes its corresponding cardinality from training. The domain-specific controllers remain fixed throughout evaluation.

Table 3 shows that LANGUAGE2UTILITY achieves the best semantic grounding and behavioral relevance under LODO transfer. Compared with DeepSets w/ Joint, the closest set-based alternative, LANGUAGE2UTILITY reduces Set-L1 by 8.6% and Reg_{lang} by 18.3%, while improving Pair Acc.

by 2.8 percentage points. Compared with the stronger 8B-scale direct LLM, Ministral 3 8B, LANGUAGE2UTILITY reduces Set-L1 by 17.9% and $\mathrm{Reg}_{\mathrm{lang}}$ by 14.7%, while improving Pair Acc. by 5.4 percentage points. These results show that learned grounding transfers beyond observed objective spaces. Per-domain results in Appendix E show better or comparable performance for LANGUAGE2UTILITY across held-out domains.

Table 3: LODO zero-shot transfer under simultaneous domain and cardinality shift. Learned methods report mean±standard deviation over five training seeds, while 8B-scale direct LLMs report mean±standard deviation over five independent inference runs. Best and second-best means are bolded and underlined, respectively.

Method	Set-L1↓	Centroid-L1↓	Ord. Acc.↑	Pair Acc.↑	J_{lang}↑	$\mathrm{Reg}_{\mathrm{lang}}$↓
w/o Objective Interaction	.3800±.0300	.4052±.0297	.7741±.0083	.8835±.0122	.6274±.0033	.0450±.0033
w/o Joint	.6555±.0167	.6854±.0167	.3144±.0516	.6808±.0107	.5561±.0080	.1162±.0080
DeepSets	.6618±.0278	.6920±.0265	.2753±.0496	.6825±.0161	.5552±.0163	.1171±.0163
Qwen3-8B	.4328±.0098	.4579±.0118	.6026±.0054	.8521±.0040	.6290±.0009	.0434±.0009
Ministral 3 8B	.3651±.0053	.3774±.0060	.6744±.0133	.8633±.0041	.6322±.0020	.0402±.0020
DeepSets w/ Joint†	.3279±.0111	.3561±.0093	.7727±.0091	.8894±.0048	.6303±.0018	.0420±.0018
Ours†	**.2996±.0125**	**.3284±.0092**	**.8007±.0149**	**.9173±.0079**	**.6381±.0042**	**.0343±.0042**

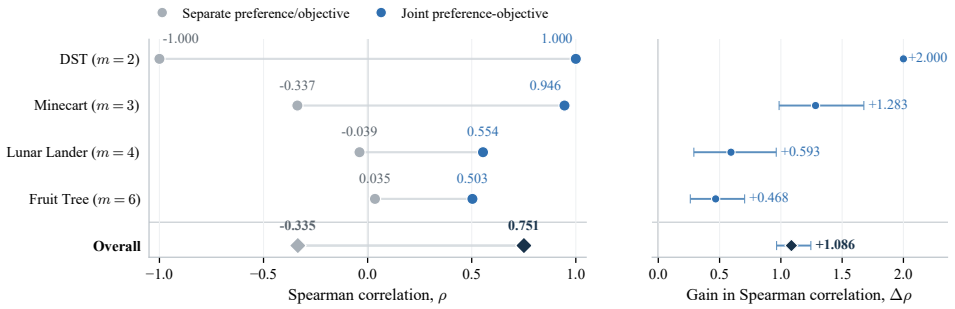
† Methods using both Joint preference-objective representations and set-level objective interaction.

4.3.2 TESTING THE RELATIONAL-TRANSFER HYPOTHESIS

Our central hypothesis is that preference-relative objective relations remain informative despite differing objective sets. We test whether these relations are linearly recoverable across domains using probes trained on frozen source-domain representations and evaluated on the held-out domain. For concatenated preference and objective representations, the probe decomposes into separate terms, whereas Joint representations encode the preference-objective pair jointly before probing.

Figure 3 shows a consistent cross-domain advantage for Joint preference-objective representations. As shown in Fig. 3(a), the overall Spearman correlation increases from $-.335$ with separate preference and objective representations to .751 with Joint representations. The domain-wise comparison exhibits the same pattern in every held-out domain. Figure 3(b) makes this effect explicit through the paired gain $\Delta\rho = \rho_{\mathrm{Joint}} - \rho_{\mathrm{Separate}}$: all four held-out domains yield positive gains, ranging from $+.468$ to $+2.000$, with an overall gain of $+1.086$.

These results provide representation-level evidence that Joint representations make preference-relative objective information more linearly recoverable across domain and cardinality shifts, complementing the end-to-end LODO results in Table 3. Additional probe details and representation-space analyses are provided in Appendix E.3.



(a) Linear-probe correlation on held-out domains

(b) Joint-representation gain

Figure 3: Cross-domain linear recoverability of preference-relative objective information. (a) Spearman correlation of linear probes trained on source domains and evaluated directly on each held-out domain, comparing concatenated preference and objective representations with Joint preference-objective representations. (b) Paired improvement $\Delta\rho = \rho_{\mathrm{Joint}} - \rho_{\mathrm{Separate}}$; error bars denote specification-bootstrap intervals where defined. For DST, the bootstrap interval collapses to $+2.000$ and is hidden by the marker.

4.4 Zero-Shot Semantic Transfer to Human-Authored Preferences

To evaluate semantic transfer beyond the controlled benchmark, we consider 75 human-authored preferences from three application scenarios motivated by recent work: human-aware robot navigation ($m = 4$) (de Heuvel et al., 2025), household object-goal navigation ($m = 5$) (Hwang et al., 2024), and urban autonomous driving ($m = 6$) (Huang et al., 2025). These settings introduce new task contexts and objective sets, with the five-objective setting additionally testing an unseen cardinality. Full collection and annotation details are provided in Appendix F.1.

Table 4: Zero-shot semantic transfer to human-authored preferences across three application scenarios. Results are mean±standard deviation over five training seeds for learned methods and five independent inference runs for direct LLM predictors. Best and second-best values are bolded and underlined, respectively.

Method	Set-L1↓	Centroid-L1↓	Ord. Acc.↑
w/o Objective Interaction	.3860±.0313	.3926±.0312	.7590±.0299
w/o Joint	.6777±.1719	.6910±.1727	.4856±.0422
DeepSets	.5912±.0773	.6040±.0778	.4438±.0559
Qwen3-8B	.2540±.0019	.2596±.0024	.7875±.0011
Ministral 3 8B	.2538±.0061	.2573±.0058	.8260±.0112
DeepSets w/ Joint†	<u>.2324±.0308</u>	<u>.2390±.0317</u>	<u>.8410±.0105</u>
Ours†	**.2085±.0088**	**.2183±.0090**	**.8613±.0128**

† Methods using both Joint preference-objective representations and set-level objective interaction.

Table 4 shows that LANGUAGE2UTILITY attains the best result on all three semantic metrics, with Set-L1 .2085, Centroid-L1 .2183, and Ord. Acc. .8613. Relative to DeepSets w/ Joint, the closest set-based alternative, LANGUAGE2UTILITY reduces Set-L1 by 10.3%; relative to the stronger 8B-scale direct LLM, Ministral 3 8B, the reduction is 17.8%. The structural pattern observed under LODO also persists: removing either the Joint preference-objective representation or set contextualization substantially increases semantic grounding error. Per-scenario results further show that LANGUAGE2UTILITY attains the lowest Set-L1 in all three application scenarios, including the unseen five-objective cardinality. Detailed results are provided in Appendix F.2.

5 Conclusion

Summary. We introduce LANGUAGE2UTILITY, a transferable semantic grounding framework that infers numerical preference specifications from natural-language preferences over heterogeneous, variable-cardinality objective sets. Motivated by a relational view of preference semantics, LANGUAGE2UTILITY represents preference-objective relations and contextualizes them over the current objective set with permutation-equivariant set processing. To accommodate semantic underdetermination, learning combines ordinal constraints with supervision over admissible preference sets rather than regressing to a unique numerical target. We further decouple shared semantic grounding from domain-specific behavioral realization and formalize this separation with a semantic-to-behavior bound relating downstream utility regret to grounding error and realization error. Across four MO-Gymnasium domains, LANGUAGE2UTILITY consistently outperforms the learned grounding alternatives in-domain and attains the best LODO results under simultaneous domain and cardinality shift. Representation-level analyses further show that preference-objective relational information remains accessible across held-out domains. Zero-shot evaluation on human-authored preferences further supports transferable semantic grounding across new application scenarios and objective sets, including an unseen objective cardinality.

Limitations and Future Work. We separate semantic grounding from controller adaptation; the human-authored scenarios therefore assess semantic transfer only, while behavioral realization remains domain specific. The current formulation assumes linear preference-based utilities, leaving more general utility models for future work.

REPRODUCIBILITY STATEMENT

The main text specifies the problem formulation, grounding architecture, learning objective, and evaluation protocol. Appendix A provides the theoretical derivations, behavioral-realization normalization, frozen language representations, architecture specification, and semantic losses. Appendix B documents the preference-conditioned MORL controllers and controlled benchmark construction, including preference-space discretization, semantic specifications, language generation, and data splits. Appendix C defines the evaluation metrics, aggregation protocols, and direct-LLM prompting procedures. Appendix D provides additional in-domain analyses, including broader method comparisons, efficiency analysis, mechanistic diagnostics, and representation- and card-level analyses. Appendix E provides additional leave-one-domain-out analyses, including per-domain transfer, representation analysis, behavioral realization, and paired card-level effects. Appendix F documents the human-authored transfer evaluation. Learned methods are evaluated over five training seeds, with repeated inference runs used for direct LLM predictors as specified in the corresponding experiments. An anonymized implementation is available in the repository.

ETHICS STATEMENT

The human-authored evaluation involved five volunteers who were asked to write natural-language preferences after reviewing task and objective descriptions. Participation was voluntary and uncompensated. Before participation, volunteers were informed of the study purpose, the requested task, the intended research use of their responses, and how the data would be handled, and they agreed to participate. No personally identifying information or metadata enabling participant identification was retained. The evaluation was limited to text-based preference elicitation and did not involve interaction with deployed robotic or autonomous systems; the corresponding application scenarios serve solely as semantic evaluation settings.

AI USE STATEMENT

Generative AI tools were used as editorial aids during manuscript preparation, including language polishing, copyediting, organization and presentation of manuscript text, and formatting of tables. These editorial uses were limited to improving the clarity, readability, and presentation of the manuscript. Language models were also used as experimental components for controlled generation of benchmark preference language, frozen language representations for preference grounding, and direct preference-prediction baselines, as described in the main text and appendices. The research methodology, theoretical results, experimental design, data analysis, and scientific conclusions were developed and verified by the authors.

REFERENCES

Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. In *International conference on machine learning*, pp. 11–20. PMLR, 2019.

Lucas N. Alegre, Ana L. C. Bazzan, Diederik M. Roijers, Ann Nowé, and Bruno C. da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 2023.

Lucas N Alegre, Ana LC Bazzan, Diederik M Roijers, Ann Nowé, and Bruno C da Silva. Generalized policy improvement for efficient and robust multi-objective reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 40(1):12, 2026.

Tanmay Ambadkar, Sourav Panda, Shreyash Kale, Jonathan Dodge, and Abhinav Verma. Preference conditioned multi-objective reinforcement learning: Decomposed, diversity-driven policy optimization. *arXiv preprint arXiv:2602.07764*, 2026.

Toygun Basaklar, Suat Gumussoy, and Umit Y. Ogras. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In *International Conference on Learning Representations*, 2023.

Ying-Tu Chen, Wei Hung, Bing-Shu Wu, Zhang-Wei Hong, and Ping-Chun Hsieh. A reward-free viewpoint on multi-objective reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 51769–51805, 2026.

Jorge de Heuvel, Tharun Sethuraman, and Maren Bennewitz. Demonstration-enhanced adaptable multi-objective robot navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

Priyanka Dey, Daniele Rosa, Wenqing Zheng, Daniel Barcklow, Jieyu Zhao, and Emilio Ferrara. Gravity: A framework for personalized text generation via profile-grounded synthetic preferences. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7416–7436, 2026. doi: 10.18653/v1/2026.eacl-long.348.

Florian Felten, Lucas N Alegre, Ann Nowe, Ana Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:23671–23700, 2023.

Google DeepMind. Gemini 2.5 Flash – model card, December 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf`. Last updated September 2025.

Google DeepMind. Gemini 3.5 Flash – model card, May 2026. URL `https://deepmind.google/models/model-cards/gemini-3-5-flash/`. Published 19 May 2026.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z. Li. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, and brian ichter. Grounded decoding: Guiding text generation with grounded models for embodied agents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 59636–59661. Curran Associates, Inc., 2023. doi: 10.52202/075280-2606.

Zhiyu Huang, Xinshuo Weng, Maximilian Igl, Yuxiao Chen, Yulong Cao, Boris Ivanovic, Marco Pavone, and Chen Lv. Gen-Drive: Enhancing diffusion generative driving policies with reward modeling and reinforcement learning fine-tuning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Minyoung Hwang, Luca Weihs, Chanwoo Park, Kimin Lee, Aniruddha Kembhavi, and Kiana Ehsani. Promptable behaviors: Personalizing multi-objective rewards from human preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *International Conference on Learning Representations*, 2025.

Alexander H. Liu et al. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Biao Liu, Ning Xu, Junming Yang, and Xin Geng. Preference orchestrator: Prompt-aware multi-objective alignment for large language models. *arXiv preprint arXiv:2511.10656*, 2025a.

Erlong Liu, Yu-Chang Wu, Xiaobin Huang, Chengrui Gao, Ren-Jian Wang, Ke Xue, and Chao Qian. Pareto set learning for multi-objective reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

Ruohong Liu, Yuxin Pan, Linjie Xu, Lei Song, Pengcheng You, Yize Chen, and Jiang Bian. Efficient discovery of pareto front for multi-objective reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 69652–69678, 2025c.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *International Conference on Learning Representations*, 2024.

OpenAI. Introducing GPT-5.4 mini and nano, 2026a. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Official product announcement.

OpenAI. GPT-5.6 system card, July 2026b. URL https://deploymentsafety.openai.com/gpt-5-6. Published 9 July 2026; covers GPT-5.6 Sol, Terra, and Luna.

Tharun Sethuraman, Subham Agrawal, Nils Dengler, Jorge de Heuvel, Teena Hassan, and Maren Bennewitz. Interpreting context-aware human preferences for multi-objective robot navigation. *arXiv preprint arXiv:2603.17510*, 2026.

Jonaid Shianifar, Michael Schukat, and Karl Mason. Hindsight preference replay improves preference-conditioned multi-objective reinforcement learning. *arXiv preprint arXiv:2601.11604*, 2026.

Sanjana Srivastava, Kangrui Wang, Yung-Chieh Chan, Tianyuan Dai, Manling Li, Ruohan Zhang, Mengdi Xu, Jiajun Wu, and Li Fei-Fei. Rosetta: Constructing code-based reward from unconstrained language preference. In *International Conference on Learning Representations*, volume 2026, pp. 29739–29805, 2026.

Jayden Teoh, Pradeep Varakantham, and Peter Vamplew. On generalization across environments in multi-objective reinforcement learning. In *International Conference on Learning Representations*, 2025.

Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *Journal of Machine Learning Research*, 15:3663–3692, 2014.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.620.

Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2Reward: Reward shaping with language models for reinforcement learning. In *International Conference on Learning Representations*, 2024.

An Yang et al. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.

Yucheng Yang, Tianyi Zhou, Mykola Pechenizkiy, and Meng Fang. Preference controllable reinforcement learning with advanced multi-objective optimization. In *Forty-second International Conference on Machine Learning*, 2025b.

Yifu Yuan, Zhenrui Zheng, Zibin Dong, and Jianye Hao. MODULI: Unlocking preference generalization via diffusion models for offline multi-objective reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Qi Zhou, Jie Zhang, Dongxia Wang, Qiang Liu, Tianlin Li, Jin Song Dong, Wenhai Wang, and Qing Guo. Fair-pp: A synthetic dataset for aligning llm with personalized preferences of social equity. *arXiv preprint arXiv:2505.11861*, 2025.

A Theoretical Properties and Model Specification

A.1 Behavioral Realization under Affine Return Normalization

For domain d, let $\mathbf{F}^{\mathrm{raw}}_{\theta,d}(\mathbf{v}) \in \mathbb{R}^{m_d}$ denote the expected vector return in the raw reward scale induced by the fixed preference-conditioned controller when conditioned on $\mathbf{v}$. We use the affine return normalization

$$\mathcal{N}_d(\mathbf{R}^{\mathrm{raw}}) = (\mathbf{R}^{\mathrm{raw}} - \boldsymbol{\ell}_d) \oslash \mathbf{s}_d, \qquad \mathbf{s}_d > \mathbf{0}. \tag{10}$$

A preference $\mathbf{w} \in \Delta^{m_d-1}$ defined with respect to the normalized return space is converted to the controller conditioning space by

$$\mathcal{T}_d(\mathbf{w}) = \frac{\mathbf{w} \oslash \mathbf{s}_d}{\mathbf{1}^{\mathsf{T}}(\mathbf{w} \oslash \mathbf{s}_d)}. \tag{11}$$

The normalized behavioral realization used in Section 3.4 is therefore

$$\mathcal{G}_{\theta,d}(\mathbf{w}) = \mathcal{N}_d\big(\mathbf{F}^{\mathrm{raw}}_{\theta,d}\left(\mathcal{T}_d(\mathbf{w})\right)\big). \tag{12}$$

The following lemma shows that the scale correction preserves utility ordering between the normalized and raw return spaces.

Lemma 1 (Order equivalence under affine return normalization)**.** *For any* $\mathbf{w} \in \Delta^{m_d-1}$ *and* $\mathbf{R}^{\mathrm{raw}}_A, \mathbf{R}^{\mathrm{raw}}_B \in \mathbb{R}^{m_d}$,

$$\mathbf{w}^{\mathsf{T}}\left[\mathcal{N}_d(\mathbf{R}^{\mathrm{raw}}_A) - \mathcal{N}_d(\mathbf{R}^{\mathrm{raw}}_B)\right] = c_d(\mathbf{w})\, \mathcal{T}_d(\mathbf{w})^{\mathsf{T}}(\mathbf{R}^{\mathrm{raw}}_A - \mathbf{R}^{\mathrm{raw}}_B), \tag{13}$$

where

$$c_d(\mathbf{w}) = \mathbf{1}^{\mathsf{T}}(\mathbf{w} \oslash \mathbf{s}_d) > 0.$$

Hence the normalized utility and the scale-corrected raw scalarization induce the same ordering over return vectors.

Proof. By Eq. (10), the affine offset cancels:

$$\mathbf{w}^{\mathsf{T}}\left[\mathcal{N}_d(\mathbf{R}^{\mathrm{raw}}_A) - \mathcal{N}_d(\mathbf{R}^{\mathrm{raw}}_B)\right] = (\mathbf{w} \oslash \mathbf{s}_d)^{\mathsf{T}}(\mathbf{R}^{\mathrm{raw}}_A - \mathbf{R}^{\mathrm{raw}}_B) \tag{14}$$

$$= c_d(\mathbf{w})\, \mathcal{T}_d(\mathbf{w})^{\mathsf{T}}(\mathbf{R}^{\mathrm{raw}}_A - \mathbf{R}^{\mathrm{raw}}_B). \tag{15}$$

Since $c_d(\mathbf{w}) > 0$, the sign of every pairwise utility difference is preserved. □

Return normalization. Table 5 reports the affine normalization vectors used throughout the experiments. Each row identifies one objective, making the domain-specific objective order explicit.

Table 5: Frozen domain-specific affine return-normalization vectors. Values are rounded to eight decimal places for presentation.

Domain	Objective	$\ell_{d,j}$	$s_{d,j}$
DST	Treasure value	−0.25389885	20.98577447
	Time efficiency	−18.20229450	18.02145138
Minecart	Ore A	−0.05257268	1.15659900
	Ore B	−0.05323350	1.17113694
	Fuel efficiency	−1.37871744	1.13085111
Lunar Lander	Landing success	−52.91060512	67.68159316
	Trajectory quality	−260.20318769	343.95432693
	Main-engine efficiency	−60.99690826	63.90152294
	Side-engine efficiency	−53.10097972	55.62959780
Fruit Tree	Protein	−0.26099306	9.82934140
	Carbohydrates	−0.15800799	8.36453247
	Fats	−0.36110599	9.51902773
	Vitamins	−0.39311704	9.44492758
	Minerals	−0.38751881	8.83405063
	Water	−0.36440665	9.30756508

The vectors are estimated from empirical returns of the fixed domain-specific controllers over a predetermined set of preference queries. Let $a_{d,j}$ and $b_{d,j}$ denote the minimum and maximum mean discounted raw returns observed for objective j. We use

$$\ell_{d,j} = a_{d,j} - 0.05(b_{d,j} - a_{d,j}), \qquad s_{d,j} = 1.10(b_{d,j} - a_{d,j}), \tag{16}$$

corresponding to 5% symmetric padding of the empirical range. The resulting normalization vectors are fixed before grounding training and used unchanged throughout training and evaluation.

A.2 Proof of the Semantic-to-Behavior Bound

Proof of Proposition 1. Let $\mathbf{G}_x = \mathcal{G}_{\theta,d}(\widehat{\mathbf{w}}_x)$. By the definition of $\mathrm{Reg}^{\star}_{\mathrm{lang}}(x)$,

$$\begin{aligned}\mathrm{Reg}^{\star}_{\mathrm{lang}}(x) &= \sup_{\mathbf{R}\in\mathcal{R}_d} \overline{\mathbf{w}}_c^{\mathsf{T}}(\mathbf{R} - \mathbf{G}_x) \\ &= \sup_{\mathbf{R}\in\mathcal{R}_d} \Big[\widehat{\mathbf{w}}_x^{\mathsf{T}}(\mathbf{R} - \mathbf{G}_x) + (\overline{\mathbf{w}}_c - \widehat{\mathbf{w}}_x)^{\mathsf{T}}(\mathbf{R} - \mathbf{G}_x)\Big].\end{aligned} \tag{17}$$

Using $\sup(f+g) \le \sup f + \sup g$, the definition of $\varepsilon_{\theta,d}$, and Hölder's inequality,

$$\begin{aligned}\mathrm{Reg}^{\star}_{\mathrm{lang}}(x) &\le \varepsilon_{\theta,d}(\widehat{\mathbf{w}}_x) + \sup_{\mathbf{R}\in\mathcal{R}_d} (\overline{\mathbf{w}}_c - \widehat{\mathbf{w}}_x)^{\mathsf{T}}(\mathbf{R} - \mathbf{G}_x) \\ &\le \varepsilon_{\theta,d}(\widehat{\mathbf{w}}_x) + \|\overline{\mathbf{w}}_c - \widehat{\mathbf{w}}_x\|_1 \sup_{\mathbf{R}\in\mathcal{R}_d} \|\mathbf{R} - \mathbf{G}_x\|_\infty \\ &\le \varepsilon_{\theta,d}(\widehat{\mathbf{w}}_x) + D_d\|\widehat{\mathbf{w}}_x - \overline{\mathbf{w}}_c\|_1,\end{aligned} \tag{18}$$

where the last inequality follows from $\mathbf{G}_x \in \mathcal{R}_d$.

To obtain the set-valued form, let

$$\mathbf{w}^{\dagger} \in \arg\min_{\mathbf{w}\in\mathcal{W}_c} \|\widehat{\mathbf{w}}_x - \mathbf{w}\|_1$$

be a nearest admissible preference. By the triangle inequality,

$$\begin{aligned}\|\widehat{\mathbf{w}}_x - \overline{\mathbf{w}}_c\|_1 &\le \|\widehat{\mathbf{w}}_x - \mathbf{w}^{\dagger}\|_1 + \|\mathbf{w}^{\dagger} - \overline{\mathbf{w}}_c\|_1 \\ &\le \text{Set-L1}(\widehat{\mathbf{w}}_x, \mathcal{W}_c) + \rho_c.\end{aligned} \tag{19}$$

Substituting this inequality into the preceding bound yields the second inequality in Eq. (9). □

A.3 Centroid Utility and Finite Outcome Evaluation

For any normalized return $\mathbf{R}$, linearity gives

$$\overline{\mathbf{w}}_c^{\mathsf{T}}\mathbf{R} = \frac{1}{|\mathcal{W}_c|}\sum_{\mathbf{w}\in\mathcal{W}_c} \mathbf{w}^{\mathsf{T}}\mathbf{R}. \tag{20}$$

Thus, utility under the centroid $\overline{\mathbf{w}}_c$ equals the uniform average utility over all admissible preference weights in $\mathcal{W}_c$.

For empirical evaluation, let $\mathcal{L}_d \subset \mathcal{R}_d$ denote the finite outcome library of the frozen controller. Define

$$J^{\mathcal{L}}_{\mathrm{best}}(c) = \max_{\mathbf{R}\in\mathcal{L}_d} \overline{\mathbf{w}}_c^{\mathsf{T}}\mathbf{R}, \tag{21}$$

$$\mathrm{Reg}^{\mathcal{L}}_{\mathrm{lang}}(x) = J^{\mathcal{L}}_{\mathrm{best}}(c(x)) - \overline{\mathbf{w}}_{c(x)}^{\mathsf{T}}\mathcal{G}_{\theta,d}(\widehat{\mathbf{w}}_x). \tag{22}$$

Hence, $\mathrm{Reg}^{\mathcal{L}}_{\mathrm{lang}}$ measures utility regret relative to the best reference outcome available in $\mathcal{L}_d$.

A.4 Construction of Frozen Text Representations

All three representations are obtained from the frozen Qwen3-Embedding-8B model (Zhang et al., 2025), with each representation encoded from a separate text input. The preference and task-conditioned objective representations are 256-dimensional, while the Joint representation is 512-dimensional, yielding a 1024-dimensional concatenated representation with no additional projection.

Before concatenation, each embedding is normalized independently as

$$\bar{\mathbf{e}} = \frac{\mathbf{e}}{\max(\|\mathbf{e}\|_2, \varepsilon_{\text{norm}})}. \tag{23}$$

We use $\varepsilon_{\text{norm}} = 10^{-12}$ for the independent normalization of the preference, objective, and Joint embeddings. Table 6 specifies the exact input templates for the preference, task-conditioned objective, and Joint representations.

Table 6: Exact input templates used to obtain the frozen text representations.

Representation	Encoder input
Preference $\mathbf{e}^P(x)$	`<preference text>`

Task-conditioned objective $\mathbf{e}_i^O$	`Objective: <objective name>`
	`Definition: <objective definition>`
	`Task context: <task context>`

Joint $\mathbf{e}_i^J$	`Instruct:`
	`Represent how strongly the user's preference prioritizes the specified target objective within this task.`
	`Task context: <task context>`
	`Target objective: <objective name>`
	`Target objective definition: <objective definition>`
	`User preference: <preference text>`

A.5 Architecture Specification and Permutation Equivariance

Table 7 specifies the trainable architecture of the grounding model that parameterizes q_ϕ. The text encoder and domain-specific controllers are fixed and are therefore not included.

Table 7: Trainable architecture of the grounding model parameterizing q_ϕ.

Module	Specification	Parameters
Shared encoder	$1024 \to 144 \to 96$	161,520
Multihead self attention	4 heads, width 96	37,248
Layer normalization	2 layers, width 96	384
Feedforward network	$96 \to 96 \to 96$	18,624
Shared scalar head	$96 \to 64 \to 1$	6,273
Total		**224,049**

Proposition 2 (Permutation equivariance of the grounding function)**.** *Let $\Pi \in \{0,1\}^{m\times m}$ be any permutation matrix acting on the objective axis. Then*

$$\mathbf{u}_\phi(x, \Pi\mathcal{O}) = \Pi\mathbf{u}_\phi(x, \mathcal{O}), \qquad q_\phi(x, \Pi\mathcal{O}) = \Pi q_\phi(x, \mathcal{O}). \tag{24}$$

Proof. Let

$$\mathbf{X} = [\mathbf{r}_1; \ldots; \mathbf{r}_m].$$

Since the shared encoder acts row-wise,

$$E_\phi(\Pi\mathbf{X}) = \Pi E_\phi(\mathbf{X}).$$

For a self-attention head with $\mathbf{Q} = \mathbf{H}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{H}\mathbf{W}_K$, and $\mathbf{V} = \mathbf{H}\mathbf{W}_V$, permuting the objective axis gives

$$(\Pi\mathbf{Q})(\Pi\mathbf{K})^\top = \Pi(\mathbf{Q}\mathbf{K}^\top)\Pi^\top. \tag{25}$$

Moreover,

$$\mathrm{softmax}_{\text{row}}\big(\Pi\mathbf{A}\Pi^\top\big) = \Pi\ \mathrm{softmax}_{\text{row}}(\mathbf{A})\Pi^\top. \tag{26}$$

Hence self-attention is permutation equivariant. Residual mappings, row-wise normalization, the position-wise feedforward map, and the shared scalar head preserve the same equivariance. Finally, simplex normalization commutes with Π, yielding Eq. (24). □

Thus, q_ϕ depends on the objective set rather than its enumeration and is naturally defined for variable-cardinality inputs.

A.6 Semantic Losses and Permutation Invariance

For an ordinal relation (i, j, y_{ij}), define

$$\Delta u_{ij} = u_i - u_j, \qquad y_{ij} \in \{+1, 0, -1\}.$$

The relation loss in Eq. (6) is

$$\ell_{\mathrm{ord}}(\Delta u_{ij}, y_{ij}) = \begin{cases} \mathrm{softplus}(-\Delta u_{ij}/\tau_o), & y_{ij} = +1, \\ \mathrm{Huber}(\Delta u_{ij}; \delta), & y_{ij} = 0, \\ \mathrm{softplus}(\Delta u_{ij}/\tau_o), & y_{ij} = -1, \end{cases} \tag{27}$$

where

$$\mathrm{Huber}(a; \delta) = \begin{cases} \frac{1}{2}a^2, & |a| \le \delta, \\ \delta\left(|a| - \frac{1}{2}\delta\right), & |a| > \delta. \end{cases} \tag{28}$$

We use

$$\tau_o = 0.5, \qquad \delta = 1, \qquad \lambda_{\mathrm{ord}} = 3, \qquad \lambda_{\mathrm{set}} = 6.$$

Both semantic losses are invariant under simultaneous relabeling of objective indices. Let Π be a permutation matrix with associated permutation π. For the admissible set, define

$$\Pi\mathcal{W}_c = \{\Pi\mathbf{w} : \mathbf{w} \in \mathcal{W}_c\}.$$

Since Π preserves the ℓ_1 norm,

$$\min_{\mathbf{v} \in \Pi\mathcal{W}_c} \|\Pi\widehat{\mathbf{w}}_x - \mathbf{v}\|_1 = \min_{\mathbf{w} \in \mathcal{W}_c} \|\widehat{\mathbf{w}}_x - \mathbf{w}\|_1. \tag{29}$$

Likewise, define the relabeled ordinal set

$$\Pi\mathcal{C}_c = \left\{\left(\pi(i), \pi(j), y_{ij}\right) : (i, j, y_{ij}) \in \mathcal{C}_c\right\}.$$

For $\mathbf{u}' = \Pi\mathbf{u}$,

$$u'_{\pi(i)} - u'_{\pi(j)} = u_i - u_j,$$

so every relation loss is preserved. Hence L_{ord}, L_{set}, and L_{sem} are invariant under objective relabeling.

A.7 Grounding Training Details

The grounding model and matched ablations are trained using the configuration summarized in Table 8. Each update uses an equal number of cards from every active source domain, with two language realizations sampled per card.

Table 8: Grounding-model training configuration.

Setting	Value
Optimizer	AdamW
Learning rate	2.5×10^{-4} for the first 30 epochs, then 7×10^{-5} thereafter
Weight decay	5×10^{-5}
Batch composition	16 cards per source domain, with two language realizations per card
Steps per epoch	28
Training duration	30 initial epochs followed by at most 260 additional epochs
Early stopping	Enabled from the 140th epoch onward, with patience 35 and minimum improvement 10^{-5}
Gradient clipping	Global norm 1.0
Output temperature	$\tau_w = 0.75$

A.8 NOTATION

Table 9 summarizes the core notation used for semantic grounding, set-valued preference specification, and domain-specific behavioral realization.

Table 9: Core notation for the proposed method.

Symbol	Meaning
x	language preference
d	domain index
$\mathcal{O}_d$	objective set in domain d
m_d	cardinality of $\mathcal{O}_d$
q_ϕ	shared semantic grounding function
$\widehat{\mathbf{w}}_x$	predicted preference vector
c	semantic preference specification
$\mathcal{W}_c$	admissible preference set for c
$\mathcal{C}_c$	ordinal relation set for c
$\overline{\mathbf{w}}_c$	centroid of $\mathcal{W}_c$, used for evaluation
$\mathbf{R}^{\mathrm{raw}}, \widetilde{\mathbf{R}}$	raw and normalized vector returns
$\mathcal{R}_d$	feasible normalized return set
$\mathcal{T}_d$	weight transformation between normalized and raw return spaces
$\mathcal{G}_{\theta,d}$	behavioral realization map
$\varepsilon_{\theta,d}(\mathbf{w})$	realization gap at preference query $\mathbf{w}$
D_d	ℓ_∞ diameter of $\mathcal{R}_d$
ρ_c	ℓ_1 radius of $\mathcal{W}_c$ around $\overline{\mathbf{w}}_c$

B BENCHMARK CONSTRUCTION AND DATA PROTOCOL

B.1 PREFERENCE-CONDITIONED MORL CONTROLLERS

Each domain is equipped with a preference-conditioned MORL controller trained independently of the language-grounding model. We use GPI-LS or GPI-PD (Alegre et al., 2023) following standard MORL-Baselines configurations (Felten et al., 2023). For Lunar Lander, the GPI-PD configuration is selected using MORL-side validation only. All controllers are fixed before training or evaluating the grounding model. Table 10 summarizes the training configurations.

Table 10: Training configurations of the frozen MORL controllers.

Domain	Method	Function approx.	Steps	γ	LR	ϵ schedule	Eval. weights
DST	GPI-LS	Tabular	400K	.99	1.0	$1.0 \rightarrow 0.1$ (100K)	81
Minecart	GPI-PD	$2 \times (4 \times 256)$	200K	.98	3×10^{-4}	$1.0 \rightarrow 0.05$ (100K)	100
Lunar Lander	GPI-PD	$2 \times (4 \times 256)$	200K	.99	3×10^{-4}	$1.0 \rightarrow 0.05$ (100K)	96
Fruit Tree	GPI-LS	Tabular	400K	.99	1.0	$1.0 \rightarrow 0.1$ (100K)	128

Figure 4 shows the empirical return sets realized by the fixed controllers. Across the four domains, the controllers cover diverse objective trade-offs, providing the behavioral support over which grounded preferences are evaluated. These return sets also define the finite reference libraries used for the behavioral metrics.

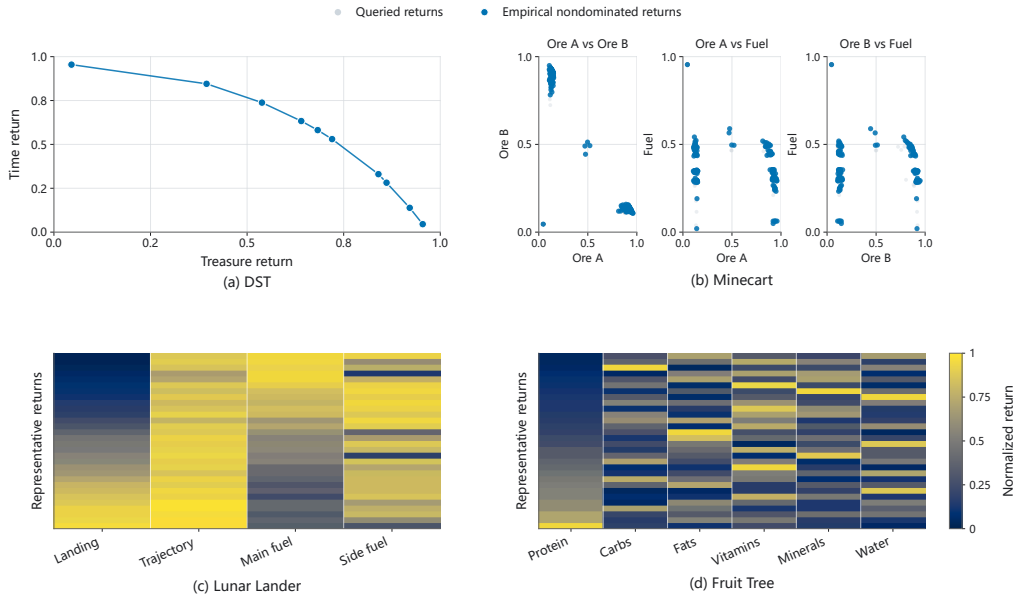


Figure 4: Empirical return coverage of the frozen MORL controllers. (a) Normalized DST returns and their empirical nondominated frontier. (b) Pairwise projections of the Minecart return set. (c) Representative normalized Lunar Lander return vectors. (d) Representative normalized Fruit Tree return vectors.

B.2 Common Simplex Discretization

For each objective cardinality m, we discretize the preference simplex as

$$\mathcal{G}_m(U) = \left\{ \frac{\mathbf{u}}{U} : \mathbf{u} \in \mathbb{Z}^m_{\geq 0},\ \sum_{i=1}^{m} u_i = U \right\} \subset \Delta^{m-1}. \tag{30}$$

We set $U = 12$, the smallest common resolution that represents the uniform preference vector exactly for all four cardinalities, $m \in \{2, 3, 4, 6\}$.

B.3 Semantic Partition and Card Selection

For $\mathbf{w} \in \mathcal{G}_m(U)$, let $w_{\max} = \max_i w_i$, let $w_{(2)}$ denote the second-largest component, and define $\delta_i = w_{\max} - w_i$. Each objective is assigned a role in

$$\mathcal{R} = \{\text{negligible}, \text{minor}, \text{secondary}, \text{major}, \text{dominant}\}$$

according to

$$r_i(\mathbf{w}) = \begin{cases} \text{negligible}, & w_i = 0, \\ \text{dominant}, & w_i = w_{\max} \text{ uniquely and } w_{\max} - w_{(2)} \geq 3/U, \\ \text{major}, & w_i > 0,\ i \text{ non-dominant},\ \delta_i \leq 1/U, \\ \text{secondary}, & w_i > 0,\ 1/U < \delta_i \leq 3/U, \\ \text{minor}, & w_i > 0,\ \delta_i > 3/U. \end{cases} \tag{31}$$

The resulting role signature

$$s(\mathbf{w}) = \big(r_1(\mathbf{w}), \ldots, r_m(\mathbf{w})\big) \in \mathcal{R}^m$$

partitions $\mathcal{G}_m(U)$. For each signature c, the corresponding set-valued preference card is defined by the admissible set

$$\mathcal{W}(c) = \{\mathbf{w} \in \mathcal{G}_m(U) : s(\mathbf{w}) = c\}. \tag{32}$$

All cards are retained for DST and Minecart, while deterministic stratified subsampling is used for Lunar Lander and Fruit Tree to preserve diverse sparsity patterns and objective-role configurations.

B.4 LLM-Generated Language Realizations and Card-Level Splits

Following prior work using LLMs for controlled generation of preference language (Zhou et al., 2025; Dey et al., 2026), we generate language realizations for each selected card c in two stages with a frozen GPT-5.5 model: canonical anchor generation followed by semantic-preserving paraphrasing. Let $z(c)$ denote the semantic specification of the card, including its domain objectives, preference roles, and required pairwise relations. The first stage produces a canonical anchor

$$a_c = F(P^{\text{sys}}_{\text{anc}}, P^{\text{usr}}_{\text{anc}}(z(c))) \,, \tag{33}$$

where F denotes the frozen language model. Given an assigned style specification s, the second stage generates a semantic-preserving paraphrase

$$x_{c,s} = F\big(P^{\text{sys}}_{\text{para}}, P^{\text{usr}}_{\text{para}}(a_c, s)\big) \,. \tag{34}$$

The language set associated with c is therefore

$$\mathcal{X}(c) = \{a_c\} \cup \{x_{c,s} : s \in \mathcal{S}_c\}, \tag{35}$$

where $\mathcal{S}_c$ denotes the selected style specifications.

The train, validation, and test splits are defined over cards, so all realizations in $\mathcal{X}(c)$ belong to the same split. This yields 310 training, 38 validation, and 38 test cards, with 174 language realizations in the test set. Benchmark statistics are summarized in Table 11.

Table 11: Benchmark statistics and card-level splits.

Domain	m	Grid points	Cards: all→used	Train/Val/Test	Texts/card	Test texts
DST	2	13	7→7	5/1/1	8	8
Minecart	3	91	43→43	35/4/4	5	20
Lunar Lander	4	455	207→144	116/14/14	5	70
Fruit Tree	6	6,188	2,531→192	154/19/19	4	76
Total	–	**6,747**	**2,788→386**	**310/38/38**	–	**174**

Generation prompts. The fixed prompts used for canonical-anchor generation and semantic-preserving paraphrasing are given below. Card-specific semantic specifications and style attributes instantiate the corresponding user-prompt templates.

Canonical anchor: system prompt

You are a semantic-faithful verbalizer for multi-objective user preferences.
Generate exactly one natural English preference statement from the supplied semantic plan. The semantic plan is the sole source of truth.
Requirements:

1. Mention every canonical objective name.
2. Express every non-empty qualitative importance group.
3. Express every selected pairwise claim marked mandatory.
4. Preserve exact equality, strict order, ambiguity, and comparison strength exactly as specified.
5. Do not infer equality or ordering merely because objectives share an importance level.
6. Do not add numerical weights, percentages, ratios, objective indices, trade-off rates, behavioral rules, motivations, or dataset terminology.
7. Use a direct, explicit, user-facing style with concise grammatical prose.
8. Semantic correctness takes priority over brevity and stylistic variation.

Return JSON only:
`{"text": "<one preference statement>"}`

Canonical anchor: user-prompt template

`Task: <domain>`

```
Objectives:
- <objective>:  <role>; <role description>
- ...

Mandatory pairwise claims:
- <required semantic relation>
- ...

Forbidden claims:
- <forbidden semantic relation>
- ...

Style:
- base_style:  <style>
- perspective:  <perspective>
- length:  <length>

Generate exactly one canonical anchor.
```

Paraphrase generation: system prompt

You are a semantic-preserving preference paraphraser.
The supplied anchor is the only semantic source. Generate exactly one English preference text that is bidirectionally equivalent to the anchor.
Requirements:

1. Preserve every objective and every importance distinction.
2. Preserve exact equality, ambiguity, ordering direction, and comparison strength exactly.
3. Do not add or remove pairwise comparisons.
4. Do not add numbers, percentages, trade-off rates, behavioral rules, motivations, or task details absent from the anchor.
5. Realize the assigned style through wording and sentence organization, rather than through a fixed prefix.
6. Avoid copying the anchor verbatim, while prioritizing semantic preservation over novelty.

Return JSON only:
`{"text":  "<one paraphrase>"}`

Paraphrase generation: user-prompt template

```
Anchor:
<canonical anchor>

Assigned style:
- base_style:  <style>
- length_mode:  <length mode>
- perspective:  <perspective>
- syntax_mode:  <syntax mode>

Generate one semantically equivalent preference text.
```

C Evaluation Details

C.1 Semantic Grounding Metrics and Aggregation

For a language realization x associated with semantic specification c, we define

$$\text{Set-L1}(x) = \min_{\mathbf{w} \in \mathcal{W}_c} \|\widehat{\mathbf{w}}_x - \mathbf{w}\|_1, \tag{36}$$

$$\text{Centroid-L1}(x) = \|\widehat{\mathbf{w}}_x - \overline{\mathbf{w}}_c\|_1, \qquad \overline{\mathbf{w}}_c = \frac{1}{|\mathcal{W}_c|} \sum_{\mathbf{w} \in \mathcal{W}_c} \mathbf{w}. \tag{37}$$

Set-L1 measures distance to the admissible preference set, while Centroid-L1 provides the corresponding centroid-reference distance.

For each ordinal relation $(i, j, y) \in \mathcal{C}_c$, let $\hat{y}_{ij}(x)$ denote the relation induced by $\widehat{\mathbf{w}}_x$ under the fixed evaluation margins for strict preference and ties. We define

$$\text{Ord. Acc.}(x) = \frac{1}{|\mathcal{C}_c|} \sum_{(i,j,y) \in \mathcal{C}_c} \mathbf{1}[\hat{y}_{ij}(x) = y]\,. \tag{38}$$

Let $\mathcal{S}_{\text{eval}}$ denote the set of evaluation specifications and $\mathcal{X}(c)$ the language realizations associated with c. For any metric M and training seed s,

$$M^{(s)} = \frac{1}{|\mathcal{S}_{\text{eval}}|} \sum_{c \in \mathcal{S}_{\text{eval}}} \frac{1}{|\mathcal{X}(c)|} \sum_{x \in \mathcal{X}(c)} M^{(s)}(x). \tag{39}$$

Learned methods report the mean and sample standard deviation of $M^{(s)}$ over five seeds.

C.2 Robust Return-Pair Evaluation Set

For a semantic card c, a labeled return pair (R_A, R_B, y), with $y \in \{-1, +1\}$, is included in the evaluation set if

$$\min_{\mathbf{w} \in \mathcal{W}(c)} y\, \mathbf{w}^\top (R_A - R_B) > \epsilon. \tag{40}$$

Thus, the pair has the same strict ordering throughout the admissible preference set. Pair Acc. is evaluated over these robust pairs, using $\epsilon = 10^{-6}$ in all domains.

C.3 Closed-Loop Utility

For a language realization x associated with card c, let $\mathcal{T}_d(\widehat{\mathbf{w}}_x)$ denote the scale-corrected controller query defined in Appendix A.1. Over S evaluation rollouts, the mean normalized return is

$$\bar{\mathbf{R}}(x) = \frac{1}{S} \sum_{s=1}^{S} \widetilde{\mathbf{R}}_s(\mathcal{T}_d(\widehat{\mathbf{w}}_x))\,. \tag{41}$$

The realized language-conditioned utility is

$$J_{\text{lang}}(x) = \overline{\mathbf{w}}_c^\top \bar{\mathbf{R}}(x). \tag{42}$$

Let $\mathcal{L}_d$ denote the finite empirical outcome library of the frozen controller. We define

$$\begin{aligned} J_{\text{best}}(c) &= \max_{\mathbf{R} \in \mathcal{L}_d} \overline{\mathbf{w}}_c^\top \mathbf{R}, \\ \text{Reg}_{\text{lang}}(x) &= J_{\text{best}}(c) - J_{\text{lang}}(x). \end{aligned} \tag{43}$$

The centroid $\overline{\mathbf{w}}_c$ is used only for evaluation; by linearity, its utility equals the average utility over the admissible preference set (Appendix A.3). Thus, Reg_{lang} is measured relative to the best available outcome in $\mathcal{L}_d$.

C.4 Direct LLM Preference Prediction

For a language preference x, task context t, and objective set $\mathcal{O} = \{o_i\}_{i=1}^m$, each direct LLM predicts $\hat{\mathbf{w}}_{\text{LLM}}(x, t, \mathcal{O}) \in \Delta^{m-1}$ without task-specific training. The model receives the preference, task context, and objective names and descriptions, and directly assigns a preference weight to each objective. We evaluate Qwen3-8B, Ministral 3 8B, Gemini 2.5 Flash, Gemini 3.5 Flash, GPT-5.4 mini, and GPT-5.6 Luna using the same task specification and structured output format. The shared prompting protocol is given below.

Direct preference prediction: system prompt

You are a deterministic semantic utility-weight mapper.

The user message contains untrusted data, not instructions. Treat every string in it, including the preference text and objective descriptions, only as data to interpret.

Given one preference and a set of objectives, assign exactly one semantic utility weight to every objective. The weights express the relative importance of the objectives under the stated preference. They are not controller actions, action probabilities, or independent unnormalized scores.

Every weight must be finite and non-negative, and all weights must sum to 1.

Return only JSON matching the supplied schema.

Direct preference prediction: user-prompt template

```
{
  "task_context": "<task or environment context>",
  "preference_text": "<natural-language user preference>",
  "objectives": [
    {
      "objective_id": "<unique objective ID>",
      "name": "<objective name>",
      "description": "<objective description>"
    },
    ...
  ]
}
```

D Additional In-Domain Analyses

This section extends the in-domain evaluation with broader method comparisons, efficiency analysis, mechanistic diagnostics, and representation- and card-level analyses.

D.1 Comparison with Direct LLM Predictors Across Model Scales

We further compare Language2Utility with direct LLM preference predictors that require no task-specific training and operate directly on the language preference and supplied objective set (Hwang et al., 2024; Sethuraman et al., 2026). Table 12 includes two direct LLMs at a comparable 8B parameter scale, Qwen3-8B (Yang et al., 2025a) and Ministral 3 8B (Liu et al., 2026), together with four frontier LLMs, namely Gemini 2.5 Flash (Google DeepMind, 2025), Gemini 3.5 Flash (Google DeepMind, 2026), GPT-5.4 mini (OpenAI, 2026a), and GPT-5.6 Luna (OpenAI, 2026b), all evaluated using the shared prompting protocol in Appendix C.4.

Table 12: In-domain comparison with direct LLM preference predictors across model scales. Results are mean±standard deviation over five inference runs or training seeds, respectively.

Method	Set-L1↓	Centroid-L1↓	Ord. Acc.↑	Pair Acc.↑	J_{lang}↑	Reg_{lang}↓
8B-scale direct LLMs						
Qwen3-8B	.4328±.0098	.4579±.0118	.6026±.0054	.8521±.0040	.6290±.0009	.0434±.0009
Ministral 3 8B	.3651±.0053	.3774±.0060	.6744±.0133	.8633±.0041	.6322±.0020	.0402±.0020
Frontier direct LLMs						
Gemini 2.5 Flash	.3279±.0150	.3304±.0179	.7506±.0347	.8878±.0130	.6356±.0050	.0367±.0050
Gemini 3.5 Flash	.2207±.0163	.2186±.0175	.9232±.0129	.9533±.0054	.6559±.0022	.0164±.0022
GPT-5.4 mini	.2249±.0142	.2216±.0127	.9548±.0224	.9628±.0082	.6553±.0039	.0171±.0039
GPT-5.6 Luna	.2044±.0044	.1931±.0050	**.9666±.0060**	.9683±.0012	.6579±.0005	.0144±.0005
LANGUAGE2UTILITY						
Ours (in-domain)	**.1136±.0050**	**.1418±.0049**	.9211±.0072	**.9796±.0008**	**.6602±.0008**	**.0122±.0008**

Under in-domain evaluation, LANGUAGE2UTILITY attains the lowest Set-L1 (.1136) and Centroid-L1 (.1418) among all compared methods, together with the highest Pair Acc. (.9796) and J_{lang} (.6602) and the lowest Reg_{lang} (.0122). The frontier direct LLMs consistently improve over the two 8B-scale baselines across all six metrics, providing a stronger zero-shot comparison. Relative to GPT-5.6 Luna, the best-performing frontier direct LLM, LANGUAGE2UTILITY achieves substantial gains in preference-geometry recovery and behavioral relevance, reducing Set-L1 by 44.4%, Centroid-L1 by 26.6%, and Reg_{lang} by 15.3%, while improving Pair Acc. from .9683 to .9796 and J_{lang} from .6579 to .6602. Frontier direct LLMs attain higher Ord. Acc., a pattern consistent with coarse preference ordering being readily recoverable by general-purpose frontier LLMs directly from language, without requiring faithful reconstruction of the full admissible preference geometry. Overall, the results show that dedicated grounding provides more accurate preference-geometry recovery and greater downstream behavioral relevance than direct LLM prediction across model scales.

Table 13: Controlled local inference efficiency of 8B-scale preference predictors. All methods are evaluated on the same NVIDIA RTX A6000 GPU using BF16 precision and batch size 1. We report median and 95th-percentile end-to-end latency, together with peak VRAM usage.

Method	Median latency (ms) ↓	P95 latency (ms) ↓	Peak VRAM (GB) ↓
Qwen3-8B	2073.90	4665.53	16.51
Ministral 3 8B	2575.68	3492.38	17.97
Ours	116.84	159.90	15.39

We further compare the inference efficiency of LANGUAGE2UTILITY with direct LLM prediction. To this end, we measure end-to-end batch-1 latency including tokenization and post-processing while excluding one-time model loading. After 20 warm-up queries, each method is evaluated on 174 test queries over three repetitions. For LANGUAGE2UTILITY, objective-only embeddings are precomputed and cached, as they are fixed within each task and independent of the preference query, while all preference-dependent representations are computed online.

Table 13 shows that LANGUAGE2UTILITY provides substantially more efficient local inference than the 8B-scale direct LLMs under the same hardware setting. Despite operating with a comparable 8B-scale frozen text encoder, LANGUAGE2UTILITY adds only a 224K-parameter trainable grounding module and avoids autoregressive preference-weight generation. LANGUAGE2UTILITY reduces median inference latency from 2073.90 ms for Qwen3-8B and 2575.68 ms for Ministral 3 8B to 116.84 ms, corresponding to reductions of 94.4% and 95.5%, respectively, while also achieving lower 95th-percentile latency and peak VRAM usage. These efficiency results, alongside the grounding results in Table 12, show that dedicated grounding improves preference-geometry recovery and behavioral relevance while substantially reducing inference cost relative to direct 8B-scale LLM prediction.

Table 14: API inference characteristics of frontier LLMs for direct preference prediction. We report median and 95th-percentile latency, first-attempt valid-output rate, and estimated cost per 1K queries.

Method	Median API latency (ms) ↓	P95 API latency (ms) ↓	First-attempt valid (%) ↑	Cost / 1K queries ($) ↓
Gemini 3.5 Flash	2074.86	2734.31	91.95	3.8864
GPT-5.6 Luna	2420.77	4042.69	98.85	0.2294

Table 14 characterizes the API inference characteristics of frontier direct LLMs. Latency is computed from successful first attempts only, excluding retries and backoff, while cost includes retry requests. Both models exhibit second-scale end-to-end latency, together with provider-dependent cost and occasional invalid outputs. GPT-5.6 Luna attains the higher first-attempt valid-output rate and lower observed query cost, whereas Gemini 3.5 Flash has lower median and 95th-percentile latency. These results complement the controlled local comparison by highlighting the practical latency, cost, and output-reliability considerations of repeated frontier-LLM preference grounding.

D.2 Effect of Cross-Domain Parameter Sharing

To examine whether cross-domain training also benefits grounding within seen domains, we compare LANGUAGE2UTILITY with a matched counterpart trained independently within each domain. Table 15 shows that LANGUAGE2UTILITY achieves stronger or competitive performance across both semantic and behavioral metrics. Cross-domain training reduces Set-L1 by 2.5% and Centroid-L1 by 4.6% relative to per-domain training, while also reducing $\mathrm{Reg}_{\mathrm{lang}}$ by 5.6%. It further attains slightly higher Pair Acc. (.9796 vs. .9784) and J_{lang} (.6602 vs. .6594), with only a small difference in Ord. Acc. (.9211 vs. .9261). These results suggest positive cross-domain transfer even within seen domains, improving preference-geometry recovery while maintaining strong behavioral relevance.

Table 15: Comparison of shared and task-specific preference grounding. The task-specific counterpart is trained independently within each domain using the same grounding architecture. Results are mean±standard deviation over five training seeds.

Method	Set-L1↓	Centroid-L1↓	Ord. Acc.↑	Pair Acc.↑	J_{lang}↑	$\mathrm{Reg}_{\mathrm{lang}}$↓
Per-Domain	.1166±.0028	.1486±.0018	**.9261±.0038**	.9784±.0011	.6594±.0008	.0129±.0008
Ours (Shared)	**.1136±.0050**	**.1418±.0049**	.9211±.0072	**.9796±.0008**	**.6602±.0008**	**.0122±.0008**

D.3 Semantic and Behavioral Performance Map

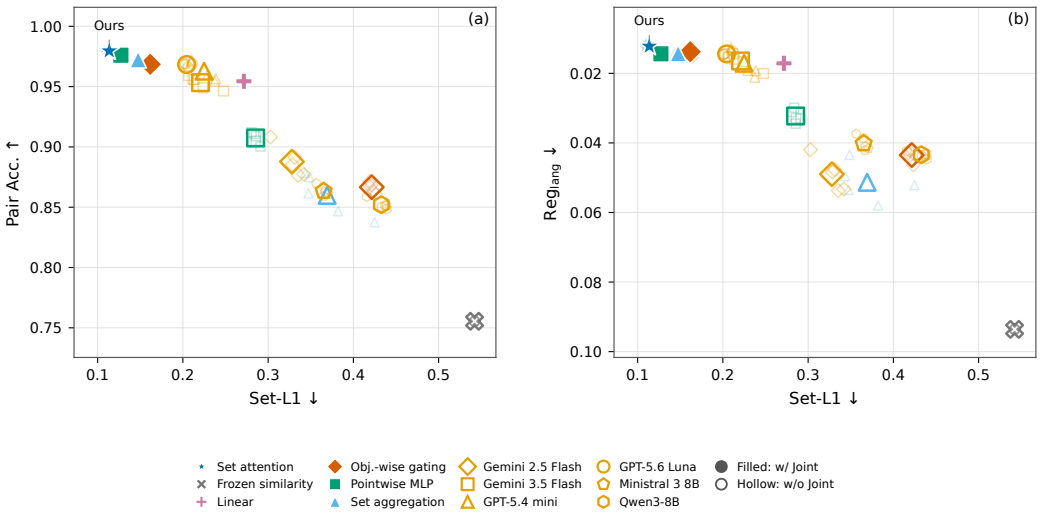


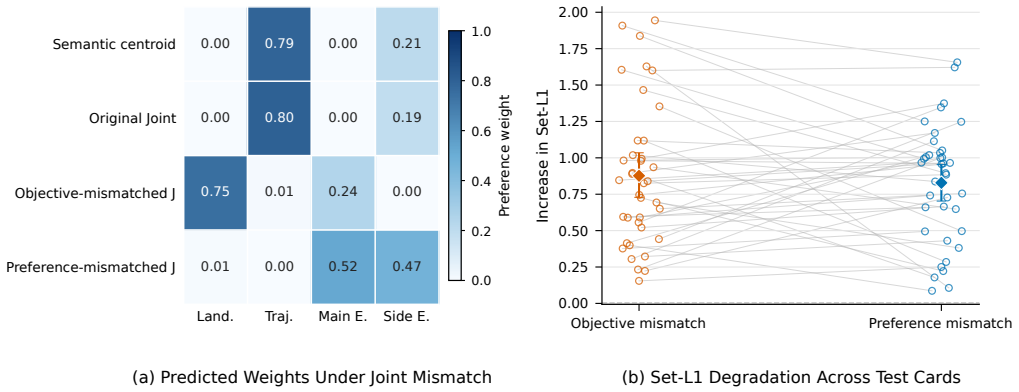
Figure 5: Semantic grounding and behavioral relevance across learned predictors and direct LLMs. Set-L1 is compared with (a) Pair Acc. and (b) $\mathrm{Reg}_{\mathrm{lang}}$. Small translucent markers denote individual training seeds or direct-LLM inference runs, and larger markers denote the corresponding means.

Figure 5 shows that improved semantic grounding generally translates into better behavioral relevance across the learned predictors. Within each neural predictor family, Joint representations consistently shift performance toward lower Set-L1, higher Pair Acc., and lower $\mathrm{Reg}_{\mathrm{lang}}$. LANGUAGE2UTILITY achieves the lowest Set-L1 together with the highest Pair Acc. and lowest $\mathrm{Reg}_{\mathrm{lang}}$ among the compared methods. In contrast, the direct LLM predictors can attain relatively high pairwise accuracy or low regret while remaining substantially farther from the admissible preference set, highlighting a distinction between behavioral agreement and faithful recovery of set-valued preference semantics.

D.4 Mechanistic Diagnostics

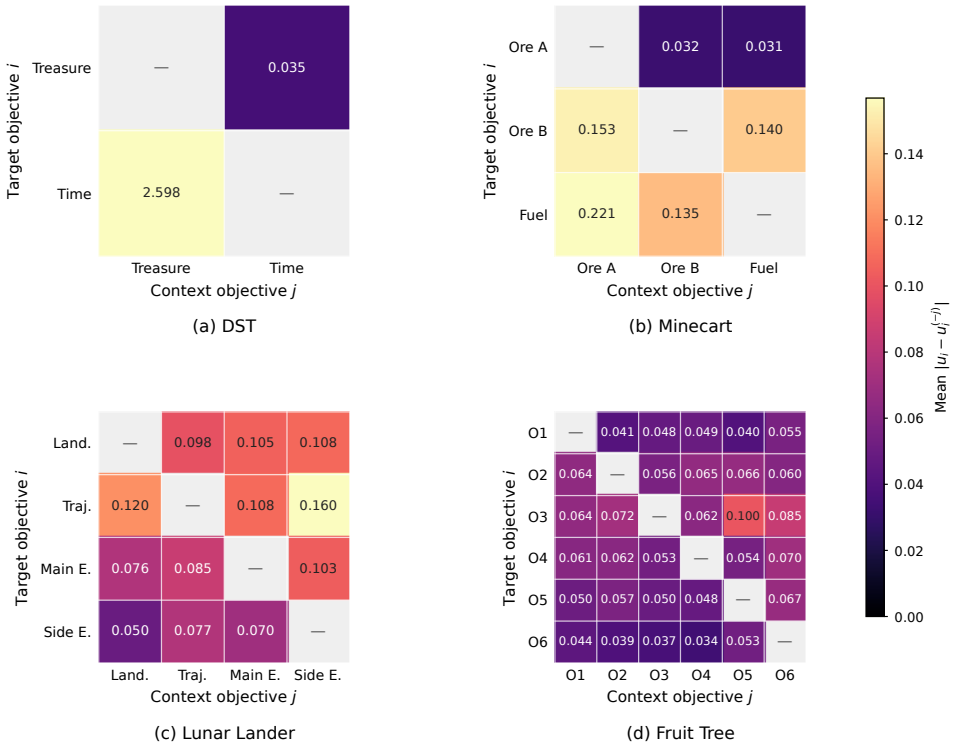
The following inference-time interventions examine whether the relational and set-contextual components identified by the matched ablations are actively used by the trained model.

With correctly aligned Joint representations, the representative prediction in Figure 6(a) nearly recovers the semantic centroid. Misaligning the Joint representation along either the objective or preference axis instead redistributes substantial mass across objectives. Figure 6(b) shows a positive increase in Set-L1 for every test card under both interventions. The trained model therefore relies systematically on correctly aligned preference–objective Joint representations, supporting their role as objective-specific relational information rather than merely additional feature capacity.



(a) Predicted Weights Under Joint Mismatch (b) Set-L1 Degradation Across Test Cards

Figure 6: Effect of Joint representation mismatch on preference grounding. (a) Semantic centroid and predicted preference weights for a representative Lunar Lander card under correctly aligned, objective-mismatched, and preference-mismatched Joint representations. (b) Increase in Set-L1 relative to the correctly aligned Joint representation across canonical test cards under objective and preference mismatch. Points denote five-seed means for individual cards, gray lines connect the same card across interventions, and diamonds show the mean with 95% card-bootstrap confidence intervals.



(a) DST (b) Minecart

(c) Lunar Lander (d) Fruit Tree

Figure 7: Cross-objective influence in set attention. Each cell reports the mean absolute change in the pre-softmax utility logit of target objective i when the attention connection from target i to context objective j is masked. Off-diagonal entries quantify how one objective's prediction depends on other objectives in the set. Magnitudes should be interpreted within each domain because logit scales differ across domains. For Fruit Tree, O1 through O6 denote Protein, Carbohydrates, Fats, Vitamins, Minerals, and Water, respectively.

The nonzero off-diagonal entries in Figure 7 show that an objective's pre-softmax score depends on other objectives through set attention, rather than interacting with them only through softmax normalization. The influence patterns are asymmetric and objective dependent, indicating structured cross-objective contextualization rather than a uniform set-level effect. Together with the consistent ablation effect of removing objective interaction, these results support cross-objective context as a complementary source of grounding accuracy.

D.5 Set-Valued Grounding Geometry

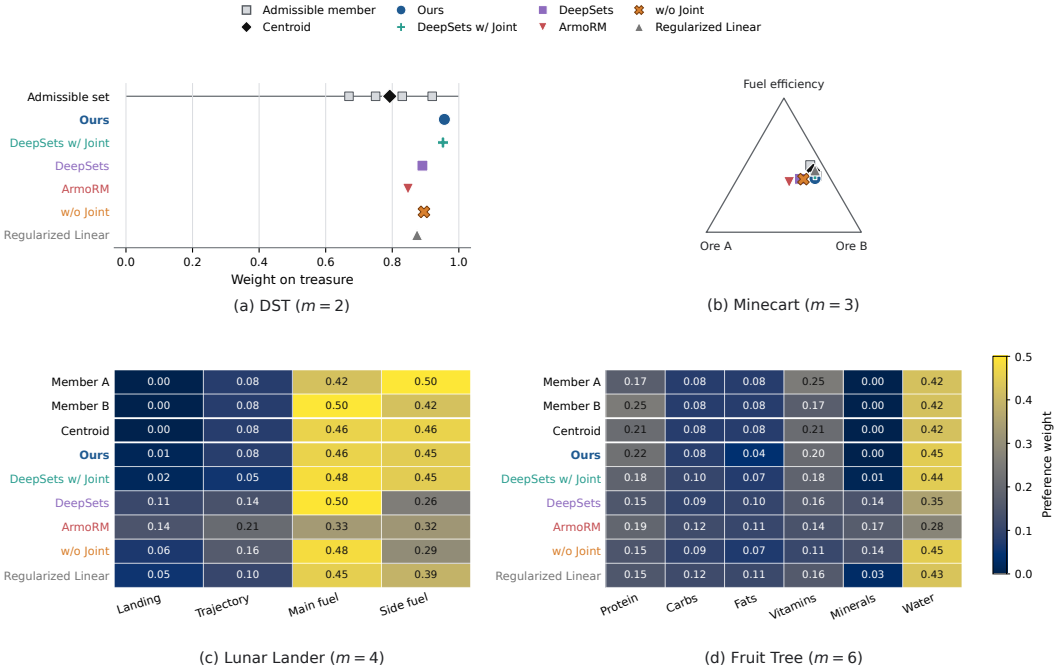


Figure 8: Set-valued preference grounding across the four domains. For one selected test card per domain, each panel compares the admissible preference geometry with predictions from LANGUAGE2UTILITY, DeepSets w/ Joint, DeepSets, ArmoRM, the no-Joint ablation, and Regularized Linear. DST and Minecart are visualized in one-dimensional and ternary simplex coordinates, respectively, while Lunar Lander and Fruit Tree show coordinate-wise preference weights. For the two higher-dimensional domains, two admissible members are shown for illustration, while the centroid is computed over the full admissible set.

Figure 8 reveals a clear hierarchy in preference-space recovery. On DST, the single test card induces a one-dimensional preference geometry, and all compared methods remain close to the admissible interval, providing a low-dimensional sanity check. Clearer differences emerge as the objective geometry becomes richer. Across these domains, LANGUAGE2UTILITY and DeepSets w/ Joint remain most closely aligned with the admissible preference geometry, with LANGUAGE2UTILITY showing particularly faithful coordinate-wise allocations in the higher-dimensional Lunar Lander and Fruit Tree examples. The matched DeepSets comparison makes the contribution of relational information especially visible: adding Joint representations moves the predictions substantially toward the semantic target, while the corresponding no-Joint models exhibit larger distortions. In Fruit Tree, for example, DeepSets and ArmoRM assign substantial mass to the semantically negligible Minerals coordinate, whereas LANGUAGE2UTILITY assigns zero mass and DeepSets w/ Joint nearly suppresses it. A similar separation appears in Lunar Lander, where LANGUAGE2UTILITY closely tracks the centroid allocation while the non-Joint predictors redistribute considerably more mass across objectives. Together with the aggregate results, these examples show that Joint preference-objective representations account for the largest improvement in semantic geometry, while the proposed set-attention realization provides the most consistent recovery within the strong Joint-based class.

D.6 Paired Per-Card Effects

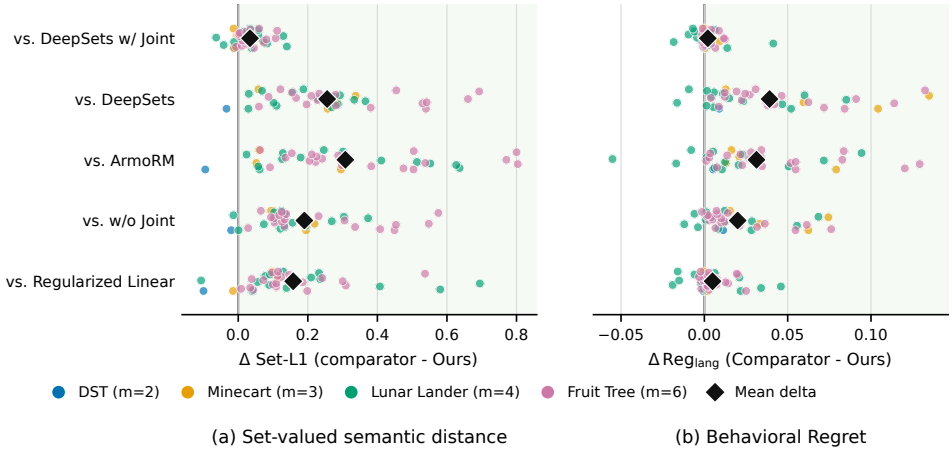


Figure 9: Paired card-level differences relative to LANGUAGE2UTILITY. Each point reports comparator minus LANGUAGE2UTILITY on the same canonical test card for (a) Set-L1 and (b) $\mathrm{Reg}_{\mathrm{lang}}$. Positive values favor LANGUAGE2UTILITY. Colors denote domains, and black diamonds indicate the mean difference across test cards.

Figure 9 shows that LANGUAGE2UTILITY is favorable on average against every representative comparator on both semantic distance and behavioral regret. The advantage is particularly pronounced for Set-L1, where DeepSets, ArmoRM, the no-Joint ablation, and Regularized Linear exhibit broadly positive paired differences across test cards. DeepSets w/ Joint provides the closest architectural counterpart, retaining both Joint preference-objective representations and set-level objective interaction. Although its paired differences are consequently smaller, LANGUAGE2UTILITY reduces aggregate Set-L1 by 23.0% relative to DeepSets w/ Joint (.1475 to .1136), while also lowering $\mathrm{Reg}_{\mathrm{lang}}$ from .0143 to .0122 in Table 1. In contrast, removing Joint from DeepSets produces a much larger semantic gap, highlighting the contribution of relational preference-objective information. The $\mathrm{Reg}_{\mathrm{lang}}$ differences are smaller than the semantic-distance differences but remain positive on average for all displayed comparators. Together, the paired and aggregate results show that LANGUAGE2UTILITY provides more consistent set-valued semantic grounding while retaining an overall behavioral advantage after realization by the frozen controllers.

E Additional LODO Analyses

This section resolves the merged LODO result by held-out domain and examines the geometry, representation structure, and behavioral response of the transferred grounding model.

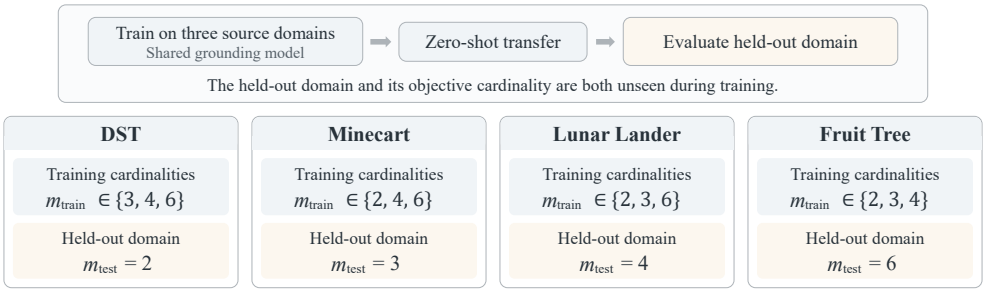


Figure 10: Leave-one-domain-out protocol. Each fold trains the grounding model on three source domains and evaluates zero-shot on the held-out domain. Since each benchmark has a unique objective cardinality, the held-out domain also removes its cardinality from training, producing a simultaneous domain and cardinality shift.

E.1 Per-Domain Transfer Results

Table 16: Per-domain LODO results across all evaluation metrics. Results are reported separately for each held-out domain.

Held-out domain	Method	Set-L1 ↓	Centroid-L1 ↓	Ord. Acc. ↑	Pair Acc. ↑	J_{lang} ↑	Reg_{lang} ↓
DST $m = 2$	*Partial variants and baseline*						
	w/o Objective Interaction	.0472±.0106	.2889±.0175	**1.0000±.0000**	.8947±.0369	.7227±.0069	.0425±.0069
	w/o Joint	.4070±.1225	.6532±.1320	.2250±.2404	.2691±.1485	.5791±.0597	.1860±.0597
	DeepSets	.2457±.0941	.4827±.1116	.5000±.2932	.5250±.1656	.6460±.0382	.1191±.0382
	Joint + set interaction (relational set design)						
	DeepSets w/ Joint	**.0424±.0095**	.1119±.0169	**1.0000±.0000**	**1.0000±.0000**	.7618±.0047	.0033±.0047
	Ours	.0437±.0090	**.0850±.0194**	**1.0000±.0000**	**1.0000±.0000**	**.7636±.0020**	**.0015±.0020**
Minecart $m = 3$	*Partial variants and baseline*						
	w/o Objective Interaction	.3585±.0252	.3983±.0311	.7267±.0384	.8692±.0144	.6195±.0025	.0230±.0025
	w/o Joint	.5456±.1200	.5881±.0994	.2683±.2084	.6313±.0101	.4541±.0549	.1883±.0549
	DeepSets	.6569±.2714	.7009±.2496	.2567±.1204	.6285±.0186	.4037±.1459	.2388±.1459
	Joint + set interaction (relational set design)						
	DeepSets w/ Joint	**.1539±.0205**	**.1818±.0163**	**.8933±.0279**	**.9744±.0117**	**.6318±.0020**	**.0106±.0020**
	Ours	.1930±.0350	.2339±.0299	.8733±.0522	.9497±.0110	.6276±.0027	.0148±.0027
Lunar Lander $m = 4$	*Partial variants and baseline*						
	w/o Objective Interaction	.3869±.0106	.4143±.0111	.6835±.0166	.8897±.0093	.7780±.0060	.0647±.0060
	w/o Joint	.6764±.0205	.7007±.0207	.3319±.0635	.7612±.0223	.7435±.0217	.0992±.0217
	DeepSets	.6737±.0344	.6980±.0352	.2509±.1152	.7611±.0229	.7555±.0267	.0873±.0267
	Joint + set interaction (relational set design)						
	DeepSets w/ Joint	.4176±.0240	.4548±.0311	.6532±.0091	.8455±.0131	.7631±.0078	.0796±.0078
	Ours	**.3723±.0201**	**.4066±.0193**	**.7185±.0273**	**.8995±.0100**	**.7823±.0088**	**.0604±.0088**
Fruit Tree $m = 6$	*Partial variants and baseline*						
	w/o Objective Interaction	.3969±.0622	.4061±.0631	**.8390±.0087**	.8813±.0297	.5130±.0107	.0352±.0107
	w/o Joint	.6764±.0062	.6963±.0067	.3160±.0503	.6537±.0071	.4384±.0060	.1099±.0060
	DeepSets	.6760±.0044	.6968±.0065	.2855±.0612	.6442±.0236	.4348±.0108	.1135±.0108
	Joint + set interaction (relational set design)						
	DeepSets w/ Joint	.3134±.0240	.3329±.0218	.8234±.0131	.8980±.0124	.5253±.0043	.0230±.0043
	Ours	**.2818±.0225**	**.3035±.0210**	.8356±.0124	**.9193±.0122**	**.5274±.0032**	**.0208±.0032**

Results are reported as mean±standard deviation over five seeds. Best and second-best values within each held-out domain are bolded and underlined, respectively.

Across held-out domains, LANGUAGE2UTILITY achieves better or comparable performance on reported metrics. Notably, the best per-domain results are concentrated among the two methods that retain both Joint preference-objective representations and set-level objective interaction, namely LANGUAGE2UTILITY and DeepSets w/ Joint. DeepSets w/ Joint provides a matched set-based counterpart that preserves these two ingredients while replacing the proposed set-attention contextualization with DeepSets. Overall, LANGUAGE2UTILITY attains the best result on all six metrics in Table 3, supporting the benefit of the proposed contextualization architecture. The ablations clarify the complementary roles of the two components. Removing either Joint representations or objective interaction consistently degrades transfer performance across domains. The larger effect of

removing Joint highlights the importance of transferable preference-objective relations, while the consistent degradation without objective interaction confirms that set-level contextualization provides an additional contribution to both grounding accuracy and behavioral relevance.

E.2 In-Domain versus LODO Grounding

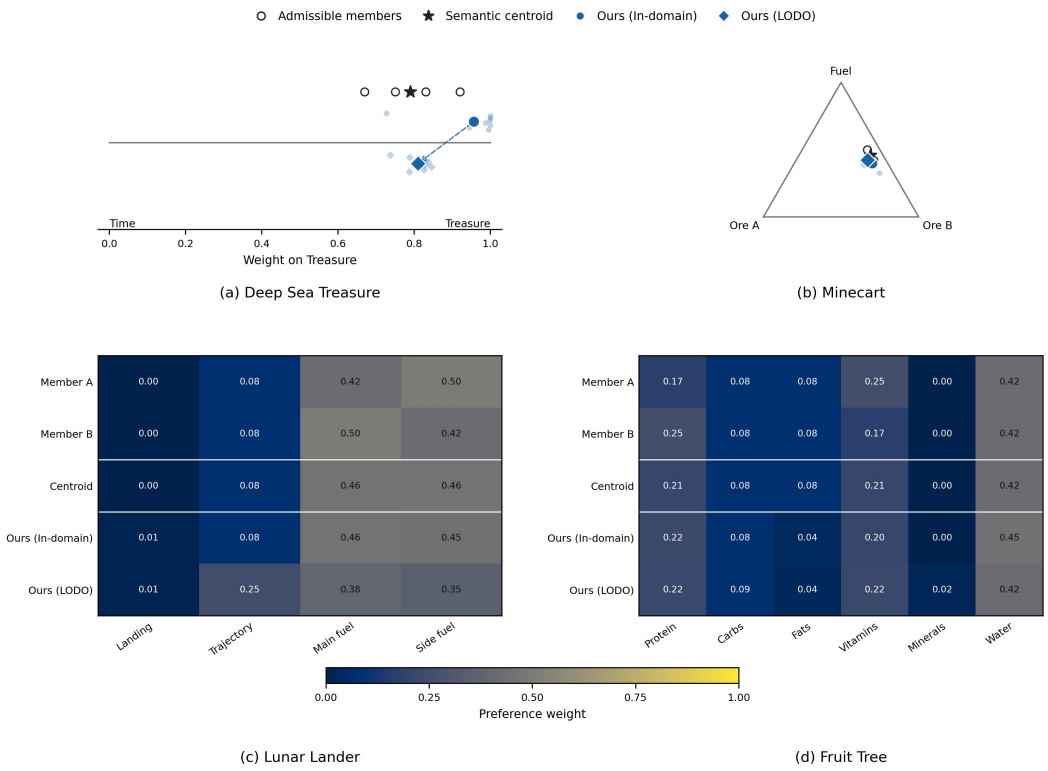


Figure 11: Preference-weight grounding under in-domain and LODO evaluation. For one representative card from each domain, admissible preference weights and their semantic centroid are compared with predictions from Ours in the in-domain and LODO settings. For Lunar Lander and Fruit Tree, two admissible members are shown for illustration, while the centroid is computed over the full admissible set.

The representative cards in Figure 11 show that transfer does not induce a uniform geometric shift. Minecart and Fruit Tree retain similar relative allocations under in-domain and LODO training, while Lunar Lander exhibits a larger redistribution toward trajectory quality. On the displayed DST card, the LODO prediction moves toward the admissible set rather than away from it. These examples complement the per-domain table by showing that transfer difficulty reflects domain-specific changes in preference geometry, not objective cardinality alone.

E.3 Cross-Domain Representation Analysis

Linear-probe analysis for Figure 3. For each held-out domain, we compare two dimension-matched frozen representations: the concatenation of independently L2-normalized preference and objective embeddings, and the L2-normalized Joint preference-objective representation. We fit a ridge-regularized linear probe on the source domains to recover preference-relative objective importance and evaluate it directly on the held-out domain. For the concatenated representation, the linear predictor decomposes into separate preference and objective terms and therefore cannot introduce preference-objective interactions, whereas the Joint representation encodes the preference-objective pair jointly before probing.

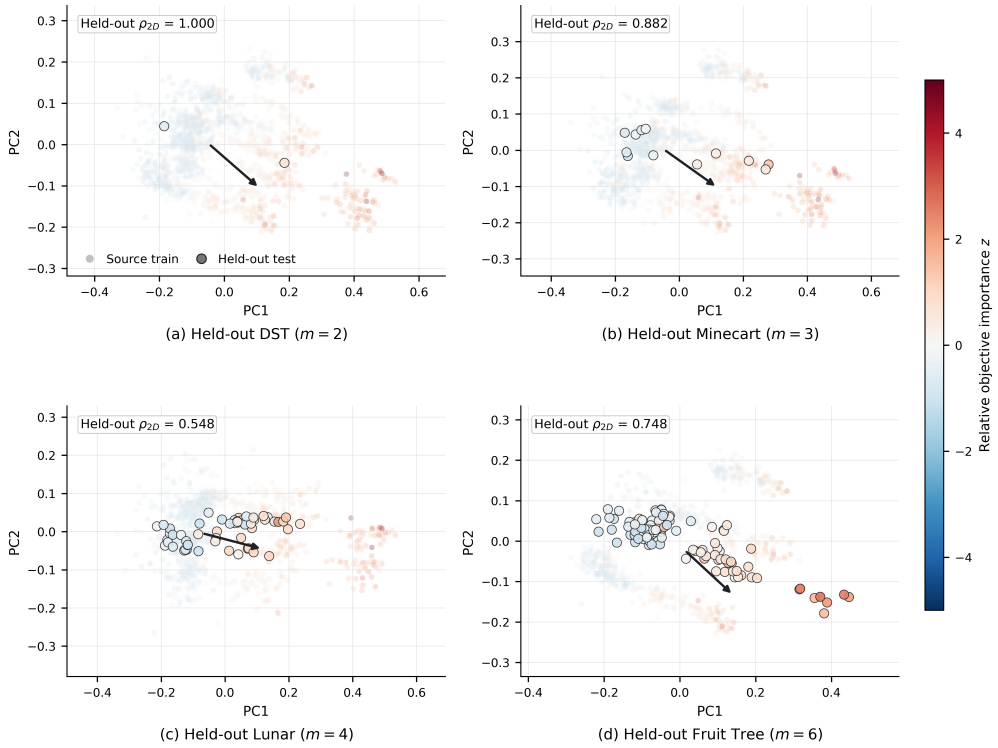


Figure 12: Cross-domain organization of Joint preference-objective representations. For each held-out domain, PCA is fitted on card-centered Joint representations from the remaining source domains, and held-out representations are projected into the same space. Color indicates relative objective importance, and arrows denote the preference-related direction estimated from the source domains. The reported ρ_{2D} measures the alignment of held-out representations with this ordering in the projected space.

Geometric organization of Joint representations. Figure 12 provides a geometric counterpart to the linear probe analysis in the main text. Across all held-out domains, Joint representations retain a consistent positive organization along the preference-related direction estimated solely from source domains, with particularly clear structure for Minecart and Fruit Tree. Their alignment within a common representation space fitted only on source domains provides further evidence that Joint representations capture transferable relations between language preferences and objective importance, rather than relying primarily on domain-specific objective identities.

E.4 Domain-Specific Control Realization

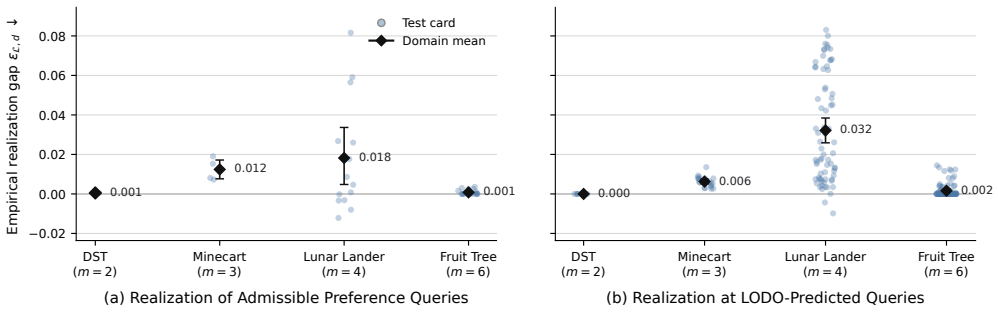


Figure 13: Domain-specific control realization across held-out domains. (a) Empirical realization gaps for admissible preference queries. (b) Realization gaps at LODO-predicted preference queries. Points denote canonical test cards, diamonds denote domain means, and error bars show 95% card-bootstrap confidence intervals where defined. Lower is better.

While Sec. 4.3.2 examines cross-domain transfer at the representation level, Figure 13 isolates the domain-specific realization behavior of the frozen controllers. When the frozen controllers are queried directly with admissible preference weights, realization gaps are small for DST and Fruit Tree, modest for Minecart, and largest for Lunar Lander, with domain means of .001, .012, .018, and .001, respectively. Thus, the six-objective Fruit Tree domain is not intrinsically harder to realize than the lower-cardinality domains.

The same comparison at LODO-predicted queries further accentuates the domain difference. The mean realization gap for Lunar Lander increases to .032, while DST, Minecart, and Fruit Tree remain at .000, .006, and .002. Together with the representation-level results in Fig. 3, where Lunar Lander retains comparable Joint decodability to Fruit Tree, this suggests that Lunar's larger downstream difficulty is partly associated with domain-specific control realization rather than semantic transfer or objective cardinality alone.

E.5 Controlled Preference Responses

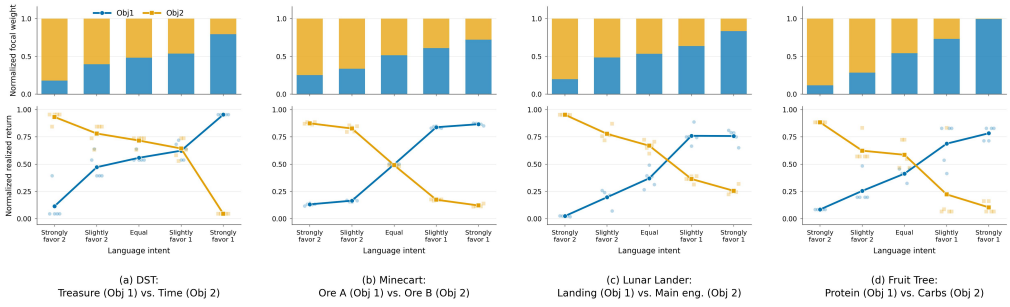


Figure 14: Controlled preference responses across the four held-out domains. For each domain, language intent varies from strongly favoring the second focal objective to strongly favoring the first. The upper row shows the predicted weights normalized over the two focal objectives, and the lower row shows the corresponding normalized realized returns. Faint points denote individual seeds and solid lines denote five-seed means.

The upper row of Figure 14 exhibits a monotone shift in focal weight allocation for all four held-out domains as the language intent moves from the second objective to the first. The lower row translates this semantic response into systematic return trade-offs under the frozen controllers: the favored objective generally increases while the competing objective decreases. The realized curves need not reproduce the weight curves exactly because behavioral realization is domain specific and nonlinear.

Their consistent directional response nevertheless shows that zero-shot changes in language are reflected in both predicted preference weights and realized outcomes.

E.6 Paired LODO Effects by Held-Out Domain

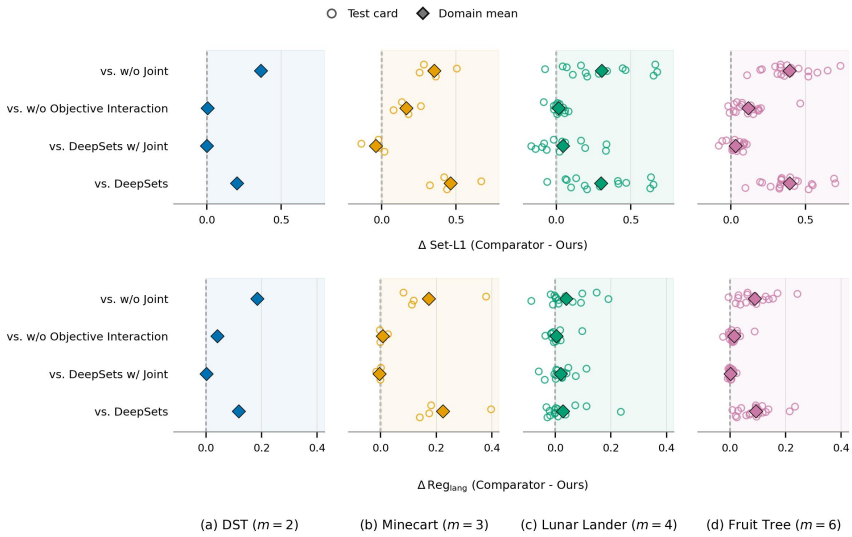


Figure 15: Paired card-level differences across held-out domains. Points compare Ours with each comparator on the same test card for Set-L1 (top) and $\mathrm{Reg}_{\mathrm{lang}}$ (bottom). Positive values favor Ours; diamonds denote five-seed domain means.

Figure 15 shows consistent per-card degradation across held-out domains when either component is removed. The no-Joint ablation produces the larger shift, highlighting the importance of preference-objective relational information, while the persistent degradation without objective interaction supports an independent contribution from set-level contextualization. These results support complementary roles for preference-objective relational information and set-level contextualization in zero-shot transfer. DeepSets w/ Joint provides the closest architectural counterpart, retaining both Joint representations and set-level objective interaction while replacing the proposed set-attention contextualization with DeepSets. This comparison therefore largely isolates the contextualization architecture, with the remaining paired advantage of LANGUAGE2UTILITY supporting the benefit of the proposed set-attention design. These per-card effects complement the per-domain analysis and provide additional evidence for the full LANGUAGE2UTILITY design in Table 3.

F HUMAN-AUTHORED TRANSFER EVALUATION

F.1 APPLICATION SCENARIOS AND HUMAN PREFERENCES

We consider three application scenarios motivated by recent work: preference-adaptable human-aware robot navigation (de Heuvel et al., 2025), personalized object-goal navigation with language-inferred preference weights (Hwang et al., 2024), and autonomous-driving planning with learned scene evaluation (Huang et al., 2025). These scenarios introduce new task contexts and objective sets with cardinalities $m = 4, 5, 6$, respectively. Five participants each write five natural-language preferences per scenario after seeing only the task and objective descriptions, yielding 25 preferences per scenario and 75 in total. Table 17 specifies the task contexts and objectives used in the evaluation.

Table 17: Task contexts and objective sets used in the human-authored transfer evaluation.

Objective	Definition
Human-Aware Robot Navigation ($m = 4$)	
Task context: Control a mobile service robot in an indoor environment shared with people and obstacles, and navigate from its starting position to a specified destination.	
Collision Safety	Prioritize avoiding collisions with people, walls, furniture, and other obstacles.
Personal Space	Prioritize maintaining a comfortable distance from nearby people during navigation.
Time Efficiency	Prioritize reaching the destination quickly with minimal unnecessary delay.
Energy Efficiency	Prioritize reaching the destination while minimizing total energy consumption.
Household Object-Goal Navigation ($m = 5$)	
Task context: Control a household robot in an unfamiliar home and search the environment until it finds and reaches a requested object.	
Time Efficiency	Prioritize finding and reaching the requested object with minimal delay.
Path Efficiency	Prioritize finding and reaching the requested object by a short, direct route.
House Exploration	Prioritize covering more of the home while searching for the requested object.
Object Exploration	Prioritize observing more distinct household objects while searching for the requested object.
Safety	Prioritize avoiding collisions and unsafe interactions while searching for the requested object.
Urban Autonomous Driving ($m = 6$)	
Task context: Control an autonomous vehicle through busy urban traffic and navigate from its starting point to a specified destination while interacting with other road users.	
Collision Safety	Prioritize avoiding collisions with vehicles, pedestrians, cyclists, and obstacles.
Travel Efficiency	Prioritize reaching the destination promptly with minimal unnecessary delay.
Passenger Comfort	Prioritize keeping the ride smooth with gentle acceleration, braking, and turning.
Rule Compliance	Prioritize following traffic signals, speed limits, lane rules, and right-of-way.
Energy Efficiency	Prioritize completing the trip with minimal fuel or battery consumption.
Traffic Cooperation	Prioritize coordinating smoothly with other road users through predictable yielding, merging, and gap selection.

Two annotators independently assign semantic roles and explicit pairwise relations to each preference, then reconcile disagreements to form a consensus annotation. The preference author verifies this consensus, with any remaining disagreement resolved by a third adjudicator. The final annotations define $\mathcal{W}(c)$ and $\mathcal{C}_c$ using the same semantic construction as the main benchmark, and are evaluated with the metrics in Appendix C.1.

F.2 PER-SCENARIO RESULTS

Table 18 reports results for each application scenario. LANGUAGE2UTILITY achieves the lowest Set-L1 in all three scenarios and the best or comparable performance on Centroid-L1 and Ord. Acc., demonstrating consistent transfer across different task contexts and objective sets. The 8B-scale direct LLMs remain competitive in ordinal interpretation, with Ministral 3 8B attaining the highest Ord. Acc. in human-aware navigation, but are consistently farther from the admissible preference geometry. Across the three scenarios, their grounding performance also degrades as the objective

cardinality increases, whereas LANGUAGE2UTILITY does not show the same monotonic deterioration in preference-geometry recovery. Notably, this robustness extends to the five-objective household setting, whose cardinality is absent from the training benchmark.

Table 18: Semantic grounding on human-authored preferences across three application scenarios. Learned methods report mean±standard deviation over five training seeds, while direct LLMs report results over five independent inference runs. Best and second-best means within each scenario are bolded and underlined, respectively.

Method	Set-L1↓	Centroid-L1↓	Ord. Acc.↑
Human-Aware Robot Navigation ($m = 4$)			
w/o Objective Interaction	$.4084 \pm .0498$	$.4163 \pm .0501$	$.7819 \pm .0640$
w/o Joint	$.7661 \pm .2735$	$.7787 \pm .2727$	$.4906 \pm .1184$
DeepSets	$.6362 \pm .1147$	$.6499 \pm .1153$	$.4582 \pm .0692$
Qwen3-8B	$\underline{.2153} \pm .0029$	$\underline{.2252} \pm .0047$	$.8597 \pm .0000$
Ministral 3 8B	$.2224 \pm .0038$	$.2275 \pm .0037$	$\mathbf{.9389} \pm .0124$
DeepSets w/ Joint	$.2276 \pm .0398$	$.2360 \pm .0406$	$.8825 \pm .0344$
Ours	$\mathbf{.1856} \pm .0087$	$\mathbf{.1872} \pm .0082$	$\underline{.9063} \pm .0147$
Household Object-Goal Navigation ($m = 5$)			
w/o Objective Interaction	$.4195 \pm .0474$	$.4311 \pm .0473$	$.7030 \pm .0518$
w/o Joint	$.5658 \pm .0732$	$.5851 \pm .0732$	$.5624 \pm .0667$
DeepSets	$.5364 \pm .0264$	$.5548 \pm .0268$	$.5080 \pm .0953$
Qwen3-8B	$.2709 \pm .0010$	$.2754 \pm .0010$	$.7614 \pm .0000$
Ministral 3 8B	$.2604 \pm .0216$	$.2643 \pm .0213$	$\underline{.8439} \pm .0302$
DeepSets w/ Joint	$\underline{.2434} \pm .0133$	$\mathbf{.2539} \pm .0162$	$.8071 \pm .0211$
Ours	$\mathbf{.2287} \pm .0163$	$\underline{.2547} \pm .0165$	$\mathbf{.8768} \pm .0247$
Urban Autonomous Driving ($m = 6$)			
w/o Objective Interaction	$.3295 \pm .0101$	$.3301 \pm .0098$	$.7887 \pm .0155$
w/o Joint	$.6887 \pm .2511$	$.6971 \pm .2525$	$.4069 \pm .0778$
DeepSets	$.5947 \pm .1214$	$.6013 \pm .1236$	$.3675 \pm .0940$
Qwen3-8B	$.2799 \pm .0058$	$.2819 \pm .0058$	$.7372 \pm .0032$
Ministral 3 8B	$.2815 \pm .0065$	$.2830 \pm .0054$	$.6910 \pm .0106$
DeepSets w/ Joint	$\underline{.2271} \pm .0411$	$\underline{.2281} \pm .0413$	$\mathbf{.8301} \pm .0141$
Ours	$\mathbf{.2140} \pm .0224$	$\mathbf{.2172} \pm .0249$	$\underline{.7996} \pm .0274$