NEXT DEMONSTRATION PREDICTION FOR VISUAL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting visual in-context demonstrations requires choosing both examples and their order. Visual nearest-neighbor retrieval can produce redundant prompts, while diversity heuristics cannot adapt their selection to a downstream vision-language model (VLM). We propose Next Demonstration Prediction (NDP), an autoregressive selector initialized by behavioral cloning of maximal marginal relevance and then trained with PPO using a frozen VLM's task reward. On 14 benchmarks with Gemma 3 4B-it at $K = 16$, NDP achieves the lowest mean regression MAE and highest mean VQA accuracy among the evaluated image-only selectors, but trails DPP on classification. Benefits vary by task and are limited at small shot counts. Reordering NDP's selected examples reduces performance in the evaluated in-domain settings, whereas cross-dataset policy transfer can underperform simple heuristics. NDP therefore offers a task-dependent, ordered alternative to static retrieval, at the cost of substantial VLM-based policy training.

1 INTRODUCTION

In-context learning (ICL) has fundamentally altered the paradigm of few-shot adaptation, empowering Vision-Language Models (VLMs) and Large Language Models (LLMs) to perform complex reasoning without updating model parameters Dong et al. (2022); Wei et al. (2022); Zhu et al. (2025). Despite its dominance, ICL is constrained by its extreme sensitivity to prompt configuration; the specific selection and ordering of demonstration examples can dictate the success or failure of a model on a given task Wang et al. (2023); Liu et al. (2024).

The default mechanism for assembling these prompts is k-Nearest Neighbors (k-NN) retrieval. Relying purely on visual feature similarity is a poor match for complex reasoning. k-NN tends to fetch highly redundant clusters, optimizing for superficial similarity at the expense of structural diversity and decision-boundary clarification.

To overcome the ceiling imposed by non-parametric retrieval, we introduce *Next Demonstration Prediction (NDP)*. Building upon recent efforts that utilize feedback loops Zhang et al. (2022; 2025) and Reinforcement Learning (RL) Wang et al. (2024b), NDP reformulates demonstration selection as a sequential Markov Decision Process. We hypothesize that optimal context construction is task-dependent: regression tasks require a diverse, scale-defining set of examples; classification requires hard-negative boundary examples; and Visual Question Answering (VQA) requires a balance of visual relevance and logical diversity to prevent heuristic copying.

As illustrated in Fig. 1, NDP trains a Proximal Policy Optimization (PPO) agent to construct demonstration sets iteratively. An autoregressive decoder cross-attends to Perceiver-encoded candidate latents and optimizes a scalar reward tied to the downstream VLM's predictive accuracy.

Our primary contributions are as follows:

- We formulate K-shot visual demonstration selection as a Markov Decision Process and train a Perceiver encoder plus autoregressive decoder with PPO on the frozen VLM's task reward, after a Maximal Marginal Relevance (MMR) behavioral-cloning warmup.
- We evaluate on 14 datasets spanning regression, classification, and VQA. The 14-dataset comparison is reported for Gemma 3 4B-it; Phi-3.5-vision and Qwen 3 are probed on a

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.

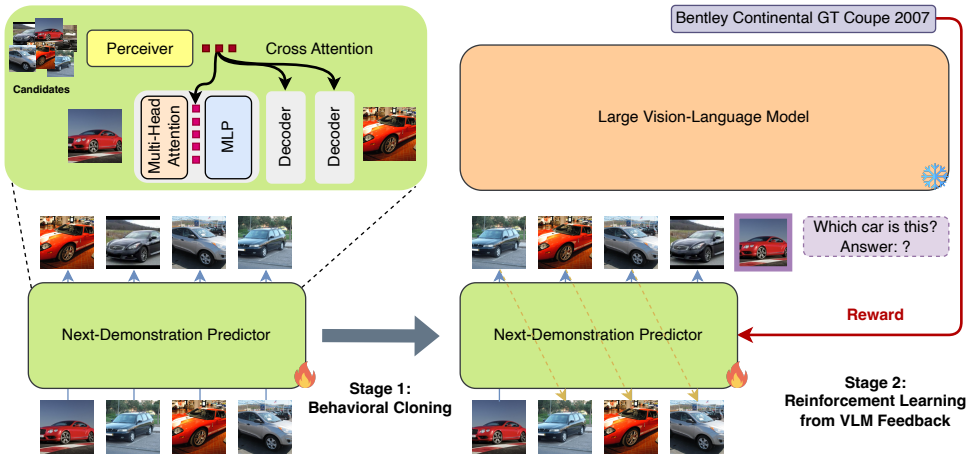


Figure 1: **NDP overview.** A Perceiver compresses the candidate pool; an autoregressive decoder selects the next demonstration. Stage 1 clones a heuristic teacher (MMR). Stage 2 applies PPO: the frozen VLM scores the finished prompt and returns a terminal task reward.

three-task slice (UTKFace, EmoSet, VQAv2) under in-domain training and Gemma-to-target transfer.

- We show that the learned order is not an unordered set: reversing or shuffling the same images degrades in-distribution accuracy. Training requires many VLM rollouts, and zero-shot cross-dataset transfer can collapse below k-NN, MMR, DPP, or even random selection.

2 Related Work

ICL is volatile: accuracy moves with the choice and permutation of demonstrations Zhang et al. (2023b); Liu et al. (2024). Robust selection matters in high-stakes settings such as histopathology classification Ferber et al. (2024). It also affects data engineering pipelines Wang et al. (2021); Khorashadizadeh et al. (2023); Ding et al. (2022), knowledge editing De Cao et al. (2021), model augmentation Ram et al. (2023), safety Panda et al. (2023); Meade et al. (2023), and subjective tasks such as sentiment analysis Zhang et al. (2023a); Xu et al. (2024); Wang et al. (2024a); Yang et al. (2024).

The usual mitigation is k-NN retrieval on the query Liu et al. (2021); Tanwar et al. (2023); Qin et al. (2023). k-NN is computationally trivial but has no task-aware supervision and is often sub-optimal Rubin et al. (2021); Ye et al. (2023a); Wang et al. (2023); Zhang et al. (2022). Similarity-first ranking narrows the reasoning context Wu et al. (2025) and yields redundant clusters rather than complementary evidence Li et al. (2024). Other construction choices include ordering by complexity Liu et al. (2021; 2024), enforcing structural conciseness An et al. (2023), and synthesizing pseudo-demonstrations with LLMs Kim et al. (2022); Yang et al. (2023a); Hao et al. (2022); Yang et al. (2023c). In the multimodal setting, retrieval-reranking pipelines Zhou et al. (2024) and representativeness metrics Guo et al. (2024) have been proposed. Global subset selection includes Determinantal Point Processes that trade quality against diversity Yang et al. (2023c) and multi-armed bandit formulations Purohit et al.. Maximizing structural diversity is important for complex reasoning Xiao et al. (2025). Broader prompting paradigms, including any-shot training Doveh et al. (2024) and contrastive reasoning bounds Yao (2024), likewise argue for contrasting boundaries rather than only similar positives. For VLMs, Mixed Modality In-Context Example Selection (MMICES) shortlists by vision and re-ranks with text-image similarity Chen et al. (2025a). MMICES is training-free and explicitly multimodal; it is the most relevant VLM-specific heuristic we do not re-run. Our main tables freeze an image-only feature space so that k-NN, MMR, DPP, IterR, LSD, and NDP differ in how they use the same candidates, not in which modality they index. That protocol is fair for regression and classification, where the label is a scalar or class name. It is weaker for VQA, where question text is available and MMICES would use it.

A complementary line of work moves from static retrieval to sequential information gain, in the spirit of hard-negative mining Li et al. (2019); Margatina et al. (2023). Dense retrievers can be trained with supervised contrastive learning Zhang et al. (2023b) or generative preference learning Zhang et al. (2025). Li introduced SabER to synthesize prompts autoregressively via distillation Li. Chen et al. formulated retrieval as a Markov Decision Process with a Gated Recurrent Unit Chen et al. (2024). Lee et al. train LSD, a Dueling DQN with a query-centric decoder, to build demonstration sets for visual regression Lee et al. (2026). Those sequential methods often distill token likelihoods or, in LSD, optimize a dense marginal MAE reward on regression only. Methods that instead edit the VLM, such as Dynamic Attention Reallocation (DARA) Chen et al. (2025b), change the model rather than the prompt and are complementary rather than interchangeable baselines. We include IterR as the GRU sequential baseline on the same cached visual embeddings, without VLM token likelihoods, and LSD as the DQN sequential baseline on the five regression tasks. NDP uses an actor-critic trained on a direct task-specific scalar reward (e.g. negative MAE), a Perceiver bottleneck over large pools, and an autoregressive Transformer decoder. This decouples the selection policy from the internal architecture of the target VLM.

3 METHOD

We formulate the problem of selecting in-context demonstrations as a sequential decision-making process. Unlike heuristic approaches that rely solely on fixed similarity metrics (e.g., k-NN), our framework, *Next Demonstration Prediction* (NDP), utilizes a parameterized policy to actively construct a sequence of K demonstrations.

Our training pipeline consists of two distinct phases: 1) Behavioral Cloning (BC) Warm-up, where the policy is pre-trained to mimic a heuristic teacher algorithm (Maximal Marginal Relevance) to acquire a foundational understanding of relevance and diversity; and 2) Reinforcement Learning (RL) Fine-tuning, where the policy is optimized via Proximal Policy Optimization (PPO) to maximize the ground-truth performance of a black-box Vision-Language Model (VLM).

3.1 PROBLEM FORMULATION

Given a query image I_q with an unknown label y_q and a candidate pool of labeled examples $\mathcal{C} = \{(I_i, y_i)\}_{i=1}^{N}$, our goal is to select an ordered sequence of K demonstrations $\mathcal{S} = \{d_1, d_2, \ldots, d_K\}$ where $d_k \in \mathcal{C}$. We model this as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$. The state s_t at step t consists of the query embedding $\mathbf{e}_q$, the embeddings of the current candidate pool $\mathbf{E}_\mathcal{C} \in \mathbb{R}^{N \times D}$, and the history of previously selected demonstrations $\{d_1, \ldots, d_{t-1}\}$. The action a_t is the selection of a specific index i from the candidate pool $\mathcal{C}$, corresponding to the next demonstration d_t. The transition updates the state deterministically by appending the selected d_t to the history and masking it from future selection to prevent duplicates. We employ a sparse, terminal reward structure r_t. The reward is zero for $t < K$. At step K, the selected sequence and query are fed into the VLM, and the reward is task dependent, e.g. negative L1 error of the prediction: $r_K = -|y_{pred} - y_{gt}|$ for regression tasks.

3.2 MODEL ARCHITECTURE

To efficiently handle large candidate pools and model the dependencies between the query and the selected sequence, we propose an Actor-Critic architecture composed of two primary modules:

3.2.1 PERCEIVER ENCODER (CONTEXT COMPRESSION).

Processing all candidate embeddings directly in the policy network is computationally prohibitive. We employ a Perceiver-style encoder Jaegle et al. (2021) to compress the variable-sized candidate pool $\mathbf{E}_\mathcal{C}$ into a fixed set of latent vectors $\mathbf{Z} \in \mathbb{R}^{L \times D}$ ($L = 64$). The encoder uses learnable latent queries and alternating Cross-Attention (Latents $\leftrightarrow$ Candidates) and Self-Attention layers. This results in a permutation-invariant, compact representation of the available demonstration pool.

3.2.2 AUTOREGRESSIVE SELECTOR (POLICY NETWORK).

The policy $\pi_\theta(a_t|s_t)$ is parameterized by a Transformer Decoder. It takes as input a sequence consisting of the query embedding followed by the embeddings of the previously selected demonstrations: $[\mathbf{e}_q, \mathbf{e}_{d_1}, \ldots, \mathbf{e}_{d_{t-1}}]$. The decoder layers employ 1) Causal Self-Attention to model the sequential relationship between the query and selected shots, utilizing Rotary Positional Embeddings (RoPE) to capture relative positions effectively; and 2) Cross-Attention to condition the selection on the available candidates, allowing the decoder to cross-attend to the global context latents $\mathbf{Z}$ produced by the Perceiver Encoder. The final hidden state $\mathbf{h}_t$ is projected to the embedding dimension D. The logits for the action distribution are computed via the dot product between $\mathbf{h}_t$ and the candidate embeddings $\mathbf{E}_\mathcal{C}$:

$$P(a_t = i|s_t) = \text{Softmax}(\mathbf{h}_t \cdot \mathbf{E}_{\mathcal{C}_i}^T)$$

3.2.3 CRITIC NETWORK.

The Value function $V_\phi(s_t)$ shares the same architectural design as the Actor (Perceiver + Decoder) but concludes with a scalar projection head instead of a dot-product action head. This allows the Critic to estimate the expected return of the current partial sequence relative to the global candidate context.

3.3 TRAINING STRATEGY

3.3.1 STAGE 1: BEHAVIORAL CLONING (BC) VIA MMR

Randomly initializing the policy for RL can lead to instability and sparse rewards. To mitigate this, we first warm up the policy using Behavioral Cloning. We treat Maximal Marginal Relevance (MMR) as a "teacher" algorithm. MMR selects examples that are relevant to the query while penalizing redundancy.

Selection problem. Let q denote the query and let $\mathcal{C} = \{(I_i, y_i)\}_{i=1}^N$ be the labeled candidate pool. The selector returns an ordered sequence of distinct indices $(a_1, \ldots, a_K)$, which determines the demonstration prompt. For VQA, the question text is included in the VLM prompt but is not used by the image-only selection policy.

$$a_t^{\text{MMR}} = \underset{i \notin \{a_1,\ldots,a_{t-1}\}}{\arg\max} \left[\lambda \, \text{sim}(e_i, e_q) - (1-\lambda) \max_{j<t} \text{sim}(e_i, e_{a_j}) \right]. \tag{1}$$

where λ controls the trade-off between relevance and diversity (set to 0.5).

During BC training, we sample a random subset of the dataset $\mathcal{C}_{sub} \subset \mathcal{C}$ (size N_{sub}) as the context. We compute the target sequence indices $y^*_{1:K}$ using the MMR heuristic on this subset. The policy is trained to minimize the Cross-Entropy loss between its predicted logits and the MMR targets:

$$\mathcal{L}_{BC} = -\sum_{t=1}^{K} \log P(a_t = y_t^*|s_t, \mathcal{C}_{sub}) \tag{2}$$

This phase creates a stable policy that already balances similarity and diversity before any interaction with the VLM.

3.3.2 STAGE 2: REINFORCEMENT LEARNING (RL) VIA PPO

After BC pre-training, we fine-tune the model using Proximal Policy Optimization (PPO) Schulman et al. (2017) to optimize for the true objective: the VLM's predictive accuracy. This phase allows the agent to deviate from the MMR heuristic and discover task-specific strategies that minimize regression error.

Trajectory Collection. During training, we perform rollouts by sampling K steps from the policy π_θ. We construct a prompt using the selected images and the query, which is processed by the frozen VLM environment (Gemma-3-4b-it). The VLM generates a numerical response, which is parsed to compute the terminal reward r_K.

Generalized Advantage Estimation (GAE). We compute advantages $\hat{A}_t$ using GAE(λ) Schulman et al. (2015) to balance bias and variance. Since our reward is sparse (occurring only at step K), the value function V_ϕ plays a critical role in assigning credit to earlier selection steps. The advantage at time t is given by:

$$\hat{A}_t = \sum_{l=0}^{K-1-t} (\gamma\lambda)^l \delta_{t+l}, \quad \text{where } \delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \tag{3}$$

Objective Function. PPO optimizes a clipped surrogate objective to ensure monotonic improvement without destructively large policy updates. Let $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ denote the probability ratio between the new and old policies. The policy loss is defined as:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right] \tag{4}$$

where ϵ is a hyperparameter (set to 0.2) controlling the clip range.

The Critic network is updated by minimizing the Mean Squared Error (MSE) between the predicted value $V_\phi(s_t)$ and the computed return $R_t = \hat{A}_t + V_{\phi_{old}}(s_t)$:

$$L^{VF}(\phi) = \mathbb{E}_t \left[(V_\phi(s_t) - R_t)^2 \right] \tag{5}$$

The final joint objective function maximizes the policy return while minimizing value error and maximizing entropy $S[\pi_\theta]$ to encourage exploration:

$$\mathcal{L}_{RL} = L^{CLIP}(\theta) - c_1 L^{VF}(\phi) + c_2 S[\pi_\theta](s_t) \tag{6}$$

where c_1 and c_2 are coefficients weighting the value loss and entropy bonus, respectively.

4 Experiments

To demonstrate the effectiveness of *Next Demonstration Prediction* (NDP), we evaluate across 14 visual reasoning tasks, compare against heuristic and learned selectors, and isolate the contribution of the PPO stage relative to its MMR teacher.

4.1 Setup

Models. We report the 14-benchmark comparison on Gemma 3 4B-it Team et al. (2025). Table 2 is a separate three-task cross-model probe (UTKFace, EmoSet, VQAv2) using Phi-3.5-vision (4.2B) Abdin et al. (2024) and Qwen 3 7B Bai et al. (2025). Its "In-Domain" rows use policies trained specifically with each target VLM's reward; its "Transfer" rows evaluate a frozen Gemma-trained policy on the target VLM without updating that policy. Cross-dataset transfer is evaluated separately in Table 3. Rollouts use vLLM with FlashAttention and a 20-token generation cap.

Datasets. Fourteen benchmarks in three groups: fine-grained regression (UTKFace Zhang et al. (2017), AVA Murray et al. (2012), SCUT-FBP5500 Liang et al. (2018), KonIQ-10k Hosu et al. (2020), KADID-10k Lin et al. (2019)), few-shot classification (EmoSet Yang et al. (2023b), CIFAR-10 Krizhevsky et al. (2009), MNIST Deng (2012), Stanford Cars Krause et al. (2013), Food-101 Bossard et al. (2014)), and VQA (VQAv2 Antol et al. (2015), OK-VQA Marino et al. (2019), GQA Hudson & Manning (2019), TextVQA Singh et al. (2019)). Candidate image features are cached before RL.

Implementation. $D = 768$, 2-layer Perceiver with 64 latents, 8-layer decoder with 8 heads. PPO uses 8 trajectories per rollout, clip $\epsilon = 0.2$, entropy 0.01, value coefficient 0.5, AdamW, actor 10^{-6} (classification/VQA) or 10^{-5} (regression), critic 5×10^{-5}. Classification and VQA use 6,000 steps; regression uses 8,000. AVA curves in Fig. 2(a) are logged to 13,000 steps. Further details are in Appendix A. Unless noted, $K = 16$.

Table 1: **Gemma 3 4B-it, 14 benchmarks ($K = 16$).** NDP has the best average regression MAE and the best average VQA accuracy among selectors with a full row, and is behind DPP on classification. LSD Lee et al. (2026) reports regression only. Best in **bold**, second underlined.

	Regression (MAE ↓)						Classification (Acc. ↑)					
Method	**UTK**	**AVA**	**SCUT**	**KonIQ**	**KADID**	**Avg.**	**EmoSet**	**Cars**	**Food**	**CIFAR**	**MNIST**	**Avg.**
Zero-Shot	**5.91**	2.55	0.65	**0.36**	0.65	2.02	59.8	50.0	74.2	88.9	75.8	69.7
Random	11.52	1.50	0.74	0.42	0.74	2.98	52.5	54.9	74.4	81.9	87.4	70.2
k-NN	6.46	1.39	**0.41**	0.42	**0.41**	1.82	59.0	76.2	61.3	87.0	95.5	75.8
MMR Ye et al. (2023b)	8.06	1.34	0.43	0.42	0.43	2.14	73.5	**80.5**	76.4	87.0	92.5	82.0
IterR Chen et al. (2024)	6.66	0.79	0.48	0.46	0.48	1.77	53.5	61.0	76.5	86.5	82.0	71.9
DPP Levy et al. (2023)	12.32	1.28	0.65	0.41	0.65	3.06	72.5	80.1	**78.8**	**89.5**	**97.5**	**83.7**
LSD Lee et al. (2026)	6.64	0.98	0.75	0.51	0.84	1.94	–	–	–	–	–	–
NDP (Ours)	6.53	**0.72**	**0.41**	0.39	**0.41**	**1.69**	**74.5**	78.3	73.5	89.0	90.3	81.1

VQA (Acc. ↑)	**Zero**	**Random**	**k-NN**	**MMR**	**IterR**	**DPP**	**NDP**
VQAv2	57.2	55.3	**58.0**	54.0	57.5	54.2	54.2
OK-VQA	29.2	38.0	37.0	38.3	33.7	35.7	**38.7**
GQA	50.5	48.0	**54.5**	54.0	51.0	53.0	**54.5**
TextVQA	65.2	68.3	62.5	67.5	66.0	65.7	**69.2**
Avg.	50.5	52.4	53.0	53.5	52.1	52.2	**54.2**

Baselines. **Zero-shot:** query and instruction only. **Random:** K draws from the pool. **Visual k-NN:** K nearest neighbors by cosine similarity. **MMR** Ye et al. (2023b): relevance vs. redundancy ($\lambda = 0.5$), and the BC teacher. **IterR** Chen et al. (2024): sequential MDP retrieval with a GRU on the same cached embeddings. **DPP** Levy et al. (2023): diversity-oriented subset selection. **LSD** Lee et al. (2026): Dueling DQN sequential selector. LSD reports only the five regression tasks; those $K = 16$ MAEs are included in Table 1, with classification and VQA left unmarked. We do not evaluate MMICES Chen et al. (2025a) or another question-aware VQA selector. The resulting comparison holds the cached visual candidate representation fixed across the reported selectors, but does not establish superiority over methods that also retrieve using question text. This distinction is particularly important for VQA, where the question is available at inference.

Cost. NDP trains a selector while keeping the VLM frozen; it is not training-free ICL. Each sampled trajectory selects K shots and requires one VLM generation to compute its terminal reward. With eight trajectories per rollout, the reported 6,000- and 8,000-step configurations correspond to approximately 48,000 and 64,000 training generations per policy, respectively, if each reported step denotes one rollout batch. These counts exclude behavioral cloning, validation, hyperparameter exploration, and repeated runs. The in-domain Phi-3.5 and Qwen 3 policies in Table 2 also require training against their respective target VLMs; evaluating a frozen Gemma policy on those models does not require target-policy updates. Candidate image features are cached. At inference, NDP encodes $N = 8{,}192$ cached vectors into 64 latents and performs K decoding steps before VLM inference. We did not measure matched GPU-hours, peak memory, selector latency, or a break-even query count. Generation counts alone therefore cannot establish that NDP is cheaper than heuristic retrieval or VLM adaptation.

4.2 Main results

Table 1 shows task-dependent gains. NDP has the lowest reported mean regression MAE (1.69 versus 1.77 for IterR), driven mainly by AVA (0.72 versus 0.79), but trails zero-shot on UTKFace and KonIQ. It has the highest mean VQA accuracy among the evaluated image-only selectors (54.2), helped by TextVQA (69.2), yet loses to k-NN on VQAv2 (54.2 versus 58.0). On classification, DPP leads on average (83.7 versus 81.1 for NDP). Thus, the aggregate results do not imply consistent gains across datasets.

The NDP vs. MMR gap is the PPO stage. MMR ($\lambda = 0.5$) supplies the behavioral-cloning targets, so NDP begins with an explicit relevance–diversity bias. The final NDP policy differs from the independently evaluated MMR heuristic on AVA (0.72 vs. 1.34 MAE), EmoSet (74.5 vs. 73.5 accuracy), and TextVQA (69.2 vs. 67.5 accuracy). This comparison does not isolate the contribution of PPO:

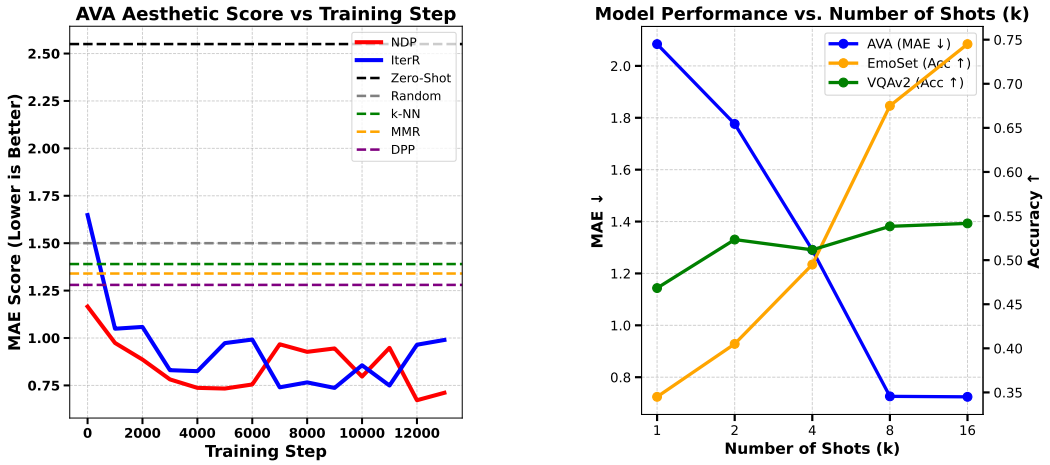


Figure 2: **Training and shot-count trends.** (a) AVA validation MAE during training. (b) AVA, EmoSet, and VQAv2 performance at $K \in \{1, 4, 8, 16\}$.

Table 2: **Cross-model probe** ($K = 16$)**.** Age = UTKFace MAE, Emo = EmoSet accuracy, VQA = VQAv2 accuracy (not a VQA average). In-domain NDP is separately trained with each target VLM's reward; Transfer evaluates a frozen Gemma-trained policy on the target VLM. NDP is not uniformly best.

	Gemma3			Phi3.5			Qwen3		
Method	*Age (↓)*	*Emo (↑)*	*VQA (↑)*	*Age (↓)*	*Emo (↑)*	*VQA (↑)*	*Age (↓)*	*Emo (↑)*	*VQA (↑)*
Zero-Shot	**5.91**	59.80	57.20	**5.35**	12.00	**75.00**	**5.49**	82.00	82.00
Random	11.52	52.50	55.30	29.18	14.00	57.20	8.37	81.70	81.70
k-NN	6.46	59.00	**58.00**	10.69	13.00	68.20	6.98	**83.50**	**83.50**
MMR Ye et al. (2023b)	8.06	73.50	54.00	14.94	50.50	61.80	7.64	78.50	78.50
IterR Chen et al. (2024)	6.66	53.50	57.50	13.44	17.50	61.20	5.95	61.00	**83.50**
DPP Levy et al. (2023)	12.32	**74.50**	54.20	26.65	56.50	61.50	8.64	78.00	81.50
NDP (In-Domain)	6.53	**74.50**	54.20	7.28	**62.50**	64.00	6.28	80.30	80.30
NDP (Transfer)†	-	-	-	8.00	58.50	61.80	5.98	77.20	77.20

† Frozen policy trained entirely on Gemma3.

we do not report evaluation of the behavioral-cloning checkpoint before RL. We therefore attribute these differences to the complete NDP pipeline, not to PPO alone. PPO uses a GAE critic and an entropy bonus (0.01); we do not use intermediate diversity rewards.

4.3 Training dynamics and K

Fig. 2(a) tracks AVA MAE during RL. Static heuristics are flat; NDP improves through those floors and through IterR by optimizing the VLM's terminal reward.

Fig. 2(b) and Appendix E vary $K \in \{1, 4, 8, 16\}$ on one dataset per group. At $K = 1$ or $K = 4$, relevance dominates and k-NN is already strong. As K grows, distance-based retrieval keeps fetching near-duplicates, while NDP uses the extra slots for scale and boundary examples: AVA MAE falls, EmoSet and VQAv2 accuracy rise. We treat $K \in \{8, 16\}$ as the operating regime and do not recommend NDP under a 1- or 4-shot cap.

4.4 Cross-model transfer

Table 2 uses two settings: in-domain (policy trained on the same VLM) and transfer (Gemma reward, frozen policy, Phi-3.5 or Qwen 3 as the environment). Target models differ in architecture and scale so the policy is not scored only on Gemma's routing.

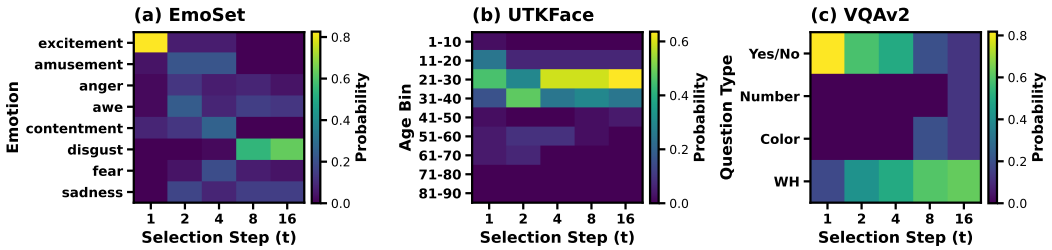


Figure 3: Temporal density of selected attributes at each step t ($K = 16$). (a) Excitement: target class, then related emotions, then contrastive boundaries. (b) Age 30 to 40: nearby bins, then a global scale. (c) "Is the ..." VQA: same question format, then counting, attribute, and object questions.

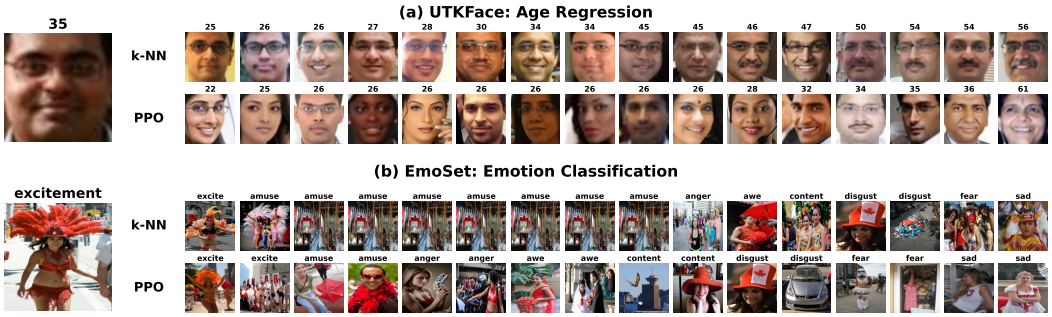


Figure 4: **Selected demonstrations.** (a) Age query 35: k-NN stays in a similar-age cluster; NDP spans young to old (up to 81) as a reference scale. (b) Excitement: k-NN stays in amusement/excitement; NDP covers a contrastive emotion set.

In-domain NDP is strongest on EmoSet for Gemma (74.50, tied with DPP in this table) and Phi-3.5 (62.50, 6.0 above the next baseline). Zero-shot and k-NN remain competitive on Age and VQAv2. ICL can hurt: Phi-3.5 VQAv2 zero-shot is 75.00 vs. 64.00 in-domain NDP. Gemma-to-Phi EmoSet transfer (58.50) still beats MMR (50.50) and DPP (56.50). On Qwen 3, the transferred policy is slightly better than in-domain NDP on Age (5.98 vs. 6.28). That is reusable visual utility, not proof that Gemma's order is optimal for Phi or Qwen. K-shot transfer curves are in Appendix D.1.

4.5 Qualitative Selection Dynamics

Fig. 3 unrolls the decoder for emotion recognition, age prediction, and VQA (question type relative to an "Is the ..." query). Three phases appear. (1) Anchoring ($t = 1$): mass on the query attribute (same emotion, age bin, or question format). (2) Boundary exploration ($t = 2$ to 4): nearby classes or formats with different semantics, which tightens classification boundaries and exposes VQA template variation. (3) Global scale ($t \geq 5$): the distribution flattens toward extremes or other reasoning structures.

Fig. 4 compares k-NN and NDP on the same queries. k-NN returns near-duplicates. NDP spans the label distribution: for age 35, a scale from young adults to the elderly; for excitement, several emotional states rather than an amusement cluster. Additional grids are in Appendix F.

4.6 OOD Transfer and Permutation

Table 3 splits two claims. Cross-dataset transfer is a negative result: KonIQ-10k $\rightarrow$ KADID-10k, EmoSet $\rightarrow$ Food-101, and VQAv2 $\rightarrow$ OK-VQA all lose to heuristics, and classification transfer loses to random (70.10 vs. 74.40). Zero-shot (0.79) beats NDP (0.94) on the IQA shift; DPP (78.80) dominates Food-101 transfer. The policy fits source prompt structure (quality scales, emotion boundaries, VQAv2 templates), not a universal selection rule. Retrain in-domain, or use MMR/DPP, when the query distribution shifts.

Table 3: **OOD transfer vs. permutation of the same shots.** Top: train on source, test on target (NDP loses to heuristics). Bottom: in-distribution predicted order vs. shuffle or reverse of the same images.

Method	Regression (MAE ↓) *KonIQ-10k → KADID-10k*	Classification (Acc. ↑) *EmoSet → Food-101*	VQA (Acc. ↑) *VQAv2 → OK-VQA*
Out-of-distribution transfer (trained on source, evaluated on target)			
Zero-Shot	**0.79**	74.20	29.20
Random	1.04	74.40	<u>38.00</u>
k-NN	<u>0.80</u>	61.30	37.00
MMR Ye et al. (2023b)	0.81	<u>76.40</u>	**38.30**
DPP Levy et al. (2023)	1.00	**78.80**	35.70
NDP (Zero-Shot Transfer)	0.94	70.10	34.60
Permutation sensitivity (trained on target, evaluated on target)			
NDP (Reversed Order)	0.83	59.70	<u>38.30</u>
NDP (Random Shuffle)	<u>0.80</u>	<u>67.00</u>	35.80
NDP (Predicted Order)	**0.62**	**73.50**	**38.70**

Permutation of the *same* in-distribution images is the opposite finding. Predicted order is best (regression MAE 0.62; Food-101 73.50; OK-VQA 38.70). Random shuffle drops Food-101 to 67.00 and VQA to 35.80. Reverse is worst on regression (0.83) and classification (59.70); VQA is slightly more robust to reverse (38.30) than to shuffle. That matches the heatmap curriculum (anchor, then boundary, then scale) and is why we model selection as an ordered MDP.

4.7 Interpretation and limitations

MMR warm start. MMR supplies NDP's behavioral-cloning targets and is also evaluated as a standalone heuristic in Table 1. NDP improves on MMR on some tasks, including AVA and TextVQA, but not uniformly. Because we do not evaluate the BC-only checkpoint, this comparison does not isolate the contribution of PPO.

Baseline scope. The main comparisons use cached image embeddings ($D = 768$): k-NN ranks by cosine similarity; MMR and DPP use pairwise similarities; and IterR is adapted as a sequential policy over the same candidate pool. LSD reports results on regression only. Question text is not used for retrieval, so the VQA results do not establish an advantage over question-aware selectors.

Order and use cases. NDP receives a terminal reward after K selections. On the tested in-domain tasks, its predicted order outperforms permutations of the same shots (Table 3), although this does not establish that PPO is necessary for the effect. Its training cost is most defensible when many queries share a domain, $K \geq 8$, and validation shows a gain over inexpensive heuristics. For small shot budgets or the examined dataset shifts, the results favor validating k-NN, MMR, or DPP rather than assuming NDP will transfer.

5 Conclusion

We framed visual demonstration selection as next-shot prediction: a small policy, cloned from MMR then trained with PPO, builds an ordered prompt for a frozen VLM. On Gemma 3 4B-it, NDP wins average regression and average VQA among image-only selectors, is behind DPP on classification, and helps mainly at $K \geq 8$ on scale and boundary tasks. The predicted order matters in-distribution. Frozen Gemma-to-target transfer helps in some cross-model settings but is not uniformly best. On the examined cross-dataset shifts, zero-shot policy transfer underperforms simpler selectors. NDP trains a selector rather than a training-free retriever, and the VQA index ignores question text. Use NDP when many queries share a domain and a moderate context budget; otherwise k-NN, MMR, or DPP remain the default.

AI USE STATEMENT

We used generative AI tools to assist with implementing and editing software code, drafting and revising manuscript text, and improving its readability. We also used these tools to obtain feedback on the presentation of our methodology and experiments and on the interpretation of reported results, including the scope and limitations of our comparisons. The authors reviewed AI-assisted text, code, and suggestions, checked the manuscript against the underlying experiments, and take responsibility for the final methods, results, claims, and submitted content.

REPRODUCIBILITY STATEMENT

Section 3 describes NDP and its training procedure, and Section 4.1 specifies the evaluation setup and baselines. Appendices A, B, and C provide implementation details, hyperparameters, and task-specific prompts and rewards. We provide code as anonymous supplementary material with this submission to support replication of the reported experiments. We plan to release the code publicly on GitHub if the paper is accepted.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. How do in-context examples affect compositional generalization? *arXiv preprint arXiv:2305.04835*, 2023.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can multimodal large language models truly perform multimodal in-context learning? in 2025 ieee/cvf winter conference on applications of computer vision (wacv). *IEEE*, 2:6000–6010, 2025a.

Shuo Chen, Jianzhe Liu, Zhen Han, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. True multimodal in-context learning needs attention to the visual context. *arXiv preprint arXiv:2507.15807*, 2025b.

Yunmo Chen, Tongfei Chen, Harsh Jhamtani, Patrick Xia, Richard Shin, Jason Eisner, and Benjamin Van Durme. Learning to retrieve iteratively for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7156–7168, 2024.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*, 2022.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Sivan Doveh, Shaked Perek, M Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision and language models. In *European Conference on Computer Vision*, pp. 250–267. Springer, 2024.

Dyke Ferber, Georg Wölflein, Isabella C Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar SM El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1):10104, 2024.

Zhaojun Guo, Jinghui Lu, Xuejing Liu, Rui Zhao, ZhenXing Qian, and Fei Tan. What makes good few-shot examples for vision-language models? *arXiv preprint arXiv:2405.13532*, 2024.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-context learning to 1,000 examples. *arXiv preprint arXiv:2212.06713*, 2022.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Hanieh Khorashadizadeh, Nandana Mihindukulasooriya, Sanju Tiwari, Jinghua Groppe, and Sven Groppe. Exploring in-context learning capabilities of foundation models for generating knowledge graphs from text. *arXiv preprint arXiv:2305.08804*, 2023.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Eugene Lee, Yu-Chi Lin, and Jiajie Diao. Learning to select visual in-context demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9455–9465, June 2026.

Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1401–1422, 2023.

Chuanhao Li, Chenchen Jing, Zhen Li, Mingliang Zhai, Yuwei Wu, and Yunde Jia. In-context compositional generalization for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17954–17966, 2024.

Meng Li, Lin Wu, Arnold Wiliem, Kun Zhao, Teng Zhang, and Brian Lovell. Deep instance-level hard negative mining model for histopathology images. In *International conference on medical image computing and computer-assisted intervention*, pp. 514–522. Springer, 2019.

Yanshu Li. Advancing multimodal in-context learning in large vision-language models with task-aware demonstrations. In *Workshop on Reasoning and Planning for Large Language Models*.

Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *2018 24th International conference on pattern recognition (ICPR)*, pp. 1598–1603. IEEE, 2018.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*, 2024.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. Active learning principles for in-context learning with large language models. *arXiv preprint arXiv:2305.14264*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tür. Using in-context learning to improve dialogue safety. *arXiv preprint arXiv:2302.00871*, 2023.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

Ashwinee Panda, Tong Wu, Jiachen Wang, and Prateek Mittal. Differentially private in-context learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Kiran Purohit, V Venktesh, Sourangshu Bhattacharya, and Avishek Anand. Sample efficient demonstration selection for in-context learning. In *Forty-second International Conference on Machine Learning*.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. *arXiv preprint arXiv:2310.09881*, 2023.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual llms are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Qianlong Wang, Hongling Xu, Keyang Ding, Bin Liang, and Ruifeng Xu. In-context example retrieval from multi-perspectives for few-shot aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8975–8985, 2024a.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*, 2021.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. In *Workshop on efficient systems for foundation models@ icml2023*, 2023.

Xubin Wang, Jianfei Wu, Yichen Yuan, Deyu Cai, Mingzhe Li, and Weijia Jia. Demonstration selection for in-context learning via reinforcement learning. *arXiv preprint arXiv:2412.03966*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wenxiao Wu, Jing-Hao Xue, Chengming Xu, Chen Liu, Xinwei Sun, Changxin Gao, Nong Sang, and Yanwei Fu. Towards reliable and holistic visual in-context learning prompt selection. *arXiv preprint arXiv:2509.25989*, 2025.

Wenyang Xiao, Haoyu Zhao, and Lingxiao Huang. The role of diversity in in-context learning for large language models. *arXiv preprint arXiv:2505.19426*, 2025.

Hongling Xu, Qianlong Wang, Yice Zhang, Min Yang, Xi Zeng, Bing Qin, and Ruifeng Xu. Improving in-context learning with prediction feedback for sentiment analysis. *arXiv preprint arXiv:2406.02911*, 2024.

Jinghan Yang, Shuming Ma, and Furu Wei. Auto-icl: In-context learning without human supervision. *arXiv preprint arXiv:2311.09263*, 2023a.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20383–20394, 2023b.

Li Yang, Zengzhi Wang, Ziyan Li, Jin-Cheon Na, and Jianfei Yu. An empirical study of multimodal entity-based sentiment analysis with chatgpt: Improving in-context learning via entity-aware contrastive learning. *Information Processing & Management*, 61(4):103724, 2024.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5443–5456, 2023c.

Liang Yao. Large language models are contrastive reasoners. *arXiv preprint arXiv:2403.08211*, 2024.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023a.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4469–4484, 2023b.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023a.

Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*, 2022.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794, 2023b.

Zheng Zhang, Shaocheng Lan, Lei Song, Jiang Bian, Yexin Li, and Kan Ren. Learning to select in-context demonstration preferred by large language model. *arXiv preprint arXiv:2505.19966*, 2025.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024.

Yan Zhu, Huan Ma, and Changqing Zhang. Exploring task-level optimal prompts for visual in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11031–11039, 2025.

A Implementation and Training Details

A.1 Environment and Infrastructure

All Reinforcement Learning (RL) training and evaluation were conducted using the `vllm` library for high-throughput, asynchronous generation (`AsyncVLLMClient`). To optimize memory overhead and ensure rapid trajectory collection, the target Vision-Language Model (VLM), `google/gemma-3-4b-it`, was served using the `enforce_eager=True` flag with a GPU memory utilization threshold of 0.8. The VLM generated a maximum of 20 new tokens per step to strictly capture numerical and categorical outputs.

A.2 Proximal Policy Optimization (PPO) Hyperparameters

The Next Demonstration Prediction (NDP) policy was trained with distinct step counts and learning rates depending on the target domain to ensure optimal convergence. Regression tasks were trained for 8,000 steps with an Actor learning rate of 10^{-5}, while Classification and VQA tasks were trained for 6,000 steps with a lower Actor learning rate of 10^{-6}. The Critic learning rate was fixed at 5×10^{-5} across all modalities.

At each rollout step, the policy selected demonstrations from a randomly sampled candidate pool subset of $N = 8,192$ (to ensure tractability while maintaining a diverse sampling frame). We used Generalized Advantage Estimation (GAE) to compute stable advantage targets.

Table 4: **PPO and Training Hyperparameters.**

Hyperparameter	**Value**
Trajectories per Rollout (Batch Size)	8
PPO Update Epochs per Batch	4
Actor Learning Rate	10^{-6} (Class/VQA), 10^{-5} (Regression)
Critic Learning Rate	5×10^{-5}
Total Training Steps	6,000 (Class/VQA), 8,000 (Regression)
PPO Clip Epsilon (ϵ)	0.2
Discount Factor (γ)	0.99
GAE Lambda (λ)	0.95
Entropy Coefficient	0.01
Value Loss Coefficient	0.5
Maximum Shots (K)	16
Candidate Pool Subset Size (N)	8,192
Optimizer	AdamW

B ARCHITECTURE DETAILS

The Actor and Critic models in the Next Demonstration Prediction (NDP) framework share a highly symmetric foundational architecture, differing primarily at their terminal projection heads. Both networks process a shared embedding space of dimension $D = 768$ and consist of two primary modules: a Perceiver Encoder for global context compression and an Autoregressive Decoder for sequential state modeling. To accelerate reinforcement learning convergence, the Critic network's encoder and decoder modules are initialized with the frozen weights of the Actor's Behavioral Cloning (BC) pre-training phase.

B.1 PERCEIVER ENCODER (CONTEXT COMPRESSION)

The Perceiver Encoder is tasked with compressing the large, variable-sized candidate pool (subset $N = 8,192$ images) into a fixed, computationally tractable latent space. Processing thousands of high-dimensional embeddings directly within a sequential decoder is computationally prohibitive; thus, this module acts as an attention-based bottleneck.

- **Latent Initialization:** The encoder initializes a fixed set of $L = 64$ learnable latent queries, forming a weight matrix $\mathbf{Z} \in \mathbb{R}^{64\times768}$. These latents are initialized from a standard normal distribution and act as the primary information bottleneck.
- **Layer Composition:** The module comprises 2 identical Perceiver blocks. Each block sequentially applies Multi-Head Cross-Attention followed by Multi-Head Self-Attention.
- **Cross-Attention (Latents $\leftarrow$ Candidates):** In this step, the L latents act as queries, while the N candidate pool embeddings act as both keys and values. This allows the latents to extract relevant structural and semantic features from the global dataset.
- **Self-Attention (Latents $\leftrightarrow$ Latents):** Following the cross-attention step, the latents attend to one another to distribute and synthesize the extracted information, resolving redundancies.
- **Attention Configurations:** Both attention mechanisms utilize 8 independent attention heads. To maintain high-fidelity feature extraction across the large candidate pool, dropout is strictly disabled ($p = 0.0$).
- **Feed-Forward Network (FFN):** The attended latents pass through a Multi-Layer Perceptron (MLP) featuring a standard 4x expansion factor ($\mathbb{R}^{768} \to \mathbb{R}^{3072} \to \mathbb{R}^{768}$) paired with a GELU activation function.
- **Pre-Normalization and Residuals:** The architecture employs a Pre-LayerNorm design, applying standard layer normalization immediately before each attention and FFN block, coupled with residual additive connections. The final output is a compressed, permutation-invariant representation of the candidate pool $\mathbf{Z}_{\text{out}} \in \mathbb{R}^{64\times768}$.

B.2 AUTOREGRESSIVE DECODER (POLICY)

The Autoregressive Selector is the core sequential decision-making engine. It models the complex dependencies between the initial query image and the chronological history of previously selected demonstrations to dictate the optimal next action.

- **Sequence Construction:** At time step t, the input sequence is formed by concatenating the query embedding $\mathbf{e}_q$ with the embeddings of the previously selected shots $\{\mathbf{e}_{d_1}, \ldots, \mathbf{e}_{d_{t-1}}\}$. This yields a dynamic sequence length of t (bounded by a maximum $K = 16$ shots, plus 1 for the query).
- **Layer Composition:** The decoder utilizes 8 custom transformer decoder layers, each operating with 8 attention heads.
- **Causal Self-Attention with RoPE:** To strictly enforce the chronological narrative of the prompt construction, the self-attention mechanism applies a causal triangular mask. Absolute positional embeddings are entirely discarded. Instead, Rotary Positional Embeddings (RoPE) are applied directly to the query and key vectors within the self-attention operation. By rotating the vectors in the complex plane according to their sequence position, RoPE allows the model to robustly capture the relative distances and order of the selected shots.

- **Cross-Attention (Sequence ← Latents):** Following causal self-attention, the sequence representations pass through a non-causal Multi-Head Cross-Attention block. Here, the sequential tokens act as queries, while the $L = 64$ global context latents ($\mathbf{Z}_{\text{out}}$ from the Perceiver) act as keys and values. This allows the autoregressive sequence to condition its next selection on the globally available data manifold.
- **Feed-Forward Network (FFN):** A Pre-LayerNorm MLP identical to the Perceiver's design expands and contracts the features ($\mathbb{R}^{768} \to \mathbb{R}^{3072} \to \mathbb{R}^{768}$) with a GELU activation.
- **Final Representation:** A terminal LayerNorm is applied to the sequence, yielding the final hidden states $\mathbf{H} \in \mathbb{R}^{t \times 768}$.

B.3 Action and Value Heads

The Actor and Critic diverge at the final projection step, utilizing the ultimate sequence token $\mathbf{h}_t$ (the hidden state corresponding to the most recent selection) to compute their respective outputs.

- **Actor Head (Policy π_θ):** The Actor treats $\mathbf{h}_t$ as the targeted query vector for the next demonstration. To compute the action logits over the entire candidate pool, a dot product is calculated between $\mathbf{h}_t$ and the transposed subset embeddings matrix $\mathbf{E}_{\mathcal{C}} \in \mathbb{R}^{N \times 768}$:

$$\mathbf{l}_t = \mathbf{h}_t \cdot \mathbf{E}_{\mathcal{C}}^T \tag{7}$$

  To strictly enforce sequence uniqueness and prevent duplicate selections, a boolean mask is applied, setting the logits of already selected candidate indices to $-\infty$. The masked logits are then passed through a Softmax function to parameterize the categorical action distribution:

$$P(a_t = i \mid s_t) = \text{Softmax}(\mathbf{l}_{t,i}) \tag{8}$$

- **Critic Head (Value Function V_ϕ):** The Critic network evaluates the expected return of the current partial sequence. Instead of a dot-product attention mechanism over the candidate pool, the Critic isolates $\mathbf{h}_t$ and projects it through a dense linear layer ($\mathbf{W}_v \in \mathbb{R}^{768 \times 1}$, $\mathbf{b}_v \in \mathbb{R}^1$) to output a continuous scalar value:

$$V(s_t) = \mathbf{W}_v \mathbf{h}_t + \mathbf{b}_v \tag{9}$$

  This scalar value is utilized during Generalized Advantage Estimation (GAE) to compute the advantage targets for the Proximal Policy Optimization (PPO) updates.

C Task-Specific Rewards and Prompts

To decouple the RL agent from the internal token likelihoods of the VLM, we utilized strict string parsing (via Regex) to extract predictions and explicitly defined terminal scalar rewards (r_K) based on downstream task accuracy.

Parsing Penalty: Across all domains, if the VLM output failed to match the expected format (e.g., failing to output a parsable integer or floating-point number), a strict negative penalty of $r_K = -50.0$ was applied to discourage the selection of prompts that induce hallucinatory or heavily conversational outputs.

C.1 Classification Tasks

- **Prompt Template:** `"You are a classifier. Choose EXACTLY ONE label from the provided category list. Output ONLY the label string, with no extra text, punctuation, or explanation. Category list: [CLASSES]"`
- **Reward Function:** Exact match accuracy.

$$r_K = \begin{cases} 1.0 & \text{if exact match} \\ 0.0 & \text{otherwise} \end{cases} \tag{10}$$

C.2 Fine-Grained Regression Tasks

We floor all regression rewards at -100.0 to prevent severe gradient explosions from extreme outliers.

- **Prompt Templates:**
  - *Age Prediction:* `"What is the age of the person in the last image? Only output the estimated age as a number."`
  - *Aesthetic Score:* `"What is the aesthetic score of the last image on a scale from 0 to 10? Only output the score as a floating number."`
  - *Quality / Beauty:* `"What is the [facial beauty / image quality] score of the last image on a scale from 0 to 5? Only output the score as a floating number."`
- **Reward Function:** Scaled negative Mean Absolute Error (MAE).
  - *Age Prediction:*

$$r_K = \max(-|y_{\text{pred}} - y_{\text{gt}}|, -100.0) \tag{11}$$

  - *Aesthetic Score:*

$$r_K = \max(-10 \times |y_{\text{pred}} - y_{\text{gt}}|, -100.0) \tag{12}$$

  - *Facial Beauty / Modified & Wild Image Quality:*

$$r_K = \max(-20 \times |y_{\text{pred}} - y_{\text{gt}}|, -100.0) \tag{13}$$

C.3 Visual Question Answering (VQA) Tasks

- **Prompt Template:** `"Answer the following question using a single word or a short phrase. Question: [QUESTION] Answer:"`
- **Reward Function:** For GQA, exact match (1.0 or 0.0). For VQAv2, OK-VQA, and TextVQA, we use the standard VQA accuracy metric based on human annotations:

$$r_K = \min\left(1.0, \frac{\text{match_count}}{3.0}\right) \tag{14}$$

D Extended Experimental Results

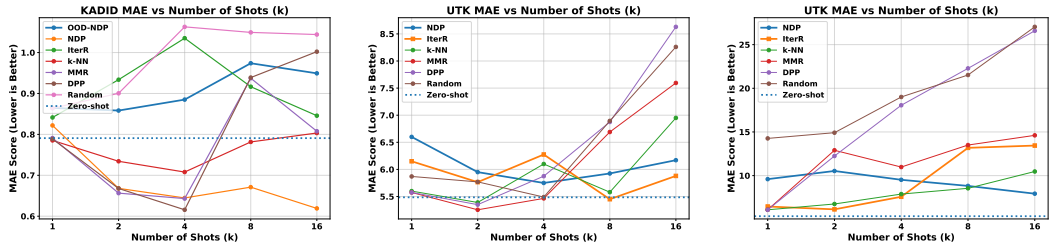
D.1 Out-of-Distribution and Cross-Model Transferability (K-Shot Dynamics)

Figure 5 reports K-shot curves for a cross-dataset shift (KonIQ-10k $\rightarrow$ KADID-10k) and two cross-model evaluations on UTKFace (a frozen Gemma-trained policy applied to Qwen 3 or Phi-3.5). The latter are distinct from the separately trained target-model policies in Table 2.

Changes with K do not remove the observed cross-dataset transfer failure: at $K = 16$, NDP obtains 0.94 MAE on KADID-10k, compared with 0.80 for k-NN and 0.81 for MMR under the transfer evaluation in Table 3. The UTKFace cross-model curves describe context-length sensitivity for the two evaluated target VLMs, not universal transfer across architectures or tasks. Table 2 likewise shows that a transferred policy is not uniformly preferable to in-domain training, zero-shot prompting, or heuristic selection. We make no claim of monotonic or exponential gains with K.

E Extended K-Shot Scaling Analysis

To provide a complete picture of our method's scalability and to supplement the analysis in the main text, we present the K-shot performance dynamics for the datasets evaluated in this study.



(a) Cross-Data (KonIQ-10k → KADID-10k)

(b) Cross-Model (Gemma 3 → Qwen 3)

(c) Cross-Model (Gemma 3 → Phi-3.5)

Figure 5: **Transfer performance by context length.** (a) A policy trained on KonIQ-10k and evaluated on KADID-10k. (b,c) A policy trained using Gemma 3 4B-it and evaluated without policy updates on Qwen 3 7B and Phi-3.5-vision, respectively, for UTKFace. These curves describe the plotted transfer settings; the $K = 16$ cross-dataset result remains below the reported heuristics.

Fine-Grained Regression. As illustrated in Figure 6, scaling the number of demonstrations does not yield a universally monotonic improvement across all methods or datasets. However, NDP demonstrates notable stability and resilience against the severe degradation at high k that plagues heuristics like DPP and Random selection (e.g., on KonIQ-10k and UTKFace). On datasets with complex continuous manifolds like AVA and KADID-10k, NDP effectively leverages the expanding context window to drive the Mean Absolute Error (MAE) progressively downward, avoiding the late-stage performance saturation or degradation seen in k-NN. Conversely, on UTKFace, increasing k beyond 4 slightly degrades NDP's performance, though it remains significantly more stable than the diversity-based baselines.

Few-Shot Classification. Figure 7 details performance on discrete categorization tasks, revealing highly dataset-dependent scaling dynamics. On subjective, boundary-sensitive datasets like EmoSet, NDP and DPP exhibit aggressive, near-monotonic climbs in accuracy up to $K = 16$, whereas k-NN struggles to provide the necessary contrastive hard negatives, resulting in lower overall peaks. Conversely, on rigid object categorization tasks like Food-101 and Stanford Cars, scaling the context window introduces noise; both NDP and k-NN experience noticeable accuracy drops as k increases, while MMR maintains better stability. On standard benchmarks like MNIST and CIFAR-10, NDP experiences intermediate performance dips (often around $k = 4$) before recovering at $K = 16$, whereas explicitly diverse heuristics like DPP maintain a more consistent upward trajectory.

Visual Question Answering (VQA). As shown in Figure 8, VQA tasks exhibit the highest volatility across all context window sizes. NDP's scaling behavior is highly non-monotonic in this domain; for instance, on GQA, it peaks sharply at $k = 4$, drops at $k = 8$, and rises again at $k = 16$. Similarly, it experiences an intermediate dip on OK-VQA before spiking. Despite this volatility, given the maximum token budget ($K = 16$), NDP achieves the highest overall accuracy on reasoning-heavy datasets like TextVQA and OK-VQA. However, standard k-NN proves exceptionally strong and competitive in the VQA domain, exhibiting a more consistent upward trend and ultimately outperforming NDP on VQAv2 and marginally on GQA.

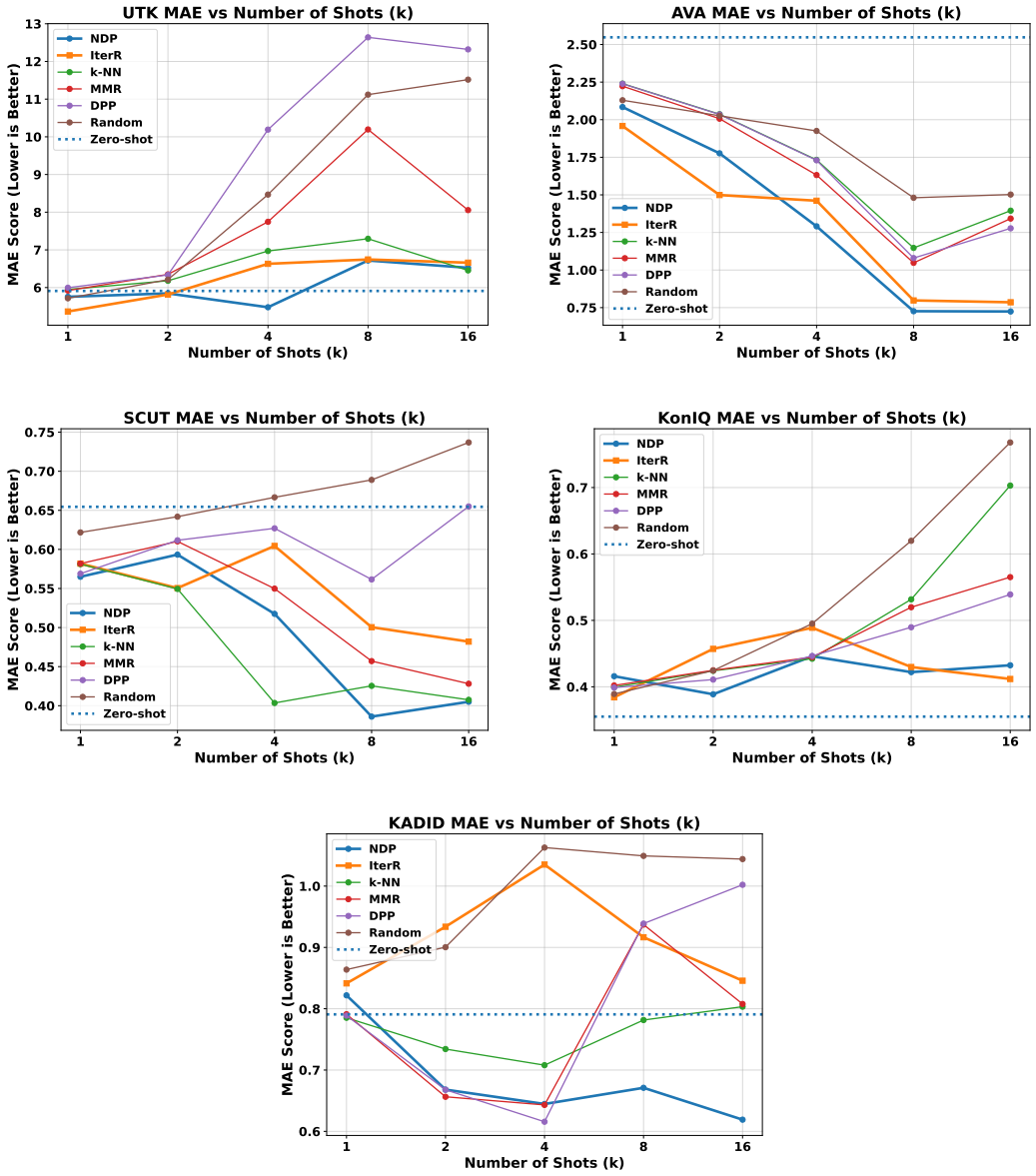


Figure 6: **Fine-Grained Regression K-Shot Scaling.** Comparison of MAE (↓) across context window sizes. While performance scaling is dataset-dependent, NDP generally avoids the severe late-stage degradation seen in random or purely diversity-driven (DPP) retrieval, effectively utilizing larger contexts on datasets like AVA and KADID-10k.

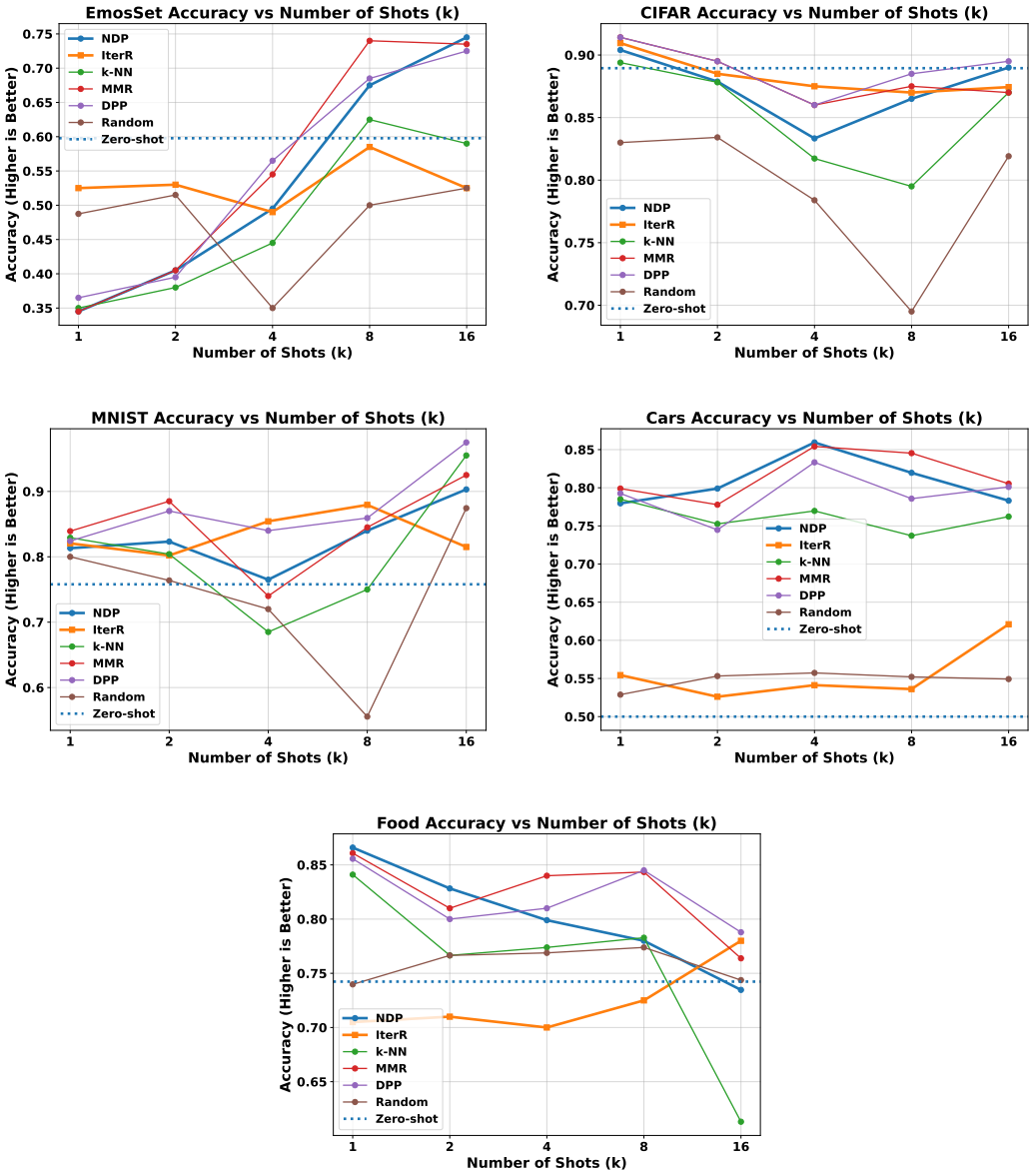


Figure 7: **Few-Shot Classification K-Shot Scaling.** Comparison of Accuracy (↑) across context window sizes. Scaling dynamics vary significantly by domain; NDP aggressively scales on subjective tasks (EmoSet) but, alongside k-NN, faces degradation on rigid categorization tasks (Food-101) as the context window expands.

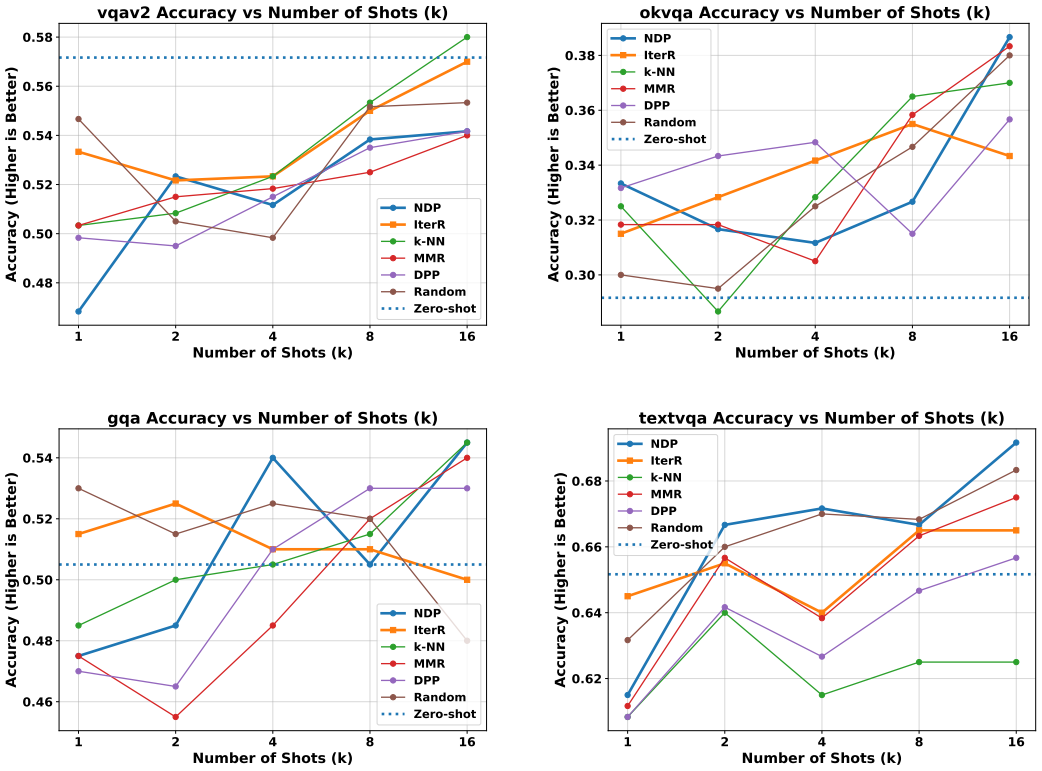


Figure 8: **Visual Question Answering K-Shot Scaling.** Comparison of Accuracy (↑) across context window sizes. The domain is characterized by high volatility. While k-NN scales smoothly and dominates on VQAv2, NDP successfully utilizes the maximum context size ($K = 16$) to peak on TextVQA and OK-VQA, despite intermediate instability.

F Extended Qualitative Study: Selected Demonstrations

Building upon the qualitative analysis in the main text, this section provides visual comparisons of the demonstration sequences constructed by standard k-NN versus our NDP policy.

The qualitative examples perfectly map to the quantitative gains observed in the previous sections. In regression (e.g., AVA, SCUT-FBP5500), k-NN fetches visually identical textures or the exact same individual, providing zero scale context. NDP (PPO), however, explicitly pulls a diverse spectrum of subjects and visual conditions. In classification (e.g., Food-101, CIFAR-10), k-NN isolates near-duplicate frames of the exact same object, whereas NDP actively curates a sequence of distinct boundary objects, establishing a clear contrastive space for the VLM to navigate.

F.1 Regression Demonstrations

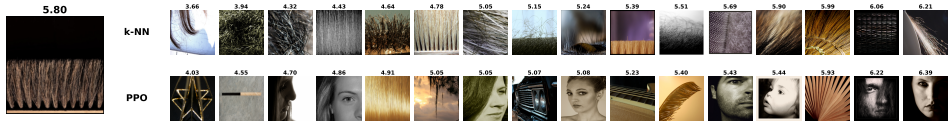


Figure 9: **AVA:** Qualitative comparison showing visual scales constructed by NDP vs. k-NN.

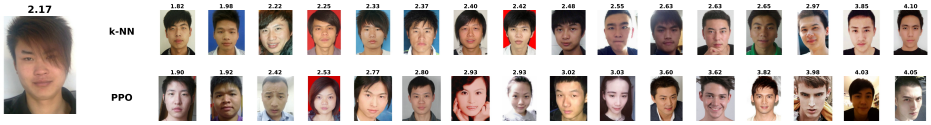


Figure 10: **SCUT-FBP5500:** Qualitative comparison showing visual scales constructed by NDP vs. k-NN.



Figure 11: **KonIQ-10k:** Qualitative comparison showing visual scales constructed by NDP vs. k-NN.

F.2 Classification Demonstrations

F.3 Visual Question Answering (VQA) Demonstrations



Figure 12: **KADID-10k:** Qualitative comparison showing visual scales constructed by NDP vs. k-NN.



Figure 13: **Food-101:** Qualitative comparison showing contrastive boundaries by NDP vs. k-NN.



Figure 14: **CIFAR-10:** Qualitative comparison showing contrastive boundaries by NDP vs. k-NN.



Figure 15: **MNIST:** Qualitative comparison showing contrastive boundaries by NDP vs. k-NN.



Figure 16: **Stanford Cars:** Qualitative comparison showing contrastive boundaries by NDP vs. k-NN.



Figure 17: **VQAv2:** Qualitative comparison showing structural diversity by NDP vs. k-NN.



Figure 18: **OK-VQA:** Qualitative comparison showing structural diversity by NDP vs. k-NN.



Figure 19: **GQA:** Qualitative comparison showing structural diversity by NDP vs. k-NN.



Figure 20: **TextVQA:** Qualitative comparison showing structural diversity by NDP vs. k-NN.