Reinforcement Learning for Retrieval Path Optimization in RAG: Tackling Sparse Feedback and Distributional Shift via Dynamic State Encoding and Meta-Policy Adaptation

1st Zhengyu Ma
Artificial Intelligence Laboratory
China Mobile Communication Group Tianjin Co., Ltd.
Tianjin, China
mazhengyu@hotmail.com

2nd Xiaojia Jin
China Mobile Communication Group Tianjin Co., Ltd.
Tianjin, China
jinxiaojia@tj.chinamobile.com

3rd Dongming Zhao
Artificial Intelligence Laboratory
China Mobile Communication Group
Tianjin Co., Ltd.
Tianjin, China
waitman_840602@163.com

4th Bo Wang
College of Intelligence and Computing
Tianjin University
Tianjin, China
bo_wang@tju.edu.cn

5th Yongzong Wang
College of Intelligence and Computing
Tianjin University
Tianjin, China
wangyongzong@inspur.com

***Abstract*—Retrieval-Augmented Generation (RAG) combines document retrieval with generation models, and its performance depends on *how* evidence is aggregated across multi-step retrieval rather than a single top-k list. We focus on reinforcement learning (RL)-based retrieval *path* optimization, addressing two key challenges: sparse, delayed long-horizon rewards from downstream QA tasks, and generalization degradation under corpus distribution shift. We propose a framework with two modules: Dynamic Multi-Scale State Encoding (DMSE) learns temporal state representations by hierarchically encoding retrieval trajectories (via BiGRU, self-attention, and descriptive statistics, fused via cross-scale attention for phase-aware decisions); Meta-Policy Guided Path Adaptation (MPGPA) uses a lightweight meta-controller to track corpus signals, adjusting exploration and ranking parameters to stabilize model behavior. Experiments on seven open-domain QA benchmarks show significant improvements: our approach outperforms baselines including RAFT [1] by 2.9–19.6 EM points (7.9 on average), reduces performance variance by 51%, and enhances exploration diversity. Ablation studies validate DMSE and MPGPA's efficacy.**

***Index Terms*—RAG, Reinforcement-Learning-based Retrieval Optimization, Dynamic Multi-Scale State Encoding, Meta-Policy-Guided Path Adaptation, Corpus Distributional Shift, Multi-Hop Reasoning, Open-Domain Question Answering**

I. Introduction

Retrieval-Augmented Generation (RAG) improves language model outputs by retrieving external documents and conditioning generation on the retrieved evidence [2]. In many realistic settings (e.g., multi-hop question answering (QA)), retrieval is not a single-shot top-k operation. Instead, the system retrieves iteratively, where earlier choices shape later candidates and ultimately affect the generator's final answer. We refer to this sequence as a *retrieval path*. Optimizing a retrieval path is a sequential decision-making problem that naturally lends itself to reinforcement learning (RL). However, applying RL to RAG retrieval remains challenging in practice: meaningful supervision often arrives only at the end of the path (e.g., Exact Match (EM) / F1 of the generated answer), and retrieval corpora evolve over time, causing *distribution shift* that can invalidate a policy trained on a static snapshot [3].

A major obstacle is the *sparse and delayed* reward signal. The policy receives informative feedback only after the final generation step, yet it must perform long-horizon credit assignment to early retrieval decisions that shaped the evidence set. A second obstacle is that the *information needed to act* changes over the trajectory. Early steps benefit from broad semantic exploration, whereas later steps require fine-grained matching to the evolving context. If the RL state representation fails to encode these phase-dependent signals explicitly, the policy tends to become either myopic (overemphasizing immediate relevance) or unstable (over-exploring without accumulating coherent evidence).

To support phase-dependent decision-making under sparse feedback, we propose *Dynamic Multi-Scale State Encoding* (DMSE). DMSE constructs the RL state from three complementary views of the retrieval process: (i) a short-term encoder capturing the most recent retrieved documents, (ii) a long-term encoder modeling full-trajectory semantics, and (iii) a local encoder summarizing statistics of the current top-k candidate set. These representations are fused via cross-scale attention to produce a single state vector for the policy. This design makes the policy explicitly aware of *where it is* in the trajectory

and which evidence signals are currently most informative, improving long-horizon credit assignment without changing the underlying RL objective.

However, even with a strong state representation, a policy trained on one corpus snapshot can generalize poorly when the corpus distribution shifts (e.g., document updates, domain drift, or injected distractors). Heavyweight remedies such as online fine-tuning are often impractical in large-scale retrieval pipelines. This motivates a lightweight inference-time mechanism that can adjust retrieval behavior based on observable signals of distributional change.

To handle corpus shift without retraining, we introduce *Meta-Policy Guided Path Adaptation* (MPGPA). MPGPA does not select documents directly; instead, it monitors retrieval-space statistics (e.g., candidate entropy or similarity histogram shifts) and modulates two control knobs of the main policy: an exploration coefficient (α) and a ranking temperature (τ). The meta-controller is trained with a variance-minimization objective across multiple corpus conditions, encouraging stable performance under drift. Together, DMSE and MPGPA form a retrieval-path optimization framework that is both trajectory-aware and distribution-aware.

In this paper, we instantiate these two innovations in the context of multi-step retrieval for open-domain QA and validate our framework across seven benchmarks: NQ, TriviaQA, PopQA, HotpotQA, 2Wiki, Musique, and Bamboogle. These datasets span diverse reasoning types—from factual recall to multi-hop and commonsense inference—providing a rigorous testbed for evaluating retrieval path optimization. Our results show that the proposed model not only improves final QA performance over competitive RL-based and contrastive retrievers, but also exhibits greater stability under both synthetic and real-world distribution shift scenarios. Ablation studies further confirm the independent and joint contributions of DMSE and MPGPA, offering insights into the roles of dynamic state encoding and policy modulation in long-horizon retrieval.

Our contributions are threefold. (1) We formulate RL-based retrieval path optimization for RAG under sparse terminal rewards and corpus distribution shift, and identify trajectory-aware state modeling and lightweight deployment-time adaptation as two key requirements for practical RL-based retrieval. (2) We propose DMSE, a dynamic multi-scale state representation that jointly encodes recent context, long-range intent, and candidate-set statistics, and fuses them via cross-scale attention to enable phase-aware retrieval decisions. (3) We introduce MPGPA, a lightweight meta-policy that modulates exploration and ranking sharpness using retrieval-space statistics and is trained to minimize performance variance across shifted corpora; we validate the framework's effectiveness and robustness via extensive experiments on seven open-domain QA benchmarks.

II. Related Work

A. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has become a standard paradigm for knowledge-intensive NLP tasks, including open-domain QA, fact verification, and dialogue systems. A typical RAG pipeline retrieves a set of documents for a query and conditions a generator on the retrieved evidence. Early systems such as RAG [2] and FiD [4] mainly relied on static top-k retrieval using DPR [5] or BM25 [6]. While effective, such static retrieval does not capture dependencies across multiple retrieval steps, which are crucial for multi-hop and iterative settings.

Subsequent work explored iterative retrieval [7], [8] and retrieval-aware reranking to refine evidence selection. However, many of these methods still treat each retrieval step independently or rely on heuristic or supervised designs, limiting their ability to optimize the joint semantics of a multi-step retrieval trajectory. In contrast, our work explicitly models retrieval as a sequential decision process and optimizes retrieval *paths*.

B. Reinforcement Learning for Document Retrieval

Reinforcement learning (RL) provides a principled framework for sequential retrieval, where an agent selects documents under delayed feedback derived from downstream metrics (e.g., EM/F1). Prior work in open-domain QA [9], [10] and subsequent RL-based retrieval methods formulate the process as an MDP, with states encoding the evolving context and actions selecting the next document. These approaches can improve over static baselines, especially for multi-hop retrieval.

Nevertheless, RL-based retrieval remains challenging due to sparse and delayed rewards and insufficient state modeling. Many methods receive supervision only after final generation, making long-horizon credit assignment difficult, and often rely on shallow or fixed state representations that fail to capture phase-dependent information needs along the trajectory. Our DMSE addresses this by constructing temporally structured multi-scale state representations tailored to retrieval trajectories.

C. Dynamic State Representations in RL

State representation is critical in RL under partial observability and long-term dependencies. Recurrent models (e.g., LSTMs/GRUs) summarize history but may overemphasize recent observations. Recent advances propose structured history modeling, including temporal attention [33] and transformer-based RL agents [34], which improve performance by capturing both short- and long-range dependencies.

Such dynamic state modeling is particularly important for retrieval, where decisions depend on evolving semantic intent and interactions across previously retrieved evidence. However, these ideas have not been systematically integrated into RL-based retrieval for RAG. Our Dynamic Multi-Scale State Encoding (DMSE) bridges this gap by jointly encoding short-term context, long-term trajectory semantics, and local candidate-set statistics, and fusing them via cross-scale attention for phase-aware decision-making.

D. Domain Adaptation and Meta-RL

Distribution shift is a persistent challenge in retrieval-augmented systems, arising from corpus updates, domain drift,

and injected distractors. Conventional remedies such as fine-tuning, domain adaptation objectives, or data augmentation can be costly or require target-domain supervision, limiting practicality in dynamic deployments. Meta-learning aims to improve generalization across environments and enable rapid adaptation, but its integration into RL-based retrieval for RAG remains underexplored.

In RL, meta-RL methods such as MAML [12] and contextual modulation improve robustness by conditioning behavior on environment signals. Inspired by this line of work, we propose Meta-Policy Guided Path Adaptation (MPGPA), a lightweight meta-controller that monitors retrieval-space statistics and modulates exploration and ranking parameters at inference time. MPGPA is trained to minimize performance variance across corpus conditions, thereby improving robustness under distribution shift.

III. Problem Formulation

We formulate retrieval path optimization in RAG as a finite-horizon Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, T)$, where an agent sequentially selects documents from a corpus to construct a retrieval trajectory. At each step $t \in \{1, \ldots, T\}$, the agent observes a state $s_t \in \mathcal{S}$, takes an action $a_t \in \mathcal{A}_t$, receives a reward $r_t \in \mathbb{R}$, and transitions to $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$. The process terminates after a fixed horizon T or when the agent selects a special STOP action (treated as trajectory termination). Since retrieval can be partially observable, we assume approximate Markovity under a state design that summarizes retrieval history and candidate-set statistics.

Notation. We use T for the maximum number of retrieval steps, k for the candidate pool size, and $\gamma \in (0, 1]$ for the discount factor. The main retrieval policy is $\pi_\theta(a_t \mid s_t)$. When MPGPA is enabled, $\alpha_t \in [0, 1]$ is an exploration mixing weight and $\tau_t > 0$ is a softmax temperature controlling ranking sharpness.

The environment consists of a document corpus $\mathcal{D}$, a retrieval engine $\mathcal{M}$ (e.g., BM25 or a dense retriever), and a question q. The action space $\mathcal{A}_t$ is the current top-k candidate set returned by $\mathcal{M}$ at step t (excluding previously selected documents), allowing the candidate pool to evolve across steps. The state s_t summarizes the retrieval history $H_{1:t-1} = [d_1, d_2, \ldots, d_{t-1}]$, the current candidate pool $C_t \subset \mathcal{D}$, and auxiliary signals (e.g., query embeddings or uncertainty measures).

The reward is sparse and delayed, defined at the end of the trajectory based on the generator output $\hat{a}$:

$$r_t = \begin{cases} 0, & t < T \\ \text{Eval}(G(q, H_{1:T}), a^{\text{gold}}), & t = T \end{cases}$$

where $\text{Eval}(\cdot)$ may correspond to exact match (EM), F1 score, or log-likelihood-based utility. The objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t\right] = \mathbb{E}_{\pi_\theta}\left[r_T\right]$$

since $r_t = 0$ for $t < T$. This setting introduces long-range credit assignment and state ambiguity due to evolving document semantics and candidate distributions, motivating temporally structured state modeling and environment-conditioned exploration.

We model the retrieval state representation s_t as a composite of multi-resolution semantic features. Let $d_t \in \mathbb{R}^h$ denote the embedding of the document selected at time t, and let $\mathcal{C}_t = \{c_{t,1}, \ldots, c_{t,k}\}$ be the set of top-k candidate embeddings at step t. The multi-scale state embedding is defined as:

$$s_t = \text{CrossAttn}(f_{\text{short}}(H_{t-w:t-1}), f_{\text{long}}(H_{1:t-1}), f_{\text{local}}(\mathcal{C}_t))$$

where:

- $f_{\text{short}}(\cdot)$: encodes recent documents using a sliding-window BiGRU,
- $f_{\text{long}}(\cdot)$: aggregates the full trajectory with positional encoding and self-attention,
- $f_{\text{local}}(\cdot)$: computes statistical descriptors of the candidate pool (e.g., entropy, cluster variance).

Cross-scale attention fuses these representations into a unified state vector used by the policy.

In deployment, the candidate distribution $\mathcal{C}_t$ may drift due to corpus updates or distribution shift. We therefore introduce an adaptive policy modulation mechanism governed by a meta-policy μ_ϕ, which observes retrieval-level statistics z_t and outputs modulation parameters α_t and τ_t:

$$\begin{aligned} (\alpha_t, \tau_t) &= \mu_\phi(z_t), \\ \pi_\theta^{\text{mod}}(a_t \mid s_t) &= (1 - \alpha_t) \cdot \text{Softmax}\left(\frac{Q_\theta(s_t, \cdot)}{\tau_t}\right)(a_t) \\ &\quad + \alpha_t \cdot \text{Unif}(\mathcal{A}_t)(a_t) \end{aligned}$$

The meta-policy is optimized to minimize the variance of the downstream QA reward across corpus distributions $\mathcal{D}^{(i)} \sim \mathcal{P}$:

$$\min_{\phi} \ \text{Var}_{\mathcal{D}^{(i)}}\left[\mathbb{E}_{\pi_\theta^{\mu_\phi}}[r_T^{(i)}]\right]$$

This formulation keeps the main policy focused on reward maximization while enabling lightweight inference-time adaptation under distributional change.

IV. Dynamic Multi-Scale State Encoding (DMSE)

A. Motivation

A key limitation of existing RL-based document retrieval methods is their oversimplified state modeling. Many approaches represent the retrieval state using flat concatenations of retrieved document embeddings or static relevance scores from dense retrievers, which are insufficient to capture the temporal dynamics and hierarchical semantics of multi-step retrieval. In RAG, retrieval paths exhibit strong cross-phase dependencies: early retrieved documents steer subsequent retrieval by shaping the semantic search space, while later decisions require awareness of both accumulated intent and the current candidate distribution. Moreover, information needs vary across phases—early steps benefit from broad semantic exploration, whereas later steps demand precise context matching. These observations motivate a phase-aware policy operating on a multi-scale state representation that integrates

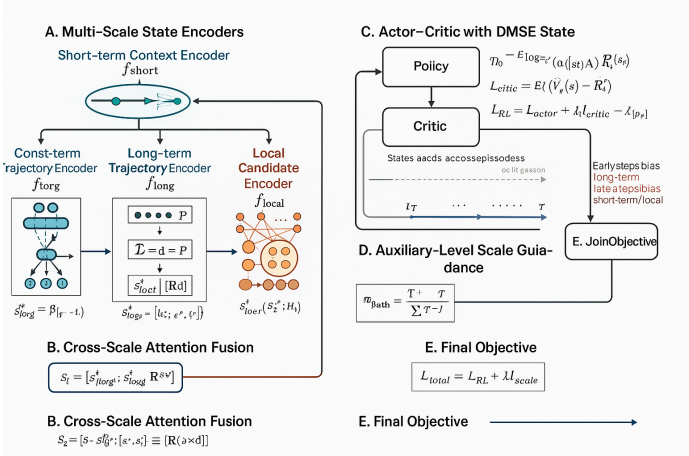

Fig. 1. Overview of Dynamic Multi-Scale State Encoding (DMSE). The retrieval horizon is T; the figure depicts a length-T retrieval trajectory that is encoded at short-term, long-term, and local (candidate-set) scales and fused via cross-scale attention.

short-term context, long-term intent, and local candidate structure (Fig. 1).

To this end, we propose Dynamic Multi-Scale State Encoding (DMSE), a structured state representation framework that provides temporally and semantically enriched observations for RL-based retrieval. At each step t, DMSE constructs a composite state embedding s_t by aggregating three sources of information: (1) recent retrieval history over a short window $H_{t-w:t-1}$, (2) the full retrieval trajectory $H_{1:t-1}$, and (3) statistical descriptors of the current candidate pool $\mathcal{C}_t$. These components are encoded by specialized modules and fused via cross-scale attention into a single state vector for action selection, enabling the policy to adaptively emphasize different information scales across retrieval phases.

B. Architecture

We define the DMSE state at step t, denoted s_t, as the fusion of three scale-specific encoders:

$$s_t = \text{CrossAttn}\left(f_{\text{short}}(H_{t-w:t-1}), f_{\text{long}}(H_{1:t-1}), f_{\text{local}}(\mathcal{C}_t)\right)$$

1) Short-term Context Encoder f_{short}: The short-term encoder captures recent retrieval patterns indicative of immediate semantic focus. We implement it with a sliding-window bidirectional GRU (BiGRU) [13] over the last w retrieved documents:

$$f_{\text{short}}(H_{t-w:t-1}) = \text{BiGRU}\left([d_{t-w}, \ldots, d_{t-1}]\right) \in \mathbb{R}^h$$

Each document $d_i \in \mathbb{R}^d$ is represented as a dense embedding (e.g., the CLS token from BERT [14], or sentence embeddings from Word2Vec [15] / GloVe [16]). The BiGRU models forward and backward semantic transitions within the window, and we use the final hidden state as the short-term representation.

2) Long-term Trajectory Encoder f_{long}: To capture long-range intent accumulation, we apply a position-aware self-attention encoder [17], [18] over the full history $H_{1:t-1}$. Let $D = [d_1, \ldots, d_{t-1}] \in \mathbb{R}^{(t-1)\times d}$ be the matrix of historical document embeddings. We add sinusoidal positional encoding $P \in \mathbb{R}^{(t-1)\times d}$ and apply multi-head self-attention:

$$D' = \text{MultiHeadAttn}(D + P) \in \mathbb{R}^{(t-1)\times d}$$

$$f_{\text{long}}(H_{1:t-1}) = \text{Aggregate}(D') = \frac{1}{t-1}\sum_{i=1}^{t-1} D'_i$$

This encoder summarizes global semantic structure along the retrieval path, including redundancy, coverage, and topical progression.

3) Local Candidate Encoder f_{local}: The local encoder captures distributional characteristics of the current candidate set $\mathcal{C}_t = \{c_{t,1}, \ldots, c_{t,k}\}$, where each $c_{t,i} \in \mathbb{R}^d$. Instead of encoding each candidate individually, we summarize candidate-set geometry with compact statistics. We compute the mean embedding

$$\mu_t = \frac{1}{k}\sum_{i=1}^{k} c_{t,i},$$

the element-wise variance embedding

$$\sigma_t^2 = \frac{1}{k}\sum_{i=1}^{k}(c_{t,i} - \mu_t)^2,$$

and a coarse cluster entropy H_t obtained by assigning candidates to K clusters and computing empirical cluster proportions p_j:

$$H_t = -\sum_{j=1}^{K} p_j \log p_j, \quad p_j = \frac{1}{k}\sum_{i=1}^{k}\mathbb{I}(c_{t,i} \in \text{cluster}_j).$$

We then concatenate these statistics into a compact vector:

$$f_{\text{local}}(\mathcal{C}_t) = [\mu_t; \sigma_t^2; H_t] \in \mathbb{R}^{2d+1}.$$

This representation informs the policy about semantic variability and uncertainty in the current retrieval space.

4) Cross-Scale Attention Fusion: To integrate the three streams, we denote $s_t^{\text{short}} = f_{\text{short}}(H_{t-w:t-1})$, $s_t^{\text{long}} = f_{\text{long}}(H_{1:t-1})$, and $s_t^{\text{local}} = f_{\text{local}}(\mathcal{C}_t)$. In practice, we project all scale-specific representations to a shared d-dimensional space before fusion; in our implementation, we set $h = d$ for the BiGRU encoder (or apply a linear projection when $h \neq d$). We then apply a learned attention fusion layer:

$$S_t = [s_t^{\text{short}}, s_t^{\text{long}}, s_t^{\text{local}}] \in \mathbb{R}^{3\times d}$$

$$\alpha_t = \text{Softmax}(W_a S_t^\top) \in \mathbb{R}^3, \quad W_a \in \mathbb{R}^{1\times d}$$

$$s_t = \sum_{i=1}^{3} \alpha_{t,i} S_{t,i}$$

The weights α_t are computed dynamically, allowing the policy to shift attention between recent decisions, long-term intent, and local candidate structure across retrieval phases.

C. Training

DMSE is integrated into an actor-critic RL framework, where the policy network π_θ takes the DMSE state s_t as input to select documents, and a critic network V_ψ estimates the expected return. We adopt the Advantage Actor-Critic (A2C) objective [19]. Given the terminal reward r_T, the return R_t and advantage A_t are:

$$R_t = \sum_{l=t}^{T} \gamma^{l-t} r_l = \gamma^{T-t} r_T, \quad A_t = R_t - V_\psi(s_t)$$

The actor loss is:

$$\mathcal{L}_{\text{actor}} = -\mathbb{E}_{\pi_\theta} \left[\log \pi_\theta(a_t|s_t) \cdot A_t \right]$$

and the critic loss is:

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{\pi_\theta} \left[\left(V_\psi(s_t) - R_t \right)^2 \right]$$

The overall RL objective includes entropy regularization $\mathcal{H}[\pi_\theta]$:

$$\mathcal{L}_{\text{RL}} = \mathcal{L}_{\text{actor}} + \lambda_c \cdot \mathcal{L}_{\text{critic}} - \lambda_e \cdot \mathcal{H}[\pi_\theta]$$

where λ_c and λ_e are weighting coefficients.

To stabilize learning under sparse terminal rewards, we further introduce an auxiliary trajectory-level supervision signal for DMSE's cross-scale attention. We compute a decayed reward signal:

$$\nabla_t^{\text{path}} = \frac{\gamma^{T-t} \cdot r_T}{\sum_{j=1}^{T} \gamma^{T-j}}, \quad t = 1, \ldots, T$$

and use it to guide the attention weights $\alpha_t \in \mathbb{R}^3$ via:

$$\mathcal{L}_{\text{scale}} = \sum_{t=1}^{T} \text{KL} \left(\alpha_t \parallel \alpha_t^{\text{target}} \right)$$

where α_t^{target} is a weak phase-aware prior (e.g., emphasizing long-term encoding in early steps and short-term encoding in late steps). The final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RL}} + \lambda_s \cdot \mathcal{L}_{\text{scale}}$$

All DMSE modules (BiGRU, attention encoders, and fusion layers) are trained end-to-end with the RL objective, supporting efficient backpropagation through time over each retrieval trajectory.

V. Meta-Policy Guided Path Adaptation (MPGPA)

A. Motivation

While reinforcement learning (RL) enables long-horizon optimization of retrieval policies, it typically assumes a stationary environment, i.e., that the data distribution remains stable between training and inference. In practical RAG deployments, this assumption is often violated: retrieval corpora evolve over time due to document updates, distribution shift, or user-specific context injection. As a result, retrieval paths learned under one distribution $\mathcal{D}_{\text{train}}$ may become suboptimal or even harmful when deployed under a different distribution $\mathcal{D}_{\text{test}}$. Such retrieval-space distribution shift can lead to substantial generalization degradation, especially in low-resource or dynamic knowledge-base settings.

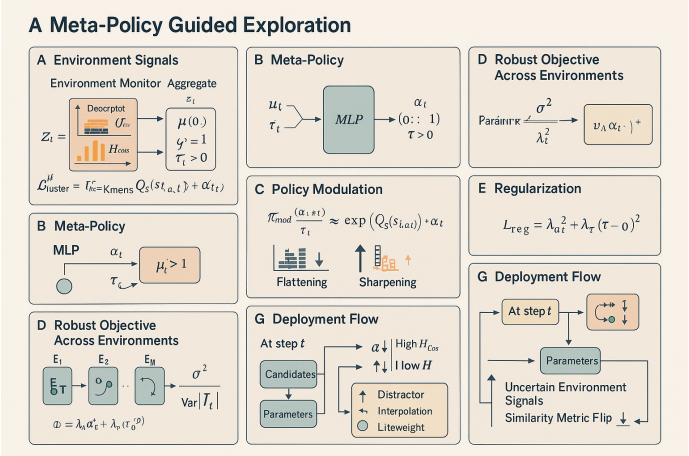

Fig. 2. Meta-Policy Guided Path Adaptation (MPGPA). The meta-controller μ_ϕ is shown from two perspectives: (i) training, where μ_ϕ is optimized with a variance-based objective across shifted corpora, and (ii) deployment, where μ_ϕ observes retrieval-space statistics and outputs α_t (exploration mixing) and τ_t (temperature) to modulate the main policy.

Conventional remedies such as online fine-tuning or replay-based updates are either computationally expensive (requiring gradient updates at inference time) or insufficiently robust under large-scale corpus drift. Moreover, these approaches typically do not explicitly leverage observable retrieval-environment signals (e.g., candidate entropy, semantic drift, or cluster dynamics), which can result in unstable behavior and brittle exploration under shift (Fig. 2).

To address this, we propose Meta-Policy Guided Path Adaptation (MPGPA), a lightweight auxiliary control mechanism that operates alongside the main retrieval policy. MPGPA decouples exploration modulation from the primary retrieval decision process and delegates it to a meta-controller that continuously monitors distributional indicators. Rather than selecting documents directly, MPGPA dynamically adjusts two inference-time control parameters of the main policy: an exploration coefficient α_t and a softmax temperature τ_t. This enables distribution-aware exploration and ranking calibration without retraining the main policy.

B. Meta-Policy Architecture

At each retrieval step t, the meta-policy μ_ϕ receives a contextual descriptor $z_t \in \mathbb{R}^m$ summarizing retrieval-environment statistics. This descriptor aggregates corpus-level signals such as candidate entropy, cluster compactness, and similarity histogram shifts. Let $\mathcal{C}_t = \{c_{t,1}, \ldots, c_{t,k}\}$ be the candidate embeddings at step t, and let μ_t denote their centroid:

$$\mu_t = \frac{1}{k} \sum_{i=1}^{k} c_{t,i}$$

We define the embedding variance σ_t^2, the K-means clustering loss $\mathcal{L}_{\text{cluster}}$ (with clusters $\mathcal{C}_t^{(j)}$ and centroids μ_j), and

cosine histogram entropy $H_{\cos}$ as:

$$\sigma_t^2 = \frac{1}{k}\sum_{i=1}^{k} \|c_{t,i} - \mu_t\|_2^2$$

$$\mathcal{L}_{\text{cluster}} = \sum_{j=1}^{K} \sum_{c \in \mathcal{C}_t^{(j)}} \|c - \mu_j\|_2^2$$

$$H_{\cos} = -\sum_{b=1}^{B} p_b \log p_b, \quad p_b = \frac{1}{k}\sum_{i=1}^{k} \mathbb{I}(\cos(c_{t,i}, \mu_t) \in \text{bin}_b)$$

These metrics are concatenated into a fixed-dimensional descriptor vector:

$$z_t = \left[\sigma_t^2; \mathcal{L}_{\text{cluster}}; H_{\cos}; \cdots\right] \in \mathbb{R}^m$$

The meta-policy μ_ϕ is implemented as an MLP that maps z_t to two continuous outputs:

$$(\alpha_t, \tau_t) = \mu_\phi(z_t), \quad \text{where } \alpha_t \in [0,1],\ \tau_t > 0$$

Here, α_t mixes uniform exploration over the current candidate set with score-based ranking induced by Q_θ, while τ_t controls the sharpness of the softmax ranking. We implement policy modulation as:

$$\pi_\theta^{\text{mod}}(a_t \mid s_t) = (1-\alpha_t) \cdot \text{Softmax}\left(\frac{Q_\theta(s_t, \cdot)}{\tau_t}\right)(a_t) + \alpha_t \cdot \text{Unif}(\mathcal{A}_t)(a_t),$$

where $Q_\theta(s_t, a_t)$ is a predicted score (e.g., an action-value proxy or policy logit). Increasing α_t injects broader exploration, while τ_t calibrates the sharpness of the softmax ranking.

This design enables dynamic behavioral control: under uncertain or rapidly shifting retrieval distributions (e.g., high $H_{\cos}$), the meta-policy can increase α_t to encourage exploration and adjust τ_t to flatten the ranking distribution. Conversely, under stable distributions, it reduces α_t and sharpens τ_t to favor exploitation.

Unlike the main retrieval policy trained via sparse downstream rewards, the meta-policy is optimized for robustness across corpus conditions. Let $\mathcal{D}^{(i)} \sim \mathcal{P}$ denote multiple simulated environments (e.g., varying document shifts or synthetic noise injections), and let $r_T^{(i)}$ be the final reward obtained from each. We define the meta-policy loss as:

$$\mathcal{L}_{\text{meta}} = \text{Var}_i\left[\mathbb{E}_{\pi_\theta^{\mu_\phi}}[r_T^{(i)}]\right]$$

This objective encourages stable performance across environments. In practice, we estimate $\mathcal{L}_{\text{meta}}$ using a batch of environment rollouts and update ϕ at a lower frequency than θ to keep the meta-controller lightweight.

To prevent over-reacting to noisy statistical signals, we regularize the outputs of the meta-policy:

$$\mathcal{L}_{\text{reg}} = \lambda_\alpha \cdot \alpha_t^2 + \lambda_\tau \cdot (\tau_t - \tau_0)^2$$

where τ_0 is a fixed temperature baseline (e.g., 1.0), and $\lambda_\alpha, \lambda_\tau$ are hyperparameters. The total meta-policy loss becomes:

$$\mathcal{L}_{\text{total}}^{\mu} = \mathcal{L}_{\text{meta}} + \mathcal{L}_{\text{reg}}$$

This formulation enables conservative yet effective adaptation, improving robustness under distribution shift without destabilizing the main policy.

VI. Experiments

A. Datasets

We evaluate on seven open-domain QA benchmarks spanning factual recall, multi-hop reasoning, compositional queries, and robustness under noisy/adversarial settings: NQ [20], TriviaQA [21], PopQA [22], HotpotQA [23], 2WikiMultiHopQA [24], Musique [25], and Bamboogle [26]. This diverse suite ensures rigorous evaluation of retrieval effectiveness and distributional robustness. We follow standard train/dev/test splits and associated corpora (e.g., Wikipedia snapshots). Each dataset is adapted to a multi-step RAG setting by first retrieving an initial candidate pool with a dense retriever (e.g., Contriever/DPR) and then applying RL-based multi-step selection.

B. Baselines

We compare against four baselines: (1) **RAG-DPR** [2], [5], a single-shot top-k RAG pipeline; (2) **RL-Retriever**, an actor–critic sequential retriever with a fixed/shallow state representation; (3) **Contrastive Rerank** [27], which re-ranks candidates via contrastive similarity without long-horizon policy optimization; and (4) **RAFT** [1], a strong recent RAG adaptation baseline. All methods use the same retriever–generator backbone when applicable and are evaluated on identical splits and metrics. For behavioral metrics (Table 2), we additionally report **RL-Dense** and **RL-Sparse**, which differ only in the candidate generator.

C. Implementation

We implement our framework in PyTorch with Hugging Face Transformers, using Contriever [27] (for strong zero-shot retrieval) as the retriever and a T5-based reader for generation (e.g., FiD-T5-base [4], [28]). At each step, the policy selects one document from a dynamic pool of size $k = 20$; after $T = 4$ steps, selected documents $H_{1:T}$ are passed to the generator $G(q, H_{1:T})$. Candidate retrieval uses MIPS with a FAISS backend [29]. We train an actor–critic policy with GAE [30] to reduce update variance; the actor consumes DMSE state representations and the critic uses a separate value head. Trajectories terminate via STOP or an entropy threshold; MPGPA modulates action probabilities with α_t and τ_t, and is updated once every 100 policy updates to minimize QA performance variance across distribution shifts, using retrieval-space statistics. We optimize with AdamW [31] and gradient clipping (max norm 1.0). Metrics include Exact Match (EM), F1 score, ROUGE [32], retrieval-path diversity, and document-selection entropy. All experiments run on 4 NVIDIA A100 GPUs with mixed precision.

VII. Results and Analysis

A. Overall Performance

We evaluate our framework on seven open-domain QA benchmarks—Natural Questions (NQ), TriviaQA, PopQA, HotpotQA, 2Wiki, Musique, and Bamboogle—covering factual retrieval, multi-hop reasoning, and robustness under noisy/adversarial settings. We evaluate models using four core metrics: Exact Match (EM), F1 score, path diversity, and document-selection entropy. EM and F1 reflect downstream answer accuracy, while path diversity and entropy characterize retrieval behavior, especially the ability to avoid myopic or repetitive retrieval patterns.

We first report the main QA results in terms of EM in Table I; F1 follows consistent trends and is omitted for brevity. As shown, **Ours (DMSE+MPGPA)** outperforms all strong baselines across datasets, demonstrating the effectiveness of our trajectory-aware and distribution-aware retrieval-path optimization. The gains are particularly pronounced on multi-hop benchmarks (HotpotQA, 2Wiki, Musique) and robustness-oriented datasets (PopQA, Bamboogle), aligning with DMSE's ability to capture semantic dependencies across multi-step trajectories and MPGPA's robustness under shift, and indicating improved evidence aggregation over multi-step retrieval paths.

TABLE I
Main experimental results on seven QA benchmarks. Best performance is highlighted in **bold**.

Dataset	RAG-DPR [2] [5]	RL-Retriever	Contrastive Rerank [27]	RAFT [1]	Ours MPGPA	**Ours** DMSE+ MPGPA
	EM (%)					
NQ	34.8	42.1	43.2	44.9	46.2	**47.8**
TriviaQA	54.4	58.3	60.1	61.8	64.1	**66.2**
PopQA	38.7	41.3	42.6	43.5	45.8	**47.3**
HotpotQA	25.5	29.7	32.1	38.9	43.1	**45.2**
2Wiki	22.6	27.4	30.2	35.1	44.3	**46.1**
Musique	4.7	6.6	8.9	15.4	20.8	**22.4**
Bamboogle	0.8	12.8	15.3	28.7	46.2	**48.3**
Average	25.9	31.2	33.2	38.3	44.4	**46.2**

To further evaluate retrieval behavior, we analyze path diversity and document-selection entropy across retrieval steps. Higher entropy indicates less over-confident selection and better-calibrated exploration, while higher diversity suggests the policy does not collapse to repetitive retrieval routes. As shown in Table II, our full method (DMSE+MPGPA) achieves higher entropy and diversity than the static retriever and two RL retrieval variants, consistent with DMSE providing richer trajectory state representations and MPGPA preserving exploration under distribution shifts.

In addition to absolute performance gains, we evaluate robustness under simulated corpus distribution shift, a core motivation for MPGPA. We simulate shifts by changing the underlying corpus slice (e.g., using different time-sliced Wikipedia snapshots) or injecting retrieval noise through paraphrased queries. Under these conditions, we compute the standard deviation (SD) of EM across 10 randomized slices per dataset. As shown in Table III, MPGPA significantly reduces performance variance, indicating its ability to stabilize retrieval behavior when corpus statistics shift. In contrast, the RL-Retriever without MPGPA exhibits larger fluctuations under shift.

TABLE II
Path diversity and document selection entropy comparison. Higher values indicate better exploration behavior and less deterministic retrieval patterns.

Model	Path Diversity ↑	Entropy of Doc Selection ↑
RAG-DPR	0.41	1.73
RL-Dense	0.48	1.85
RL-Sparse	0.50	1.92
Ours	0.58	2.14

TABLE III
Performance variance (standard deviation) under simulated corpus distribution shift. Lower values indicate better robustness.

Model	NQ (SD)	PopQA (SD)	HotpotQA (SD)	Musique (SD)
RL-Retriever (no MPGPA)	3.7	4.4	5.1	5.5
Ours (w/ MPGPA)	1.9	2.1	2.4	2.7

B. Component Analysis

We analyze the contributions of DMSE and MPGPA using the models reported in Tables I–III. First, the improvement from **RL-Retriever** to **Ours (MPGPA)** in Table I highlights the benefit of distribution-aware exploration modulation, showing that MPGPA yields consistent EM gains, particularly on multi-hop and robustness-focused benchmarks. Second, the gap between **Ours (MPGPA)** and **Ours (DMSE+MPGPA)**—holding MPGPA constant—isolates the effect of DMSE's multi-scale trajectory encoding, indicating that richer phase-aware state representations further improve long-horizon retrieval planning. Third, Table II shows that the full model maintains the highest path diversity and selection entropy, suggesting that combining phase-aware state encoding with adaptive exploration prevents premature collapse to overly deterministic retrieval routes. Table III further supports MPGPA's robustness benefit by showing substantially reduced EM variance under simulated corpus distribution shift.

VIII. Conclusion

This paper proposes a reinforcement learning framework for retrieval path optimization in Retrieval-Augmented Generation (RAG), addressing two practical challenges: sparse and delayed long-horizon rewards, and performance degradation under corpus distribution shift in dynamic retrieval environments. To tackle sparse feedback and long-range dependencies, we introduce Dynamic Multi-Scale State Encoding (DMSE), which hierarchically encodes retrieval trajectories into temporally-aware state representations to support phase-aware decision-making. To improve robustness under distribution shift, we propose Meta-Policy Guided Path Adaptation (MPGPA), a lightweight meta-controller that modulates exploration and

ranking behavior based on corpus-level distribution signals. Extensive experiments on seven open-domain QA benchmarks demonstrate consistent and significant gains over strong baselines in answer accuracy, exploration behavior, and robustness, and ablation studies confirm the complementary contributions of DMSE and MPGPA. Overall, our framework offers a practical solution for long-horizon retrieval-path learning in RAG that is both trajectory-aware and robust to corpus distribution shift.

ACKNOWLEDGEMENTS

The authors thank the reviewers for their valuable comments. We acknowledge that some English expressions were calibrated with chatgpt; all core research content is independently completed by the authors. We also thank the Artificial Intelligence Laboratory of China Mobile Communication Group Tianjin Co., Ltd. and Tianjin University for technical support, and colleagues for helpful discussions.

REFERENCES

[1] Y. Wang, X. Ma, G. Chen, Y. Wu, C. Huang, F. Huang, *et al.*: RAFT: Adapting language models to domain-specific RAG. arXiv preprint arXiv:2403.10131, 2024.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, *et al.*: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020).

[3] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, *et al.*: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2024.

[4] G. Izacard, E. Grave: Leveraging passage retrieval with generative models for open-domain question answering. arXiv preprint arXiv:2007.01282, 2020.

[5] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, *et al.*: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906, 2020.

[6] S. Robertson, H. Zaragoza: The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* **3**(4), 333–389 (2009).

[7] A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, H. Hajishirzi: XOR QA: Cross-lingual open-retrieval question answering. arXiv preprint arXiv:2010.11856, 2021.

[8] P. Qi, J. Lee, O. van Beek, C.D. Manning: Answering open-domain questions of varying reasoning steps from text. arXiv preprint arXiv:2010.12527, 2021.

[9] D. Chen, A. Fisch, J. Weston, A. Bordes: Reading Wikipedia to answer open-domain questions. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879. ACL, Vancouver (2017).

[10] W.-T. Yih, M. Richardson, C. Meek, M.-W. Chang, J. Suh: The value of semantic parse labeling for knowledge base question answering. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–206. ACL, Berlin (2016).

[11] G. Salton, A. Wong, C.-S. Yang: A vector space model for automatic indexing. *Communications of the ACM* **18**(11), 613–620 (1975).

[12] C. Finn, P. Abbeel, S. Levine: Model-agnostic meta-learning for fast adaptation of deep networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1126–1135. PMLR, Sydney (2017).

[13] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio: Learning phrase representations using RNN encoder–decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[15] T. Mikolov, K. Chen, G. Corrado, J. Dean: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[16] J. Pennington, R. Socher, C.D. Manning: GloVe: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. ACL, Doha (2014).

[17] D. Bahdanau, K. Cho, Y. Bengio: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[18] M.-T. Luong, H. Pham, C.D. Manning: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025, 2015.

[19] R.S. Sutton, A.G. Barto: *Reinforcement Learning: An Introduction*. MIT Press, Cambridge (2018).

[20] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, *et al.*: Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7**, 453–466 (2019).

[21] M. Joshi, E. Choi, D.S. Weld, L. Zettlemoyer: TriviaQA: A large-scale distantly supervised challenge dataset for reading comprehension. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. ACL, Vancouver (2017).

[22] C. Sciavolino, Z. Zhong, J. Lee, D. Chen: Simple entity-centric questions challenge dense retrievers. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6138–6148. ACL, Punta Cana (2021).

[23] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W.-W. Cohen, R. Salakhutdinov, C.D. Manning: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2369–2380. ACL, Brussels (2018).

[24] X. Ho, A.K.D. Nguyen, S. Sugawara, A. Aizawa: Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 6609–6625. ACL, Barcelona (2020).

[25] H. Trivedi, N. Balasubramanian, T. Khot, A. Sabharwal: MusiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics* **10**, 539–554 (2022).

[26] O. Press, M. Zhang, S. Min, L. Schmidt, N.A. Smith, M. Lewis: Measuring and narrowing the compositionality gap in language models. arXiv preprint arXiv:2210.03350, 2022.

[27] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave: Unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118, 2021.

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, *et al.*: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(140), 1–67 (2020).

[29] J. Johnson, M. Douze, H. Jégou: Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* **7**(3), 535–547 (2019).

[30] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel: High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015.

[31] I. Loshchilov, F. Hutter: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[32] C.-Y. Lin: ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, Barcelona (2004).

[33] T. Zahavy, A. Barreto, D.J. Mankowitz, S. Hou, B. O'Donoghue, I. Kemaev, S. Singh: Discovering a set of policies for the worst-case reward. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[34] E. Parisotto, F. Song, J. Rae, R. Pascanu, C. Gulcehre, S. Jayakumar, M. Jaderberg, *et al.*: Stabilizing transformers for reinforcement learning. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 7487–7498. PMLR (2020).