SYMSELECT: RECOVERING SYMMETRY ALIGNMENT WHEN VALIDITY DEPENDS ON THE SAMPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Candidate transformations can provide useful structural relations even when their validity varies across samples. When pair validity is unobserved, aligning every candidate can impose invalid constraints, while discarding transformed pairs loses the benefit carried by valid ones. Existing adaptive invariance methods can vary how transformations are used, but they do not determine how much of the benefit of valid pair alignment remains recoverable when validity labels are unavailable. We study this question in controlled settings where application validity can be hidden during training but recovered exactly for evaluation, and ask what a selected alignment update must preserve to retain the oracle improvement. We propose SymSelect, which selects candidate relations by representation similarity and aligns only the most similar pairs. On MNS and I-RAVEN, SymSelect recovers 77%–92% of the accuracy gain from supervised valid pair alignment without observing pair validity. Our analysis shows that selecting mostly valid pairs is not sufficient to inherit the oracle benefit: the retained valid subset must also preserve a useful alignment update. Local and finite batch results separate invalid pair contamination from retained-valid bias and characterize the effects of retention and contribution variability. Matched controls show that aggregate validity enrichment alone does not explain SymSelect's recovery, while checkpoint measurements along one I-RAVEN trajectory show higher batch precision, lower selected-valid gradient discrepancy, and positive local alignment effects.

1 INTRODUCTION

Structural transformations can provide useful supervision when they preserve the relation required by a task. Their validity, however, need not be uniform across samples. A candidate transformation family can therefore contain useful alignment relations without identifying which individual applications should be aligned. When this validity is unobserved, enforcing every candidate mixes valid and invalid relations, while discarding transformed pairs gives up potentially useful structural supervision. The resulting question is whether the benefit of aligning valid pairs can still be recovered without observing which pairs are valid. We study this question in controlled structured reasoning settings where pair validity can be hidden during training and recovered exactly for evaluation.

Prior work has already made invariance adaptive. Input specific augmentation learns transformation policies that vary across samples, while partial and conditional equivariance allow the symmetry structure used by a model to depend on the input or target setting (Miao et al., 2023; Kim et al., 2024; Chaudhuri et al., 2024). Work on selective and reweighted training likewise shows that different training signals can have different learning value (Lee, 2013; Han et al., 2018; Ren et al., 2018; Killamsetty et al., 2021). These approaches establish that both transformation use and training signal can adapt to the data. They do not, however, characterize recovery relative to an oracle that knows which candidate relations are valid. Moreover, selecting mostly valid pairs need not preserve the update produced by the full valid population.

To address this gap, we use supervised alignment over valid pairs as an exact validity reference and introduce SymSelect, which selects candidate relations by representation similarity without observing their validity. We then ask what a selected alignment update must preserve relative to this oracle. Our analysis shows that high validity precision alone is insufficient because the retained valid subset can induce a different update from the full valid population. On MNS and I-RAVEN, SymSelect recovers

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.

77%–92% of the supervised alignment gain, while matched controls and checkpoint diagnostics support the distinction between validity enrichment and preservation of the selected update.

Our main contributions are summarized below.

- We formulate symmetry alignment with hidden pair validity and introduce SymSelect, which uses representation similarity to select candidate relations for alignment without observing which transformation applications preserve the label.
- We show that selecting mostly valid pairs is not sufficient to inherit the benefit of valid pair alignment. Our analysis separates invalid pair contamination from retained-valid bias, while a finite batch bound characterizes how retention and contribution variability affect the selected alignment update.
- Without observing pair validity, SymSelect recovers 77%–92% of the accuracy gain from supervised valid pair alignment on MNS and I-RAVEN across the evaluated backbones.
- Matched controls show that validity enrichment alone does not explain SymSelect's recovery, while checkpoint measurements show higher batch precision, lower selected-valid gradient discrepancy, and positive selected alignment steps along the evaluated I-RAVEN trajectory.

2 Related Work

Classical invariant and equivariant models exploit transformations that are assumed to preserve task semantics (LeCun & Bengio, 1995; Jaderberg et al., 2015; Worrall et al., 2017; Marcos et al., 2017; Cohen & Welling, 2016), with theory characterizing their generalization benefits under globally valid symmetries (Zhu et al., 2021; Elesedy & Zaidi, 2021; Sannai et al., 2021; Soleymani et al., 2025). Data augmentation need not be uniformly beneficial across the data, and theory can require distinguishing original from transformed examples when transformation assumptions are only partially valid (Balestriero et al., 2022; Shao et al., 2022). Recent analyses further ask when partial augmentation retains the statistical benefits of full group augmentation and when distributional symmetry breaking changes the value of symmetry-aware learning (Tahmasebi et al., 2026; Lawrence et al., 2026).

Other work learns the symmetry structure itself or how much transformation invariance to impose (Desai et al., 2022; Yang et al., 2023a; Zhou et al., 2021; Schwöbel et al., 2022; Tombs & Lester, 2022; Moskalev et al., 2023). Augerino and related approaches learn augmentation distributions or invariance parameters, while differentiable Laplace objectives, invariance-constrained augmentation, and instance-wise transformation predictors adapt transformation use from data (Benton et al., 2020; Immer et al., 2022; Hounie et al., 2023; Singhal et al., 2023). InstaAug learns input specific augmentation distributions, VP G-CNN adapts partial equivariance across samples, and conditional invariance adapts the invariance objective to the target domain (Miao et al., 2023; Kim et al., 2024; Chaudhuri et al., 2024). These methods learn transformations, policies, distributions, or invariance strength. Our question is different: we fix the candidate family and study hidden application validity relative to a supervised valid pair alignment oracle. The goal is to characterize how much of that oracle benefit can be recovered and what a selected alignment update must preserve when individual application validity is unobserved.

Consistency and contrastive objectives align transformed views, while view construction determines which relations are encouraged during learning (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Tian et al., 2020). Imperfect supervision and data selection similarly retain or reweight training signals according to reliability or induced updates (Lee, 2013; Han et al., 2018; Ren et al., 2018; Killamsetty et al., 2021). SymSelect selects an additional relation rather than the original supervised example, and its analysis compares the resulting update with alignment over all valid relations. Unlike example selection, SymSelect selects auxiliary alignment relations and evaluates them relative to the full valid pair oracle. This makes validity enrichment and preservation of the oracle update distinct questions.

MNS, RAVEN, and I-RAVEN provide controlled settings for relational and abstract visual reasoning (Zhang, 2020; Zhang et al., 2019; Hu et al., 2021). Existing methods improve reasoning through relational modules, transformers, probabilistic inference, predictive errors, algebraic representations, and task decomposition (Santoro et al., 2017; Barrett et al., 2018; Dosovitskiy et al., 2021; Zhang

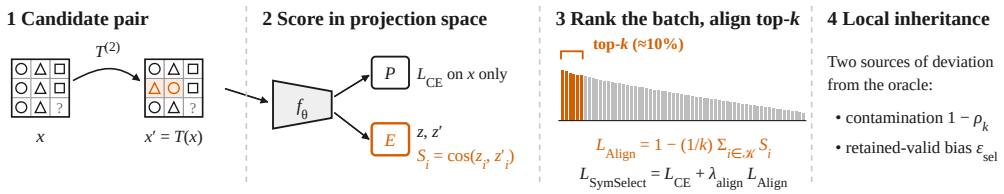

Figure 1: Overview of SymSelect under hidden pair validity. (1) A candidate transformation $T \in \mathcal{T}$ forms a pair $(x, T(x))$, whose validity is not observed by SymSelect during training. (2) The shared backbone feeds a prediction head P, used for classification on the anchor x, and a projection head E, used to score the candidate pair by cosine similarity. (3) Candidate pairs are ranked within each mini batch and only the top k contribute to the alignment loss. (4) The selected alignment update can depart from the valid pair oracle through invalid pair contamination $1 - \rho_k$ and bias among the retained valid pairs ϵ_{sel}.

et al., 2021; Yang et al., 2023b; Zhang et al., 2022; Kwiatkowski & Krawiec, 2024). We use these tasks not to introduce a new reasoning architecture, but because their explicit structural rules allow pair validity to be hidden during training and recovered exactly for evaluation, providing a controlled reference for measuring alignment recovery.

3 Methods

We first define the valid pair alignment available when application validity is observed. We then construct SymSelect from an observable representation signal and ask when its selected update retains the local benefit of the oracle. Figure 1 summarizes this progression from candidate pair construction to selection and local inheritance.

3.1 From Candidate Transformations to a Valid Pair Oracle

To separate a candidate structural relation from a valid alignment constraint, we attach validity to each transformation application. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a structured visual reasoning dataset and $\mathcal{T}$ a candidate transformation family. We define

$$Z(x, T) = \mathbf{1}\{y(T(x)) = y(x)\}. \tag{1}$$

Training observes $\mathcal{T}$ but not Z. If Z were available, the model could align exactly the valid transformed pairs. We use this supervised alignment as the oracle reference. It keeps the candidate family and pairwise alignment objective fixed while revealing which applications are valid.

3.2 Selecting Alignment Relations from Representation Similarity

Without access to Z, SymSelect needs an observable signal for choosing which candidate relations to align. For each anchor x_i in a mini batch of size B, we sample one $T_i \in \mathcal{T}$ and form $x_i' = T_i(x_i)$. A backbone f_θ feeds a classification head P and a separate projection head E. Only the original anchors contribute to the classification objective, while

$$z_i = E(f_\theta(x_i)), \qquad z_i' = E(f_\theta(x_i'))$$

define the alignment space. Thus selection changes the additional relation used for alignment without assigning the anchor label to a transformed input whose validity is unknown.

For each candidate pair, SymSelect computes

$$S_i = \frac{z_i \cdot z_i'}{\|z_i\| \, \|z_i'\|}. \tag{2}$$

Let $\mathcal{K} = \text{TopK}(\{S_i\}_{i=1}^{B}, k)$. The selected alignment loss is

$$L_{\text{Align}} = 1 - \frac{1}{k} \sum_{i \in \mathcal{K}} S_i, \tag{3}$$

and training minimizes

$$L_{\text{SymSelect}} = L_{\text{CE}} + \lambda_{\text{align}} L_{\text{Align}}, \tag{4}$$

where L_{CE} is computed from the original anchors. The top-k indices are fixed for the backward pass, while gradients propagate through the selected similarities. Appendix A.11 gives the equivalent subset objective and finite batch gradient form.

For comparison, supervised valid alignment uses the same candidate sampling rule as SymSelect: each anchor draws exactly one $T_i \in \mathcal{T}$. Let

$$\mathcal{V}_{\mathcal{B}} = \{i : Z(x_i, T_i) = 1\}.$$

Its mini batch alignment loss is

$$L_{\text{oracle},\mathcal{B}} = \begin{cases} \dfrac{1}{|\mathcal{V}_{\mathcal{B}}|} \displaystyle\sum_{i \in \mathcal{V}_{\mathcal{B}}} (1 - S_i), & |\mathcal{V}_{\mathcal{B}}| > 0, \\ 0, & |\mathcal{V}_{\mathcal{B}}| = 0. \end{cases} \tag{5}$$

The classification loss remains the anchor-only L_{CE} in Equation (4). Align all instead averages the same pairwise loss over all B sampled candidates. The oracle, align all, and SymSelect therefore share the candidate distribution and differ only in which sampled relations contribute to alignment.

3.3 When Selected Alignment Inherits the Oracle Improvement

Figure 1 separates two ways in which a selected alignment update can depart from the valid pair oracle. The selected set can contain invalid pairs, and its retained valid pairs can induce a different average update from the full valid population. We first characterize these deviations at the population level. Let A_k denote the marginalized event that a candidate pair is selected by the mini batch top-k rule and let

$$\rho_k = \mathbb{P}(Z = 1 \mid A_k = 1).$$

At a fixed checkpoint θ_0, the top-k assignment is computed once and treated as fixed for the local update below. If the kth and $(k+1)$th scores are strictly separated and the scores vary continuously with the parameters, this assignment is locally constant around θ_0. With this assignment fixed, define

$$L_{\text{oracle}}(\theta) = \mathbb{E}[\ell_{\text{align}} \mid Z = 1], \qquad L_{\text{sel}}(\theta) = \mathbb{E}[\ell_{\text{align}} \mid A_k = 1].$$

Assumption 1 (Oracle alignment margin)**.** *Let $J(\theta)$ be a differentiable diagnostic for which larger values are better. Oracle alignment improves J with positive first order margin*

$$\mu_J = -\langle \nabla_\theta J(\theta), \nabla_\theta L_{\text{oracle}}(\theta) \rangle > 0. \tag{6}$$

Theorem 1 (Local inheritance from oracle alignment)**.** *Assume J has L_J-Lipschitz gradient with $\|\nabla_\theta J(\theta)\| \le H_J$, and individual alignment gradients satisfy $\|\nabla_\theta \ell_{\text{align}}(x, T(x))\| \le G$. Define*

$$\epsilon_{\text{sel}} = \left\| \mathbb{E}[\nabla_\theta \ell_{\text{align}} \mid A_k = 1, Z = 1] - \nabla_\theta L_{\text{oracle}}(\theta) \right\|.$$

Then one selected alignment update $\theta^+ = \theta - \eta \nabla_\theta L_{\text{sel}}(\theta)$ satisfies

$$J(\theta^+) - J(\theta) \ge \eta \big[\mu_J - H_J(\epsilon_{\text{sel}} + 2(1 - \rho_k)G) \big] - \frac{L_J}{2} \eta^2 \|\nabla_\theta L_{\text{sel}}(\theta)\|^2. \tag{7}$$

For sufficiently small η, selected alignment inherits the oracle improvement whenever

$$\mu_J > H_J(\epsilon_{\text{sel}} + 2(1 - \rho_k)G).$$

Theorem 1 characterizes what a selected alignment rule must preserve relative to the valid pair oracle. The condition is selector independent; SymSelect instantiates it through similarity based top-k selection. Invalid pair contamination contributes through $1 - \rho_k$, while ϵ_{sel} measures how far the retained valid pair gradient moves from the full valid pair gradient. High validity precision controls only the first term, so precision alone does not determine whether the oracle improvement is inherited. Appendix A.1 proves the result.

Proposition 1 (Scope of the inheritance guarantee)**.** *Theorem 1 applies to any model parameterization and candidate transformation family with a fixed sampling rule for which its quantities are well defined, provided Assumption 1 and the stated regularity conditions hold. The bound then uses the corresponding quantities for that instance.*

Appendix A.2 proves Proposition 1.

The population result separates these effects in expectation. Actual training selects a finite top-k set, so we next ask how its composition and contribution variability affect the realized first order update. Let

$$a_i = -\langle \nabla_\theta J, \nabla_\theta \ell_i \rangle \tag{8}$$

be the first order contribution of aligning pair i. For a selected set $\mathcal{K}$, define

$$a_{\mathcal{K}} = \frac{1}{k} \sum_{i \in \mathcal{K}} a_i. \tag{9}$$

Suppose a batch contains n_1 valid and n_0 invalid candidates. For $z \in \{0, 1\}$, let $\bar{a}_z$ and σ_z^2 denote the mean and variance of a_i within validity group z.

Theorem 2 (Finite batch bound for the selected alignment update)**.** *Let $\mathcal{K}$ be the top-k set selected by SymSelect, and let r of the selected pairs be invalid. Then*

$$\begin{aligned} a_{\mathcal{K}} \geq \frac{k-r}{k}\bar{a}_1 + \frac{r}{k}\bar{a}_0 \\ - \frac{\sigma_1\sqrt{(k-r)(n_1-k+r)} + \sigma_0\sqrt{r(n_0-r)}}{k}. \end{aligned} \tag{10}$$

In particular, if every selected pair is valid,

$$a_{\mathcal{K}} \geq \bar{a}_1 - \sigma_1\sqrt{\frac{n_1}{k} - 1}. \tag{11}$$

The finite batch bound is a general characterization of any selected subset with the stated composition. The selected invalid count r controls contamination, while the variance terms quantify how a retained subset can depart from its validity group average. Even when $r = 0$, retaining only part of the valid population leaves a deviation that grows with contribution variability and with the fraction omitted. Appendix A.11 gives the proof.

A complementary decomposition makes the distinction from validity enrichment explicit. Let R be a uniformly random k-subset of the same batch, let $\pi = n_1/B$ and $\rho = (k-r)/k$, and let b_z denote the selected mean contribution minus the full validity group mean within class z. Then

$$a_{\mathcal{K}} - \mathbb{E}_R[a_R] = (\rho - \pi)(\bar{a}_1 - \bar{a}_0) + \rho b_1 + (1-\rho)b_0. \tag{12}$$

The first term captures validity enrichment, while the remaining terms capture which pairs are retained within each validity class. Section 4.3 tests this distinction by matching selected validity composition while randomizing pair identity.

Proposition 2 (Perfect validity ranking is insufficient)**.** *There exists a smooth cosine similarity instance in which every valid pair is ranked above every invalid pair, yet top-k selection has a negative first order effect while alignment over the full valid population has a positive one.*

Appendix A.11 gives an explicit construction.

For SymSelect, similarity enters through the ranked set it produces. Let N_{inv} count invalid-valid inversions, where an invalid candidate is ranked above a valid one. If $n_1 \geq k$, selecting r invalid candidates implies

$$N_{\text{inv}} \geq r(n_1 - k + r). \tag{13}$$

Similarity ranking errors therefore constrain the contamination channel, while retention and variation among valid pair contributions remain separate. The experiments examine this distinction through randomization within validity groups and direct measurements of the selected valid gradient and local alignment update. Appendix A.11 gives the corresponding bound on r together with random selection and joint update extensions.

These results isolate the alignment component of one update at a fixed checkpoint. Appendix A.11 gives the corresponding decomposition when the classification gradient is included, while recovery after repeated optimization is evaluated separately in Section 4.

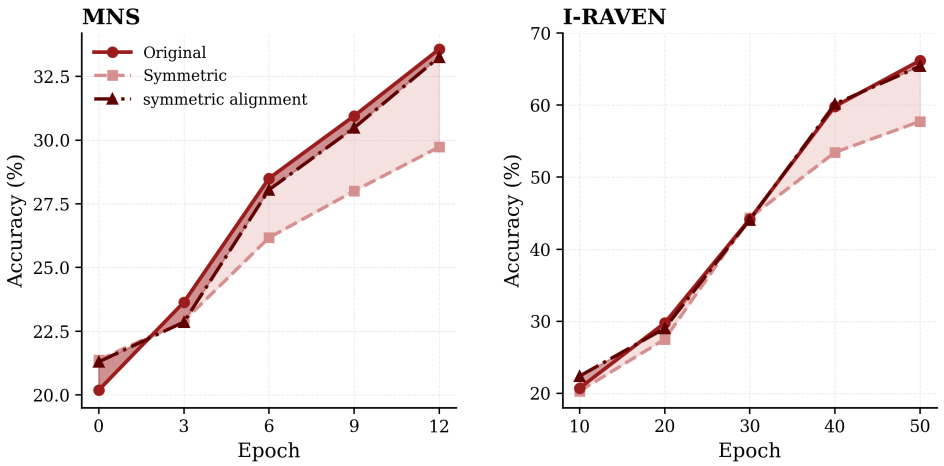

Figure 2: Accuracy on original samples and valid transformed counterparts during training. Base training develops a counterpart gap on MNS and I-RAVEN, while supervised valid pair alignment substantially improves accuracy on valid transformed counterparts.

4 Experiments

We first establish whether supervised valid pair alignment provides a benefit worth recovering and measure how much of it SymSelect retains. We then test whether the identity of the selected relations matters beyond the alignment budget and selected validity composition. Finally, we examine the actual batch selector and the local update quantities identified by the theory.

The main experiments use Machine Number Sense (MNS) (Zhang, 2020) and I-RAVEN (Hu et al., 2021). On MNS, candidate transformations swap operand positions and are valid for commutative operators, giving a validity rate of approximately 30%. On I-RAVEN, $\mathcal{T} = \{T^{(1)}, T^{(2)}, T^{(3)}\}$ contains three row swaps, with approximately 26% of candidate pairs valid. Training samples one transformation per anchor before within-batch top-k selection. CCT is the main backbone, with additional MNS results for ResNet-MLP (He et al., 2016), ViT-B/16, and CLIP ViT-B/16. Models use $B = 64$ and $k = 6$, retaining $6/64 \approx 9.4\%$ of each batch. Training, validation, and test splits are kept separate. Hyperparameters, including k and λ_{align}, and checkpoint selection use the validation split under the same protocol for all methods; the held out test split is used only for final reporting. The test-set trajectories in Figure 2 are post hoc diagnostics from fixed training trajectories and are not used for hyperparameter or checkpoint selection. Unless otherwise noted, reported test accuracies are means over 10 independent seeds and $\pm$ denotes standard deviation. Appendix B gives the CCT architecture, and Appendix C gives the full training specification.

4.1 Valid Pair Alignment Provides a Recoverable Benefit

Before asking how much benefit can be recovered without validity labels, we verify that valid pair alignment provides a nontrivial reference. If ordinary task training already behaved consistently on valid transformed counterparts, supervised alignment would add little. Figure 2 shows the opposite. Base training develops a gap between original samples and their valid counterparts on both benchmarks, while supervised alignment substantially improves valid-counterpart accuracy.

This establishes a supervised alignment benefit that can be used as the recovery target in the remaining experiments.

4.2 SymSelect Recovers Most of the Supervised Alignment Benefit

We compare Base, Sym. Sample, Sup. Align, and SymSelect. Sym. Sample adds ground truth valid transformed counterparts without an alignment objective, while Sup. Align aligns ground truth valid pairs and defines the exact validity reference.

Table 1: Accuracy for the solvers evaluated with SymSelect. Base is standard training. Sym. Sample adds valid transformed counterparts without alignment. Sup. Align aligns ground truth valid pairs. SymSelect aligns pairs selected without validity labels. Entries are mean ± standard deviation over 10 independent seeds.

Dataset	Solver	Base	Sym. Sample	Sup. Align	SymSelect
MNS	ResNet-MLP	25.29±0.44	27.03±0.52	30.05±0.56	28.94±0.40
	ViT-B/16	22.78±0.34	24.33±0.27	27.63±0.23	27.22±0.39
	CLIP ViT-B/16	23.46±0.32	26.77±0.38	28.36±0.37	27.59±0.37
	CCT	35.74±0.46	37.86±0.45	40.23±0.41	39.18±0.49
I-RAVEN	CCT	75.62±0.42	77.25±0.46	79.34±0.53	78.97±0.58

Table 2: Behavior on held out valid counterparts, relative to Base.

Dataset	Method	$\Delta v(x')$	$\Delta\gamma$	$\Delta\mathrm{Acc}(x')$
I-RAVEN	Sup. Align − Base	+0.864	−0.528	+0.095
	SymSelect − Base	+0.776	−0.483	+0.083
MNS	Sup. Align − Base	+0.541	−0.323	+0.061
	SymSelect − Base	+0.473	−0.286	+0.052

Table 1 shows substantial recovery without validity labels. Using the corresponding seed-mean accuracies, we measure the recovered fraction of the Base-to-oracle improvement by

$$\text{Recovery} = \frac{\text{Acc}_{\text{SymSelect}} - \text{Acc}_{\text{Base}}}{\text{Acc}_{\text{Sup.Align}} - \text{Acc}_{\text{Base}}}.$$

With CCT, SymSelect recovers 77% of the supervised gain on MNS and 90% on I-RAVEN, with 77%–92% recovery across the additional MNS backbones. SymSelect also exceeds Sym. Sample, so access to valid transformed examples without alignment does not reproduce the recovered gain. Full pool alignment remains 1.64–4.11 accuracy points below SymSelect across the four evaluated MNS backbones; the complete comparison is reported in Appendix A.4.

We also evaluate the valid transformed counterparts directly. For a valid pair $x' = T(x)$, let $\gamma = v(x) - v(x')$ denote the anchor-counterpart confidence gap.

Relative to Base, both supervised alignment and SymSelect increase correct class confidence and accuracy on valid counterparts and reduce the anchor-counterpart gap on both benchmarks. The recovered task benefit is therefore accompanied by improved behavior on the relations that supervised alignment is designed to preserve. Appendix A.5 gives the full confidence decomposition.

4.3 Similarity Selection Outperforms Randomization Within Validity Groups

Recovery alone does not show whether the identity of the retained pairs matters beyond the alignment budget and selected validity composition. We therefore compare selectors that retain the same number of candidate pairs. Random-k samples k candidates uniformly, while Bottom-k retains the lowest-similarity pairs. For Validity-matched random-k, at each mini batch and optimization step the control model first computes the valid and invalid counts that its own current top-k similarity ranking would select, then samples the same counts uniformly within the two validity groups. Validity labels are used only to construct this diagnostic control.

Random selection and Bottom-k provide little recovery relative to SymSelect. The validity-matched control improves substantially over Random-k, showing that selected validity composition matters in this comparison, but it remains 1.66 accuracy points below SymSelect on MNS and 1.76 points below on I-RAVEN. In these experiments, the matched control and SymSelect have the same aggregate selected precision and valid retention on both benchmarks. Preserving selected validity composition while randomizing pair identity therefore does not reproduce the SymSelect result. Aggregate validity composition alone does not account for the observed performance gap.

Table 3: Matched selector controls. $\widehat{\rho}_k$ is selected validity precision and Ret. is the fraction of valid candidates retained. Precision and retention are computed within each seed and then averaged across the 10 independent seeds; accuracy is mean $\pm$ standard deviation over the same seeds.

	MNS			I-RAVEN		
Selector	$\widehat{\rho}_k$	Ret.	Acc.	$\widehat{\rho}_k$	Ret.	Acc.
Base	–	–	35.74 ± 0.46	–	–	75.62 ± 0.42
Bottom-k	0.12	0.04	35.91 ± 0.38	0.07	0.02	75.48 ± 0.37
Random-k	0.28	0.10	36.27 ± 0.44	0.29	0.08	75.96 ± 0.45
Validity-matched random-k	0.74	0.23	37.52 ± 0.41	0.81	0.29	77.21 ± 0.39
SymSelect	0.74	0.23	39.18 ± 0.49	0.81	0.29	78.97 ± 0.58

Table 4: Local oracle reference on I-RAVEN at a fixed trained checkpoint.

Measurement	Result
Oracle margin $\mu_J = -\langle \nabla J, \nabla L_{\text{oracle}} \rangle$	22.14 ± 9.07
$\cos(\nabla L_{\text{sel}}, \nabla L_{\text{oracle}})$	0.57
Oracle step, predicted ΔJ at $\eta = 3 \times 10^{-5}$	6.6×10^{-4}
Oracle step, measured ΔJ at $\eta = 3 \times 10^{-5}$	5.0×10^{-4}

4.4 Selected Alignment Steps Improve the Diagnostic Across Checkpoints

The local inheritance theorem is stated for any differentiable diagnostic J. To instantiate it, we use a transformation-specific interaction diagnostic tied to the regions acted on by each candidate relation. Let

$$v_\theta(x; y) = \log \frac{p_\theta(x)_y}{1 - p_\theta(x)_y}$$

be the correct class log odds, and let $\mathcal{S}(T)$ denote the region pair acted on by candidate transformation T. Define

$$I_\theta(x, T; y) = \frac{1}{|\mathcal{S}(T)|} \sum_{(i,j) \in \mathcal{S}(T)} \big[v_\theta(x; y) - v_\theta(x_{\setminus i}; y) - v_\theta(x_{\setminus j}; y) + v_\theta(x_{\setminus \{i,j\}}; y) \big]. \tag{14}$$

The diagnostic averages this interaction over valid candidate pairs,

$$J(\theta) = \mathbb{E}_{(x,y),T}[I_\theta(x, T; y) \,|\, Z(x, T) = 1], \tag{15}$$

where T follows the same candidate distribution used by the diagnostics. On I-RAVEN, $\mathcal{S}(T^{(1)}) = \{(0, 1)\}$, $\mathcal{S}(T^{(2)}) = \{(3, 4)\}$, and $\mathcal{S}(T^{(3)}) = \{(6, 7)\}$; on MNS, $\mathcal{S}(T)$ is the operand pair swapped by T. Appendix A.7 gives the masking protocol.

Before examining the selected trajectory, we verify that the supervised oracle direction is locally beneficial for this J. At the reference checkpoint, the oracle margin is positive and a single oracle alignment step produces an increase close to its first order prediction.

The positive oracle margin instantiates Assumption 1 at this checkpoint, while the finite step remains close to the first order prediction. The gradient cosine provides a direct orientation comparison between the selected and oracle alignment directions. Appendix A.3 gives the full protocol and remaining reference quantities.

We next measure the actual within-batch selector at six I-RAVEN checkpoints. Besides batch validity precision and valid retention, we compute the relative selected-valid gradient discrepancy

$$\widehat{D}_{\text{valid}} = \frac{\|\widehat{g}_{\text{sel,valid}} - \widehat{g}_{\text{oracle}}\|}{\|\widehat{g}_{\text{oracle}}\|}, \tag{16}$$

where $\widehat{g}_{\text{sel,valid}}$ averages alignment gradients over selected valid candidates and $\widehat{g}_{\text{oracle}}$ averages them over all valid candidates in the evaluation set. We also measure the selected alignment first order effect $-\langle \nabla J, \nabla L_{\text{sel}} \rangle$ and the change in J after the corresponding selected alignment step with $\eta = 3 \times 10^{-5}$.

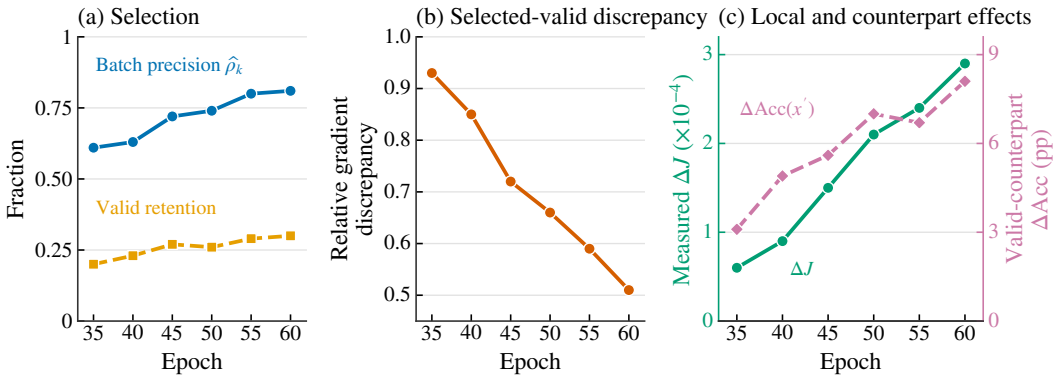

Figure 3: Selector dynamics along one I-RAVEN trajectory. (a) Batch precision increases while valid retention remains near 0.20–0.30. (b) Selected-valid gradient discrepancy decreases. (c) Measured alignment steps improve J, while valid-counterpart accuracy improves overall.

From epoch 35 to epoch 60, actual batch validity precision rises from 0.61 to 0.81, valid retention rises from 0.20 to 0.30, and the relative discrepancy between the selected valid gradient and the full valid gradient falls from 0.93 to 0.51. Every measured selected alignment step improves J, while the valid-counterpart accuracy gain rises overall from 3.1 to 8.1 percentage points. The selected alignment first order effect likewise increases from 2.4 to 11.4. Along this evaluated trajectory, stronger validity enrichment therefore accompanies a smaller relative valid-gradient discrepancy and positive local alignment effects. Appendix A.8 reports the exact checkpoint values together with the global ranking diagnostic.

These measurements characterize the selected alignment update along one training trajectory rather than verifying every term in the sufficient condition of Theorem 1. Together with the local oracle reference above, they show that both the oracle and the realized selected alignment directions improve the same measured diagnostic, while recovery after repeated optimization remains the separate empirical result in Section 4.2.

5 Discussion and Conclusion

Knowing a candidate transformation family does not determine which of its applications should be aligned. Within the controlled candidate families studied here, hidden pair validity does not eliminate the benefit of valid pair alignment: SymSelect recovers a large fraction of the supervised valid pair alignment gain without observing pair validity. The theory adds a broader distinction: selecting mostly valid pairs is not enough when the retained pairs induce an update that differs from the full valid population. Useful selection therefore depends not only on the validity composition of the selected set, but also on how its retained valid pairs shape the alignment update.

The selection controls support this distinction after full training. Matching aggregate validity composition does not reproduce SymSelect, while valid-counterpart results show that the recovered benefit appears on the relations the oracle is designed to preserve. Along the evaluated I-RAVEN trajectory, increasing batch precision and decreasing selected-valid gradient discrepancy accompany positive local alignment effects. Together, these results connect the quantities isolated by the analysis to the behavior of the actual selector without identifying either as a complete explanation of training.

These conclusions are relative to the candidate families, models, and evaluation quantities studied here rather than a global property of arbitrary transformations. The validity-matched diagnostic control uses exact validity for its within-group resampling and evaluation, while SymSelect itself never observes Z. The checkpoint measurements describe one I-RAVEN trajectory, and the inheritance analysis concerns local updates; recovery after repeated optimization remains an empirical result. Within these controlled settings, however, the results separate two aspects of selective alignment that are easy to conflate: enriching valid relations and preserving a useful update among the relations retained.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, to check the correctness and internal consistency of mathematical derivations and proofs, and to assist with experimental plotting.

REPRODUCIBILITY STATEMENT

The main text reports the central objectives, comparisons, and evaluation quantities. The appendix provides proofs, architecture and training details, diagnostic protocols, and supplementary results supporting reproduction of the reported analyses.

REFERENCES

Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2745.

D. Barrett, F. Hill, A. Santoro, A. Morcos, and T. Lillicrap. Measuring abstract reasoning in neural networks. In *Proceedings of the International Conference on Machine Learning*, 2018.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning invariances in neural networks from training data. In *NeurIPS*, 2020.

Abhra Chaudhuri, Serban Georgescu, and Anjan Dutta. Learning conditional invariances through non-commutativity. In *International Conference on Learning Representations*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the International Conference on Machine Learning*, 2016.

Krish Desai, Benjamin Nachman, and Jesse Thaler. Symmetry discovery with deep learning. In *Physical Review D*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969, 2021.

J.-B. Grill, F. Strub, F. Altché, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Bo Han, Quanming Yao, Xiaokang Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Ignacio Hounie, Luiz F. O. Chamon, and Alejandro Ribeiro. Automatic data augmentation via invariance-constrained learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13410–13433, 2023.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Alexander Immer, Tycho F. A. van der Ouderaa, Gunnar Rätsch, Vincent Fortuin, and Mark van der Wilk. Invariance learning in deep neural networks with differentiable laplace approximations. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0904.

M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5464–5474, 2021.

Hyunsu Kim, Yegon Kim, Hongseok Yang, and Juho Lee. Variational partial group convolutions for input-aware partial equivariance of rotations and color-shifts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 24194–24209, 2024.

Jakub Kwiatkowski and Krzysztof Krawiec. Learning abstract visual reasoning via task decomposition: A case study in raven progressive matrices. *International Journal of Applied Mathematics and Computer Science*, 34(2):309–321, 2024. doi: 10.61822/amcs-2024-0022.

Hannah Lawrence, Elyssa Hofgard, Vasco Portilheiro, Yuxuan Chen, Tess Smidt, and Robin Walters. To augment or not to augment? diagnosing distributional symmetry breaking. In *International Conference on Learning Representations*, 2026.

Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, pp. 3361–3391. MIT Press, 1995.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013.

D. Marcos, M. Volpi, N. Komodakis, and D. Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5048–5057, 2017.

Ning Miao, Tom Rainforth, Emile Mathieu, Yann Dubois, Yee Whye Teh, Adam Foster, and Hyunjik Kim. Learning instance-specific augmentations by capturing local invariances. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 24720–24736, 2023.

Artem Moskalev, Anna Sepliarskaia, Erik J. Bekkers, and Arnold W. M. Smeulders. On genuine invariance learning without weight-tying. In *Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning*, volume 221, pp. 218–227, 2023.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4334–4343, 2018.

Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant / equivariant deep networks via quotient feature spaces. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 771–780, 2021.

A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, 2017.

Pola Schwöbel, Martin Jørgensen, Sebastian W. Ober, and Mark van der Wilk. Last layer marginal likelihood for invariance learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 3542–3555, 2022.

Han Shao, Omar Montasser, and Avrim Blum. A theory of PAC learnability under transformation invariances. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1017.

Utkarsh Singhal, Carlos Esteves, Ameesh Makadia, and Stella X. Yu. Learning to transform for generalizable instance-wise invariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6211–6221, 2023.

Ashkan Soleymani, Behrooz Tahmasebi, Stefanie Jegelka, and Patrick Jaillet. Learning with exact invariances in polynomial time. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 56101–56121, 2025.

Behrooz Tahmasebi, Melanie Weber, and Stefanie Jegelka. Data augmentation: A fourier analysis perspective. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6114–6155, 2026.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Rupert Tombs and Christopher G. Lester. A method to challenge symmetries in data with self-supervised learning. *Journal of Instrumentation*, 17(08):P08024, 2022. doi: 10.1088/1748-0221/17/08/P08024.

D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5028–5037, 2017.

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *ICML*, 2023a.

Lingxiao Yang, Hongzhi You, Zonglei Zhen, Dahui Wang, Xiaohong Wan, Xiaohua Xie, and Ru-Yuan Zhang. Neural prediction errors enable analogical visual reasoning in human standard intelligence tests. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39572–39583, 2023b.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. RAVEN: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Chi Zhang, Sirui Xie, Baoxiong Jia, Ying Nian Wu, Song-Chun Zhu, and Yixin Zhu. Learning algebraic representation for systematic generalization in abstract reasoning. In *Computer Vision – ECCV 2022*, pp. 692–709, 2022. doi: 10.1007/978-3-031-19842-7_40.

Wenhe Zhang. Machine Number Sense: A dataset of visual arithmetic problems for abstract and relational reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Allan Zhou, Tom Knowles, and Chelsea Finn. Meta-learning symmetries by reparameterization. In *ICLR*, 2021.

Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. *Advances in Neural Information Processing Systems*, 34: 4328–4341, 2021.

A Proofs and Diagnostic Details

This appendix provides proofs, finite-batch extensions, diagnostic details, and supplementary experiments supporting the main text. Appendix A.1 proves Theorem 1, and Appendix A.2 proves Proposition 1. Appendix A.11 gives the finite-batch proof, ranking and random-selection extensions, the joint classification update, and the cosine-loss counterexample. Appendix A.3 gives the oracle local measurement protocol. The remaining sections report full-pool and counterpart controls, diagnostic construction details, the complete checkpoint trajectory, additional MNS supervised-alignment baselines, and the candidate-family check.

A.1 Proof of the Local Inheritance Theorem

Population and mini batch selection. Theorem 1 is stated at the population level using a marginalized selection event A_k. Training instead selects the top k candidates jointly within each mini batch, so the selected status of one pair also depends on the other batch members. The finite batch result in Section 3.3 and Appendix A.11 analyzes this actual ranked-set structure directly. The population theorem and finite batch result therefore answer complementary questions rather than identifying the mini batch event with an independent per-pair selector.

We prove Theorem 1. Recall that

$$L_{\mathrm{oracle}}(\theta) = \mathbb{E}\big[\ell_{\mathrm{align}}(x, Tx) \,\big|\, Z = 1\big],$$

$$L_{\mathrm{sel}}(\theta) = \mathbb{E}\big[\ell_{\mathrm{align}}(x, Tx) \,\big|\, A_k = 1\big].$$

The top-k selection event A_k is computed at the current checkpoint and treated as fixed during the subsequent update.

Let $g(x, T) = \nabla_\theta \ell_{\mathrm{align}}(x, Tx)$. Conditioning on whether a selected pair is valid gives

$$\begin{aligned}\nabla_\theta L_{\mathrm{sel}} &= \mathbb{E}\big[g(x,T) \,\big|\, A_k = 1\big] \\ &= \rho_k \,\mathbb{E}\big[g(x,T) \,\big|\, A_k = 1, Z = 1\big] \\ &\quad + (1-\rho_k)\,\mathbb{E}\big[g(x,T) \,\big|\, A_k = 1, Z = 0\big].\end{aligned} \tag{17}$$

Define

$$g_{\mathrm{oracle}} = \nabla_\theta L_{\mathrm{oracle}} = \mathbb{E}\big[g(x,T) \,\big|\, Z = 1\big],$$

$$g_{\mathrm{valid}|A} = \mathbb{E}\big[g(x,T) \,\big|\, A_k = 1, Z = 1\big],$$

$$g_{\mathrm{invalid}|A} = \mathbb{E}\big[g(x,T) \,\big|\, A_k = 1, Z = 0\big].$$

Then

$$\nabla_\theta L_{\mathrm{sel}} - g_{\mathrm{oracle}} = \rho_k(g_{\mathrm{valid}|A} - g_{\mathrm{oracle}}) + (1-\rho_k)(g_{\mathrm{invalid}|A} - g_{\mathrm{oracle}}). \tag{18}$$

Taking norms and using the definition of ϵ_{sel},

$$\big\|\nabla_\theta L_{\mathrm{sel}} - \nabla_\theta L_{\mathrm{oracle}}\big\| \le \rho_k\,\epsilon_{\mathrm{sel}} + (1-\rho_k)\big\|g_{\mathrm{invalid}|A} - g_{\mathrm{oracle}}\big\|. \tag{19}$$

Since individual alignment gradients satisfy $\|g(x,T)\| \le G$, we also have $\|g_{\mathrm{invalid}|A}\| \le G$ and $\|g_{\mathrm{oracle}}\| \le G$, so $\|g_{\mathrm{invalid}|A} - g_{\mathrm{oracle}}\| \le 2G$. Therefore,

$$\big\|\nabla_\theta L_{\mathrm{sel}} - \nabla_\theta L_{\mathrm{oracle}}\big\| \le \epsilon_{\mathrm{sel}} + 2(1-\rho_k)G. \tag{20}$$

We now connect the gradient deviation in Equation (20) to the diagnostic improvement. By the L_J-Lipschitz gradient assumption on J, the descent lemma gives, for any update direction u:

$$J(\theta - \eta u) - J(\theta) \ge -\eta\langle \nabla_\theta J(\theta), u\rangle - \frac{L_J}{2}\eta^2\|u\|^2. \tag{21}$$

Setting $u = \nabla_\theta L_{\mathrm{sel}}(\theta)$, the first-order term decomposes as

$$-\langle \nabla_\theta J, \nabla_\theta L_{\mathrm{sel}}\rangle = -\langle \nabla_\theta J, \nabla_\theta L_{\mathrm{oracle}}\rangle - \langle \nabla_\theta J, \nabla_\theta L_{\mathrm{sel}} - \nabla_\theta L_{\mathrm{oracle}}\rangle. \tag{22}$$

The first term on the right equals μ_J by Assumption 1. For the second term, the Cauchy–Schwarz inequality combined with $\|\nabla_\theta J\| \le H_J$ and Equation (20) gives

$$\big|\langle \nabla_\theta J, \nabla_\theta L_{\mathrm{sel}} - \nabla_\theta L_{\mathrm{oracle}}\rangle\big| \le H_J\,\big(\epsilon_{\mathrm{sel}} + 2(1-\rho_k)G\big). \tag{23}$$

Combining Equations (21)–(23) yields

$$J(\theta^+) - J(\theta) \ge \eta\big[\mu_J - H_J(\epsilon_{\mathrm{sel}} + 2(1-\rho_k)G)\big] - \tfrac{L_J}{2}\eta^2\|\nabla_\theta L_{\mathrm{sel}}(\theta)\|^2,$$

which is Equation (7). For sufficiently small η, the second-order term is dominated by the first, and the right-hand side is positive whenever $\mu_J > H_J(\epsilon_{\mathrm{sel}} + 2(1-\rho_k)G)$. □

A.2 Scope of the Local Inheritance Argument

We prove Proposition 1.

Fix an arbitrary model parameterization a, checkpoint $\theta_a \in \Theta_a$, candidate transformation family $\mathcal{T}$, and fixed sampling rule over its candidates. Define

$$g^a(x, T) = \nabla_{\theta_a} \ell^a_{\text{align}}(x, T(x)).$$

The oracle and selected gradients are

$$\nabla_{\theta_a} L^a_{\text{oracle}} = \mathbb{E}[g^a(x, T) \mid Z = 1],$$

and

$$\nabla_{\theta_a} L^a_{\text{sel}} = \mathbb{E}[g^a(x, T) \mid A_k = 1].$$

Conditioning the selected gradient on Z gives

$$\begin{aligned}\nabla_{\theta_a} L^a_{\text{sel}} = \rho^a_k \, \mathbb{E}[g^a \mid A_k = 1, Z = 1] \\ + (1 - \rho^a_k) \, \mathbb{E}[g^a \mid A_k = 1, Z = 0].\end{aligned}$$

Hence the same decomposition used in the proof of Theorem 1 yields

$$\|\nabla_{\theta_a} L^a_{\text{sel}} - \nabla_{\theta_a} L^a_{\text{oracle}}\| \le \epsilon^a_{\text{sel}} + 2(1 - \rho^a_k) G^a.$$

Applying the L^a_J-smoothness bound for J^a and the oracle margin μ^a_J gives the bound in Theorem 1 with the corresponding quantities for that instance. Since the model parameterization and candidate transformation sampling rule were arbitrary, this proves Proposition 1. □

A.3 Direct Gradient Measurement Protocol

Table 4 reports the main oracle reference quantities. The complete measurements are computed on I-RAVEN at a fixed trained checkpoint using $1{,}000$ held out examples, which yield $3{,}000$ candidate transformed pairs under the three row transformations. Let $\mathcal{V} = \{(x, y, T) : Z(x, T) = 1\}$ denote the held out valid evaluation pairs. We instantiate Equation (15) by

$$\widehat{J}(\theta) = \frac{1}{|\mathcal{V}|} \sum_{(x,y,T) \in \mathcal{V}} I_\theta(x, T; y). \tag{24}$$

We compute oracle and selected alignment gradients from the same model parameters and evaluate the corresponding first order quantities directly in parameter space. Uncertainty in the oracle margin is estimated with $1{,}000$ bootstrap resamples of the held out evaluation set.

Table 5: Original local reference measurements on I-RAVEN at a fixed trained checkpoint.

Measurement	Result
Oracle margin $-\langle \nabla J, \nabla L_{\text{oracle}} \rangle$	22.14 ± 9.07
$\|\nabla J\|$	26.9
$\cos(\nabla L_{\text{sel}}, \nabla L_{\text{oracle}})$	0.57
$-\langle \nabla J, \nabla L_{\text{sel}} \rangle$	positive, 4.5 bootstrap s.d. above zero
Oracle step, predicted ΔJ at $\eta = 3 \times 10^{-5}$	6.6×10^{-4}
Oracle step, measured ΔJ at $\eta = 3 \times 10^{-5}$	5.0×10^{-4}

The gradient cosine describes orientation relative to the oracle and is not substituted for ϵ_{sel} in Theorem 1. The oracle finite step is used only to check the local approximation for that reference direction. The selected trajectory measurements in Figure 3 use the actual within-batch selector at six checkpoints and are reported separately in the main text.

A.4 Full Pool Alignment Control

Full pool alignment uses the same candidate pool and pairwise loss as SymSelect but aligns every candidate. It remains below SymSelect on all four evaluated MNS backbones.

Table 6: Full pool alignment comparison on MNS.

Solver	Align all	SymSelect	SymSelect − Align all
ResNet-MLP	26.03	28.94	+2.91
ViT-B/16	23.11	27.22	+4.11
CLIP ViT-B/16	24.98	27.59	+2.61
CCT	37.54	39.18	+1.64

A.5 Behavior on Valid Counterparts

Table 2 reports the main counterpart quantities. Here we give the full decomposition. For a valid pair $x' = T(x)$, define

$$\gamma_\theta(x) = v_\theta(x) - v_\theta(x').$$

Comparing two checkpoints gives

$$\Delta v(x') = \Delta v(x) - \Delta\gamma. \tag{25}$$

This identity separates counterpart improvement into the change on the anchor and the change in the anchor-counterpart gap.

Table 7: Counterpart confidence improvement on held-out valid pairs, relative to Base.

Dataset	Method	$\Delta\mathrm{Sim}$	$\Delta v(x)$	$\Delta v(x')$	$\Delta\gamma$	$\Delta\mathrm{Acc}(x')$
I-RAVEN	Sup. Align − Base	+0.033	+0.336	+0.864	−0.528	+0.095
	SymSelect − Base	+0.027	+0.292	+0.776	−0.483	+0.083
MNS	Sup. Align − Base	+0.024	+0.218	+0.541	−0.323	+0.061
	SymSelect − Base	+0.019	+0.187	+0.473	−0.286	+0.052

Relative to Base, both SymSelect and supervised alignment increase confidence and accuracy on held-out valid counterparts and reduce the anchor-counterpart gap on both benchmarks. The identity above follows directly from $v_\theta(x') = v_\theta(x) - \gamma_\theta(x)$.

A.6 Structural Interaction Decomposition

Using the interaction I_θ from Equation (14), define

$$B_\theta(x, T; y) = \frac{1}{|\mathcal{S}(T)|} \sum_{(i,j)\in\mathcal{S}(T)} \big[v_\theta(x_{\setminus i}; y) + v_\theta(x_{\setminus j}; y) - v_\theta(x_{\setminus\{i,j\}}; y)\big].$$

Then $I_\theta(x, T; y) = v_\theta(x; y) - B_\theta(x, T; y)$, so between two checkpoints

$$\Delta I = \Delta v(x) - \Delta B. \tag{26}$$

Table 8: Structural interaction decomposition, relative to Base.

Dataset	Method	$\Delta v(x)$	ΔB	ΔI	$\Delta B/\Delta v(x)$
I-RAVEN	Sup. Align − Base	+0.336	+0.029	+0.307	0.087
	SymSelect − Base	+0.292	+0.022	+0.270	0.075
MNS	Sup. Align − Base	+0.218	+0.024	+0.194	0.110
	SymSelect − Base	+0.187	+0.018	+0.169	0.096

Under the same masking intervention used to define J, both supervised alignment and SymSelect produce positive ΔI on both benchmarks, while ΔB is much smaller. The masking baselines and transformed region sets are specified in Appendix A.7.

A.7 Diagnostic Construction Details

Panel indexing and transformation (I-RAVEN). Panels are indexed in row major order:

$$(0, 1, 2), \quad (3, 4, 5), \quad (6, 7, \square).$$

We use the same family of three candidate transformations for both training and diagnostics. Each transformation swaps the first two observed panels within exactly one row and leaves all other panels unchanged:

$$T^{(1)} : (0, 1), \quad T^{(2)} : (3, 4), \quad T^{(3)} : (6, 7).$$

During the held out diagnostics, applying all three members of this family to an evaluation sample yields three candidate transformed pairs. The interaction diagnostic is indexed by the same candidate transformation, with $\mathcal{S}(T^{(1)}) = \{(0, 1)\}$, $\mathcal{S}(T^{(2)}) = \{(3, 4)\}$, and $\mathcal{S}(T^{(3)}) = \{(6, 7)\}$. Restricting the swaps to the first two positions avoids the asymmetric semantic role of the third column, which contains the target panel for each of the first two rows and the missing answer position in the last row.

Validity label. For each candidate pair $(x, T(x))$, validity is defined by

$$Z(x, T) = \mathbf{1}\{y(T(x)) = y(x)\}.$$

Training SymSelect does not use this validity label; it is used only in held out diagnostics to measure selection precision. The marginal validity prior $\pi_0 = \mathbb{P}(Z = 1)$ used in the diagnostics is computed empirically over the candidate distribution. On MNS, $\pi_0 = 0.301$ reflects the fraction of operand swap candidates falling on commutative operators (addition and multiplication) under the dataset's operator distribution, rather than a uniform 50% across all four operators. On I-RAVEN, $\pi_0 = 0.259$ reflects the fraction of candidate swaps within rows whose latent rule (Constant, Arithmetic, or Distribute Three) is insensitive to the corresponding order change.

Similarity score. The held out selection diagnostic uses the same similarity score as SymSelect training on both benchmarks. For each candidate pair, we compute the projection embeddings $z = E(f_\theta(x))$ and $z' = E(f_\theta(T(x)))$ and rank pairs by the cosine similarity S defined in Equation (2). Validity labels are used only after ranking to evaluate precision and are never used to form the similarity score or the ranking.

Log odds computation. All prediction quantities use the anchor label $y = y(x)$:

$$v_\theta(x; y) = \log \frac{p_\theta(x)_y}{1 - p_\theta(x)_y}.$$

For numerical stability, probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$ before computing log odds.

Masking protocol. On I-RAVEN, masked panels are replaced by a baseline defined by the mode of each image channel. On MNS, masked digit regions are replaced by the panel background color. For a region index i, $x_{\setminus i}$ denotes the input where region i is replaced; for a pair (i, j), $x_{\setminus \{i,j\}}$ denotes the input where both are replaced. The numerical identity $I(x, T) = v(x) - B(x, T)$ is verified to within 10^{-7} on all evaluated candidate pairs.

Global ranking diagnostic. The complete checkpoint table in Appendix A.8 includes the held out global top-10% precision used in the earlier ranking diagnostic. This statistic is separate from the actual within-batch precision reported in the main text; training always samples one transformation candidate per anchor and applies top-k selection within each mini batch.

Top-k selection ratio. Changing the retained fraction affects both contamination and the size of the valid subset retained. The main selector control reports the realized batch precision and valid retention at the training operating point, while Appendix A.8 reports their evolution along the evaluated I-RAVEN trajectory.

Table 9: Complete selector and behavior diagnostics across the evaluated I-RAVEN trajectory. The checkpoint quantities are descriptive measurements from this trajectory rather than across-seed estimates; measured ΔJ uses $\eta = 3 \times 10^{-5}$ throughout.

Epoch	Global $\rho_{10\%}$	Batch $\widehat{\rho}_k$	Valid retention	$\widehat{D}_{\mathrm{valid}}$	Selected effect	Measured ΔJ	$\Delta\mathrm{Acc}(x')$
35	0.631	0.61	0.20	0.93	+2.4	0.6×10^{-4}	+0.031
40	0.724	0.63	0.23	0.85	+3.7	0.9×10^{-4}	+0.049
45	0.771	0.72	0.27	0.72	+6.3	1.5×10^{-4}	+0.056
50	0.819	0.74	0.26	0.66	+8.2	2.1×10^{-4}	+0.070
55	0.852	0.80	0.29	0.59	+9.3	2.4×10^{-4}	+0.067
60	0.858	0.81	0.30	0.51	+11.4	2.9×10^{-4}	+0.081

A.8 Diagnostics Across Checkpoints

Table 9 gives the complete I-RAVEN checkpoint series underlying Figure 3, including the global held out ranking diagnostic and the change in valid-counterpart accuracy relative to Base.

The global and within-batch precision measures follow similar overall trends but are not the same statistic. The valid-counterpart accuracy improvement is mildly nonmonotonic, while the selected local effect and measured ΔJ remain positive at every evaluated checkpoint.

A.9 Additional MNS Supervised Alignment Baselines

The main performance table includes only solvers evaluated with SymSelect. For completeness, Table 10 reports the additional MNS solvers used in the supervised-alignment comparison but not evaluated with SymSelect.

Table 10: Additional MNS supervised-alignment baselines.

Solver	Base	Sym. Sample	Sup. Align
CNN-MLP	22.71 ± 0.11	24.50 ± 0.16	25.07 ± 0.14
CNN-LSTM	22.16 ± 0.52	24.44 ± 0.58	25.97 ± 0.51
RN	22.96 ± 0.62	24.28 ± 0.61	25.17 ± 0.64

A.10 Candidate Family Size Check

The main experiments use compact candidate families. Expanding the MNS candidate set to $|\mathcal{T}| = 2, 3, 4$ gives SymSelect accuracies of 39.53%, 40.29%, and 40.14%, respectively. Performance is therefore similar across these tested candidate counts.

A.11 Finite Batch Analysis of Top-k Selection

For a fixed mini batch, let $\ell_i(\theta) = 1 - S_i(\theta)$, $g_i = \nabla_\theta \ell_i(\theta)$, and $a_i = -\langle \nabla_\theta J, g_i \rangle$. For a selected set $\mathcal{K}$,

$$a_{\mathcal{K}} = \frac{1}{k} \sum_{i \in \mathcal{K}} a_i = - \left\langle \nabla_\theta J, \frac{1}{k} \sum_{i \in \mathcal{K}} g_i \right\rangle . \tag{27}$$

For $z \in \{0, 1\}$, let $\bar{a}_z$ and σ_z^2 denote the mean and variance of a_i within validity group z. We first prove the finite-set lemma used in Theorem 2.

Lemma 1 (Subset mean deviation)**.** *Let $u_1, \ldots, u_n$ have mean μ and population variance*

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (u_i - \mu)^2.$$

For any subset A of size $t \in \{1, \dots, n\}$, with subset mean μ_A,

$$|\mu_A - \mu| \le \sigma \sqrt{\frac{n-t}{t}}. \tag{28}$$

Proof. Let μ_C be the mean of the complement. Since

$$t(\mu_A - \mu) + (n-t)(\mu_C - \mu) = 0,$$

we have

$$\mu_C - \mu = -\frac{t}{n-t}(\mu_A - \mu)$$

when $t < n$; the claim is immediate when $t = n$. Dropping the nonnegative within-group variation from the total squared deviation gives

$$\begin{aligned} n\sigma^2 &\ge t(\mu_A - \mu)^2 + (n-t)(\mu_C - \mu)^2 \\ &= \frac{nt}{n-t}(\mu_A - \mu)^2. \end{aligned}$$

Rearranging yields Equation (28). Equality can occur when the values are constant within the selected subset and constant within its complement. □

Proof of Theorem 2. Let $\mathcal{K}_1 = \{i \in \mathcal{K} : Z_i = 1\}$ and $\mathcal{K}_0 = \{i \in \mathcal{K} : Z_i = 0\}$. Their sizes are $k - r$ and r, respectively. Applying Lemma 1 to the valid group gives, when $k - r > 0$,

$$\frac{1}{k-r} \sum_{i \in \mathcal{K}_1} a_i \ge \bar{a}_1 - \sigma_1 \sqrt{\frac{n_1 - k + r}{k-r}}.$$

Likewise, when $r > 0$,

$$\frac{1}{r} \sum_{i \in \mathcal{K}_0} a_i \ge \bar{a}_0 - \sigma_0 \sqrt{\frac{n_0 - r}{r}}.$$

Weighting the two selected-group means by $(k-r)/k$ and r/k gives Equation (10). Setting $r = 0$ gives Equation (11). □

Ranking inversions and contamination. Let N_{inv} count ordered invalid-valid pairs for which the invalid candidate is ranked above the valid candidate under the same total ordering used by top-k selection. Suppose the selected set contains r invalid pairs. It therefore contains $k - r$ valid pairs, leaving $n_1 - k + r$ valid candidates outside the top k. Every selected invalid candidate is ranked above every unselected valid candidate, which proves Equation (13). For $n_1 \ge k$, solving the resulting quadratic inequality for the nonnegative root gives

$$r \le \frac{\sqrt{(n_1 - k)^2 + 4N_{\text{inv}}} - (n_1 - k)}{2}. \tag{29}$$

Uniform random selection. Let R be uniformly distributed over all k-element subsets of the batch. Since each candidate is included with probability k/B,

$$\mathbb{E}_R[a_R] = \frac{1}{B} \sum_{i=1}^{B} a_i. \tag{30}$$

Writing the selected and full batch means by validity group, with $\pi = n_1/B$, $\rho = (k-r)/k$, and b_z the selected mean contribution minus the full validity group mean within class z, gives Equation (12).

Joint classification and alignment step. Let

$$c = \nabla_\theta L_{\text{CE}}, \qquad \nu_{\text{CE}} = -\langle \nabla_\theta J, c \rangle,$$

and

$$u = c + \lambda_{\text{align}} \frac{1}{k} \sum_{i \in \mathcal{K}} g_i.$$

By L_J-smoothness,

$$J(\theta - \eta u) - J(\theta) \ge \eta \left[\nu_{\text{CE}} + \lambda_{\text{align}} a_{\mathcal{K}} \right] - \frac{L_J}{2} \eta^2 \|u\|^2. \tag{31}$$

This places the selected alignment contribution inside the full classification and alignment step without turning the local result into a guarantee for the complete optimization trajectory.

Proof of Proposition 2. Let the parameter be scalar and set

$$J(\theta) = \theta.$$

For three candidate pairs, realize the cosine similarity by the unit vectors

$$z = (1, 0), \qquad z_i'(\theta) = \big(\cos(\phi_i + \omega_i\theta),\, \sin(\phi_i + \omega_i\theta)\big),$$

so that

$$S_i(\theta) = z^\top z_i'(\theta) = \cos(\phi_i + \omega_i\theta), \qquad \ell_i(\theta) = 1 - S_i(\theta).$$

Choose $\phi_i \in (0, \pi)$ so that

$$(\cos\phi_1, \cos\phi_2, \cos\phi_3) = (0.9, 0.8, 0.1).$$

Let pairs 1 and 2 be valid and pair 3 invalid. Set the desired first-order contributions to

$$(a_1, a_2, a_3) = (-1, 3, 0)$$

and choose

$$\omega_i = -\frac{a_i}{\sin\phi_i}.$$

At $\theta = 0$,

$$\frac{d\ell_i}{d\theta} = \omega_i \sin\phi_i,$$

so, because $J'(0) = 1$,

$$-J'(0)\frac{d\ell_i}{d\theta} = a_i.$$

The similarity ranking places both valid pairs above the invalid pair. With $k = 1$, SymSelect selects pair 1 with validity precision one, but

$$a_{\mathcal{K}} = -1,$$

whereas oracle alignment over the two valid pairs has

$$\bar{a}_1 = \frac{-1 + 3}{2} = 1.$$

Thus perfect validity ranking can coexist with a selected update whose first-order effect has the opposite sign from the oracle, proving Proposition 2.

B Architectural Details of CCT

We provide the full architectural specification of the Context-Choice Transformer (CCT) used in our experiments. CCT uses the same four stages on both benchmarks: a vision head applied to each panel, assembly of the reasoning sequence, a transformer encoder with context normalization, and two heads (P, E) for classification and representation alignment. The instantiations on I-RAVEN and MNS differ only in components dictated by the task interface (8-way candidate selection versus K-way classification). We describe the shared design first and then the specialization for each task.

Shared design. Each input panel is independently encoded by a convolutional vision head consisting of stacked 5×5 and 3×3 convolutions with ReLU activations. The resulting feature map is augmented with a soft positional embedding obtained by linearly projecting a normalized 2D coordinate grid (concatenated with its complement, following the SoftPositionEmbed formulation) and added to the feature map before flattening into a set of panel tokens.

The panel tokens are assembled into a reasoning sequence (specified per benchmark below), to which a learnable `[CLS]` token is prepended. Before entering the transformer, the sequence passes through a context normalization step. Features are normalized along the token axis, rescaled by a learnable affine transform (γ, β), and convexly combined with the unnormalized features through a coefficient α that is linearly annealed across training,

$$\tilde{x} = \alpha(t)\,\big(\gamma \cdot \mathrm{Norm}(x) + \beta\big) + \big(1 - \alpha(t)\big)\,x, \tag{32}$$

where $\alpha(t) = 1$ for the first few epochs and decays linearly to a small constant thereafter. This encourages the model to rely on relations across panels early in training while allowing it to retain absolute features later on.

The transformer encoder is a standard pre-norm stack with GELU activations, multi-head self-attention, and feed-forward expansion ratio 4. From its outputs we form a backbone representation r (the precise pooling is specific to each task; see below). Two heads operate on r: a scoring head P, implemented as $\mathrm{LayerNorm} \to \mathrm{Linear}$, produces task logits; and a projection head E, implemented as a small MLP with batch normalization and ℓ_2 output normalization, produces alignment features. P and E share the backbone but are otherwise independent.

I-RAVEN instantiation. Each panel is reduced to a single vector by attention pooling: a learned linear scorer over panel tokens produces softmax weights, and the weighted sum gives the panel embedding. For the 16 panels of an I-RAVEN problem (8 context + 8 candidates), each candidate is paired independently with the 8 context panels to form a length-9 reasoning sequence; this is realized by expanding the batch dimension from B to $B \cdot 8$. A grid positional embedding indicating each panel's position in the underlying 3×3 matrix (encoded as a 6-dimensional row/column one-hot and projected linearly) is added to the corresponding panel token. After context normalization and `[CLS]` prepending, the sequence is processed by the transformer encoder, and the backbone representation is formed by concatenating the `[CLS]` output with the mean of the remaining token outputs. The scoring head P maps this representation to a scalar per candidate; the $B \cdot 8$ scalars are reshaped to logits in $\mathbb{R}^{B\times8}$ for 8-way candidate selection.

MNS instantiation. Panel tokens are kept at their full spatial resolution rather than pooled. The three panels of an MNS problem contribute their spatial tokens to a single concatenated sequence; a spatial positional embedding is added to each token to encode its location within a panel. After context normalization and `[CLS]` prepending, the sequence is processed by the transformer encoder, and the `[CLS]` output (optionally combined with the mean of the token outputs) serves as the backbone representation. The scoring head P maps this representation directly to logits in $\mathbb{R}^{B\times K}$ for K-way classification.

Summary of differences required by the tasks. The two instantiations differ in three places required by the task interface. I-RAVEN represents each panel by a single vector obtained through attention pooling, whereas MNS keeps a full set of spatial tokens. I-RAVEN forms a length-9 sequence for each answer candidate and adds grid positional embeddings, whereas MNS uses one concatenated sequence with spatial positional embeddings. Finally, P produces a scalar score for each I-RAVEN answer candidate and a direct K-way output for MNS. The vision head, transformer encoder with context normalization and annealed $\alpha(t)$, and the two heads (P, E) are otherwise shared across the benchmarks.

C Training Details and Design Choices

This appendix records the training hyperparameters and the design choices behind the main components of SymSelect.

Training hyperparameters. All SymSelect experiments use the AdamW optimizer with learning rate 10^{-4}, weight decay 10^{-2}, and a cosine learning rate schedule decaying from 10^{-4} to 10^{-6} over the full training horizon. Models are trained for 100 epochs on both MNS and I-RAVEN, with a mini batch size of $B = 64$ throughout. Hyperparameter and checkpoint selection use the validation split, and the held out test split is reserved for final reporting; test-set trajectories are computed only as post hoc diagnostics from fixed training trajectories.

The selection rule operates within each mini batch, so it is worth confirming that a typical batch carries more valid pairs than the top-k slots being filled. With $B = 64$, $k = 6$, and $\pi_0 \in [0.25, 0.30]$ on the two benchmarks, the expected number of valid pairs per batch is $B\pi_0 \approx 16$–19, and a binomial calculation gives $\mathbb{P}(V \geq k) \geq 1 - 10^{-4}$. Thus, under the empirical validity priors used here, the selected set is not forced to fill its top-k slots from a shortage of valid candidates. The local analyses depend on the realized composition of the selected set, while the effect of changing batch size or selection ratio must still be evaluated through the resulting selection statistics.

On I-RAVEN, the expansion from B to $B \cdot 8$ described in Appendix B comes from the eight answer candidates in the task interface, not from eight transformation candidates in SymSelect. Panels are

encoded before this reasoning expansion over answer candidates. This answer candidate dimension is therefore distinct from the transformation family $\mathcal{T}$ used by the selection rule.

Alignment loss weight We set $\lambda_{\text{align}} = 1$ throughout. The alignment loss L_{Align} is bounded in $[0, 2]$ by construction as one minus a mean cosine similarity.

Projection head. The projection head E is a two layer MLP with a linear projection of the backbone representation, a ReLU activation, and a final linear layer to the alignment dimension. Its outputs are ℓ_2 normalized before the cosine similarity in Equation (2).

D Baseline Behavior: Detailed Training Curves

Tables 11 and 12 report the numerical values underlying Figure 2. On both datasets, base training produces a measurable accuracy difference between original samples and their valid symmetric counterparts. Supervised oracle alignment substantially reduces this difference throughout the reported training interval. These curves motivate oracle alignment as a useful reference for the recovery experiments; the gradient space oracle margin itself is measured directly in Section 4.4.

Table 11: Training curves on MNS. Base / Original: accuracy on standard test samples under base training. Base / Symmetric and Sup. Align / Symmetric: accuracy on valid operand swap counterparts under each training scheme.

	Epoch 0	Epoch 3	Epoch 6	Epoch 9	Epoch 12
Base / Original	20.19%	23.64%	28.50%	30.95%	33.57%
Base / Symmetric	21.37%	22.90%	26.17%	28.00%	29.73%
Sup. Align / Symmetric	21.30%	22.86%	28.05%	30.48%	33.26%

Table 12: Training curves on I-RAVEN. Base / Original: accuracy on standard test samples under base training. Base / Symmetric and Sup. Align / Symmetric: accuracy on valid counterparts formed by swapping the first two panels within a row under each training scheme.

	Epoch 10	Epoch 20	Epoch 30	Epoch 40	Epoch 50
Base / Original	20.71%	29.75%	44.18%	59.82%	66.20%
Base / Symmetric	20.30%	27.50%	44.37%	53.44%	57.75%
Sup. Align / Symmetric	22.38%	29.01%	44.07%	60.16%	65.41%

E Random Augmentation versus Valid Symmetric Counterparts

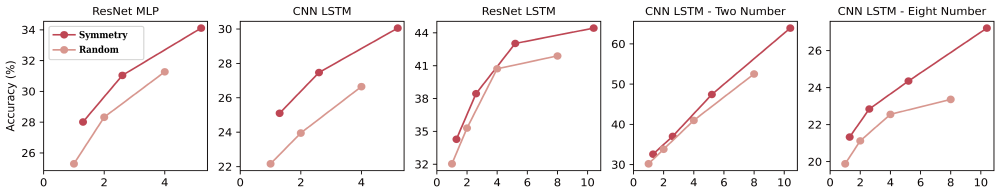

Figure 4: Random augmentation versus valid symmetric counterparts on MNS. Across architectures and difficulty levels, valid symmetric counterparts yield larger gains than an equivalent volume of random samples.

We compare two augmentation strategies under matched data volume. Random augmentation adds randomly generated samples, while the symmetry based strategy adds valid symmetric counterparts of existing samples. Since approximately 30% of MNS samples admit valid symmetric transformations,

adding these counterparts increases the training set to about $1.3\times$ the original size; the random baseline uses an equivalent amount of additional data.

As shown in Figure 4, augmentation with valid symmetric counterparts consistently outperforms random augmentation across architectures and difficulty levels, with the gap especially clear on more complex settings where random augmentation provides limited improvement. The comparison shows that the source of the additional training samples matters: matched data volume alone does not reproduce the gain obtained from valid symmetric counterparts.

This experiment is complementary to the align all comparison in Table 6. The augmentation experiment compares the information carried by valid symmetric examples with matched random data, whereas align all directly tests whether applying the same alignment loss without selection can match SymSelect on the noisy candidate pool.