EAVI: EVIDENCE-AWARE CROSS-MODAL ALIGNMENT WITH VALUE-GUIDED INTERACTION FOR CHAT-BASED PERSON RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Chat-based person retrieval (**ChatPR**) identifies a target pedestrian in a gallery by progressively acquiring appearance evidence through multi-round question answering. This setting creates two requirements: a retriever must make effective use of evidence accumulated across turns, and a controller must acquire additional evidence selectively under a limited budget. We propose **EAVI** to address both. Its Dialogue Evidence Set Retriever (DESR) complements global matching with anonymous evidence sets aligned by entropy-regularized optimal transport and reliability-weighted aggregation, while keeping gallery representations cacheable. Its Value-Guided Interaction Agent (VGIA) selects from the recorded question pool using only the observable retrieval state, predicting each candidate's one-step retrieval value and deciding when to stop with an independent head. Because ranking questions within a dialogue and comparing values across dialogues rely on different properties of the value, VGIA treats question ordering, stopping, and shared-budget allocation as separate decisions. On ChatPedes, DESR improves R@1 by 1.7–2.4 points over the released DiaNA checkpoint under identical dialogue prefixes, with 3.0× fewer loaded parameters and about 6× faster full-test evaluation. A fitted question prior remains competitive for within-dialogue ordering, whereas retrieval-state feedback improves shared-budget allocation by 5.01 R@1 points on average over a question-and-progress control across four budgets on a fixed acquisition path. A development-selected stopping head reduces additional QA acquisitions by 37.7% at a 0.67-point R@1 cost.

1 INTRODUCTION

Text-based person retrieval (TBPR) identifies a target person in an image gallery from a natural-language description Jiang & Ye (2023); Qin et al. (2024); Radford et al. (2021), with applications in intelligent surveillance and public security Ye et al. (2021). Conventional TBPR assumes that the user can give a complete description in a single query Jiang & Ye (2023); Park et al. (2024), which rarely holds when the initial description is incomplete, vague, or ambiguous. Chat-based person retrieval (ChatPR) Bai et al. (2025) instead acquires complementary appearance evidence through multi-turn question answering (QA) to disambiguate visually similar candidates. This progressive setting creates two requirements: the retriever must make effective use of accumulated evidence, and a controller must decide how to acquire additional evidence under a limited budget.

The first requirement concerns matching accumulated evidence. Identity-discriminative cues may be spread across dialogue turns and correspond to localized visual regions. As the dialogue grows, a single global representation may mix them with redundant context, making fine-grained correspondence between dialogue and visual evidence harder to establish Bai et al. (2025); Park et al. (2024); Li et al. (2026). A ChatPR retriever should therefore align local evidence across modalities while keeping gallery-level retrieval efficient.

The second requirement concerns acquiring additional evidence. Candidate questions differ in retrieval value, yet the standard ChatPedes protocol reveals them in their recorded order Bai et al. (2025), so informative evidence may arrive late while redundant questions consume interaction. Estimating the value of a question Levy et al. (2023); Lee et al. (2024) is only part of the problem.

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.

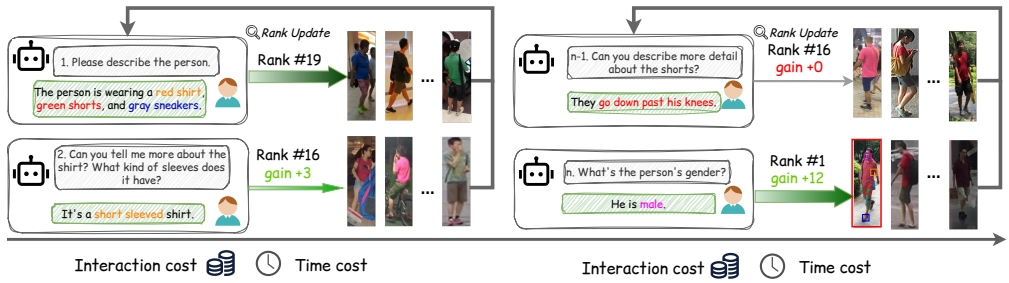

Figure 1: Motivation of evidence-aware and value-guided interaction in ChatPR. Multi-turn QA reveals sparse identity cues, while different questions provide unequal retrieval gains.

Ranking questions within one dialogue depends only on the relative order of their values, whereas deciding when to stop, or which of several dialogues sharing a budget should receive the next interaction, compares values across different retrieval states. A value that orders questions well therefore need not support the other two decisions, and interaction control should treat them separately.

We propose **EAVI**, a framework combining Evidence-Aware Cross-Modal Alignment with Value-Guided Interaction for ChatPR. Its retriever represents and matches the accumulated evidence, and its controller uses the resulting retrieval state to acquire more.

Dialogue Evidence Set Retriever (DESR). DESR complements a global representation of the dialogue Jiang & Ye (2023); Bai et al. (2025) with compact, anonymous evidence sets over dialogue tokens and visual patches. Content-Adaptive Evidence Transport (CAET) establishes soft correspondences between the two sets without predefined matching rules, and reliability-weighted aggregation lets weakly supported local matches contribute less. Gallery representations remain query-independent and cacheable, so local alignment does not make full-gallery retrieval query-dependent.

Value-Guided Interaction Agent (VGIA). VGIA selects questions from the recorded QA pool using only the observable retrieval state, never the target identity or an unrevealed answer. It predicts the one-step correction benefit minus correction harm of each candidate question from a question descriptor, a fitted question prior, and the retrieval state, and an independent head estimates whether the current top-1 prediction is correct in order to decide when to stop. The same values can rank questions within a dialogue or rank dialogues competing for a shared budget, and we evaluate question ordering, stopping, and allocation as separate operating modes rather than treating one score as a universal interaction rule.

Our contributions are as follows. **(1) Evidence-aware retrieval.** DESR adds anonymous evidence sets and CAET to global matching while keeping gallery representations cacheable. Under identical dialogue prefixes it improves R@1 over the released DiaNA checkpoint by 1.7–2.4 points, with 3.0× fewer loaded parameters and about 6× faster full-test evaluation. **(2) Decision-specific interaction control.** VGIA separates question ordering, stopping, and shared-budget allocation over one frozen retriever. A development-selected stopping head reduces additional QA acquisitions by 37.7% at a 0.67-point R@1 cost relative to acquiring all available QA. **(3) Decision-specific use of retrieval feedback.** Controlled ChatPedes experiments show that a fitted question prior remains competitive for within-dialogue ordering, whereas retrieval-state feedback substantially improves shared-budget allocation: on a fixed acquisition path, it raises R@1 by 5.01 points on average over a question-and-progress control across four budgets, and by 4.54 points over allocating to the least confident dialogue first.

2 Related Work

2.1 Text- and Chat-based Person Retrieval

Text-based person retrieval (TBPR) identifies a person in an image gallery from a natural-language description. Early methods learn global cross-modal representations Jiang & Ye (2023); Qin et al. (2024), while subsequent work models fine-grained correspondence through optimal transport Li

et al. (2026), slot-based part discovery Park et al. (2024), or language-model-assisted attribute mining Tan et al. (2024). These methods assume a static query available before ranking begins, where the retriever cannot acquire new evidence.

Chat-based person retrieval (ChatPR) replaces this one-shot setting with multi-turn interaction. DiaNA Bai et al. (2025) introduced ChatPedes, where appearance evidence is revealed through question–answer turns, followed by work on instruction tuning and question generation Lu et al. (2025); Niu et al. (2025). Existing ChatPR studies often collapse dialogue representation and question formulation into a single score. We separate them: DESR matches evidence from observed dialogue, whereas VGIA controls active evidence acquisition. Both are evaluated under a common recorded-QA protocol, keeping retrieval and interaction claims distinct.

2.2 Interactive Retrieval and Sequential Interaction Control

Interactive retrieval methods acquire feedback over multiple steps to update rankings. Many approaches focus on generating informative questions, including open-ended clarification, candidate-conditioned questioning, and foundation-model-driven interaction Levy et al. (2023); Lee et al. (2024); Lu et al. (2025); Niu et al. (2025). In these settings, question wording is part of the action, so retrieval performance conflates interaction value with question generation quality.

The recorded-QA protocol fixes the question pool and corresponding answers in advance, deciding which legal question to reveal rather than how to generate one. This isolates evidence acquisition from question generation and exposes both beneficial and harmful questions, as in active feature acquisition settings Shim et al. (2018); Li & Oliva (2021); Wang et al. (2015); Melville et al. (2005). It consequently motivates a signed value estimate: a question may improve, preserve, or degrade the current ranking.

A separate line of work studies when to stop an interaction using uncertainty, confidence, or trajectory-level criteria Zhang et al. (2025); Park et al. (2025); Qin et al. (2023). Question selection and stopping are distinct decisions: the optimal next question may yield small expected gain, while stopping rules cannot select which question to ask.

Budgeted interaction introduces control when several dialogues share a common budget. The system must compare the next-step value across active dialogues to allocate interactions, a cross-instance scheduling problem distinct from single-trajectory optimization and connected to offline policy evaluation from logged data Kallus & Zhou (2022); Dudík et al. (2014). Selection and stopping are studied primarily within individual trajectories, whereas cross-instance allocation remains largely absent from ChatPR evaluations.

EAVI positions itself at the intersection of these lines of work. It keeps the retriever fixed while separating three operating decisions: selecting the next recorded question within a dialogue, deciding whether to stop, and allocating a shared budget across dialogues. This separation allows these decisions to be evaluated independently under the same retriever and recorded QA pool.

3 Method

We present **EAVI**, a conversational person retrieval framework that separates evidence matching from evidence acquisition. The **Dialogue Evidence Set Retriever (DESR)** scores a gallery image using the dialogue observed so far, while the **Value-Guided Interaction Agent (VGIA)** chooses the next item from a recorded question–answer (QA) pool and decides when the current ranking is ready to submit. Thus, VGIA is a target-blind controller for selecting recorded questions; it does not generate open-ended questions or inspect an answer before that question is selected.

3.1 Problem Formulation

Let $\mathcal{G} = \{I_j\}_{j=1}^{|\mathcal{G}|}$ be the gallery and let $\mathcal{D} = \{(Q_i, A_i)\}_{i=1}^{N}$ be the recorded QA pool associated with a query. One QA pair is observed at initialization. Let σ_0 denote its index and set $\mathcal{X}_0 = [(Q_{\sigma_0}, A_{\sigma_0})]$. At step t, let $\boldsymbol{\sigma}_{1:t}$ index the additional questions asked after this initial pair. The observed evidence,

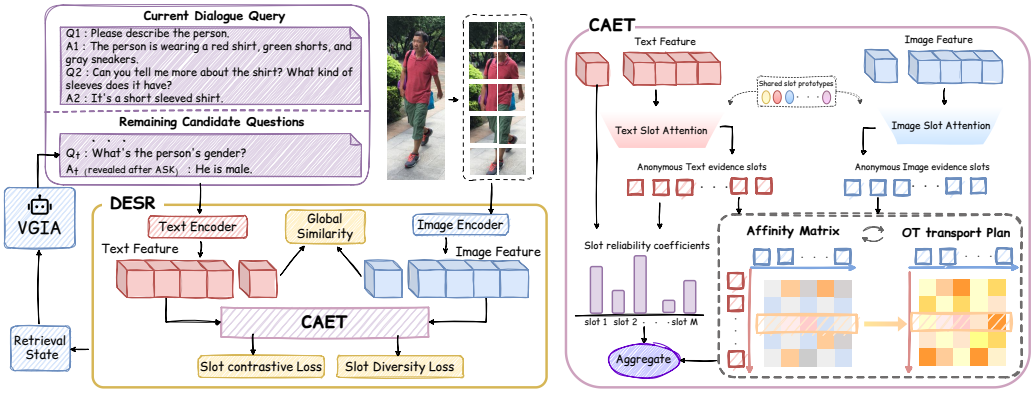

Figure 2: Overview of DESR. Dual encoders give global dialogue and image features. CAET extracts anonymous evidence slots in both modalities, aligns them by entropy-regularized optimal transport, and aggregates the transported similarities with slot reliability. The global and evidence-set scores are combined for retrieval, and gallery features are computed once and cached.

the indices of the remaining pairs, and the legal questions are

$$\begin{aligned} \mathcal{X}_t &= \mathcal{X}_0 \,\|\, [(Q_{\sigma_1}, A_{\sigma_1}); \ldots; (Q_{\sigma_t}, A_{\sigma_t})], \\ \mathcal{I}_t &= \{1, \ldots, N\} \setminus \{\sigma_0, \sigma_1, \ldots, \sigma_t\}, \qquad \mathcal{Q}_t = \{(i, Q_i) : i \in \mathcal{I}_t\}, \end{aligned} \tag{1}$$

where $\|$ denotes concatenation. Questions are indexed rather than identified by their text, because the same wording can occur more than once in a recorded pool. At each step, the controller executes $\texttt{ASK}(q)$ for $q \in \mathcal{Q}_t$, which reveals the recorded answer of that pair and appends the pair to $\mathcal{X}_t$, or executes $\texttt{SUBMIT}$, which returns the current gallery ranking. The interaction cap K counts additional $\texttt{ASK}$ actions and is unrelated to the number of DESR slots below; after K additional asks no further question is legal and the controller must submit.

3.2 Dialogue Evidence Set Retriever (DESR)

DESR combines global dual-encoder matching with fine-grained alignment of anonymous evidence sets. A text encoder produces contextualized dialogue tokens $\mathbf{W}_t = [\mathbf{w}_{t,0}, \ldots, \mathbf{w}_{t,L_t-1}]$, and an image encoder produces a class token $\mathbf{v}_{\text{cls}}$ and patch features $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_H]$. Their normalized global representations are

$$\mathbf{w}_t = \text{Norm}(\Phi_D(\mathbf{w}_{t,0})), \qquad \mathbf{v} = \text{Norm}(\Phi_I(\mathbf{v}_{\text{cls}})). \tag{2}$$

The image representation is encoded once and cached; the dialogue representation is recomputed when a new QA pair is appended.

To align local evidence without imposing a turn-to-region correspondence, we add role and turn embeddings to the dialogue tokens and extract content- dependent slots using a shared prototype initialization. We use M slots in each modality:

$$\begin{aligned} \widehat{\mathbf{W}}_t &= \Psi_{\text{txt}}(\mathbf{S}^{(0)}, \widetilde{\mathbf{W}}_t) = [\widehat{\mathbf{w}}_{t,1}, \ldots, \widehat{\mathbf{w}}_{t,M}], \\ \widehat{\mathbf{V}} &= \Psi_{\text{img}}(\mathbf{S}^{(0)}, \mathbf{V}) = [\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_M]. \end{aligned} \tag{3}$$

Here $\widetilde{\mathbf{W}}_t$ includes the role and turn embeddings. The slots are anonymous: their indices do not encode a fixed dialogue turn or image part. We call the resulting alignment procedure *Content-Adaptive Evidence Transport (CAET)*. Let $A_{mn} = \cos(\widehat{\mathbf{w}}_{t,m}, \widehat{\mathbf{v}}_n)$ be the text-slot/visual-slot affinity. CAET obtains a soft correspondence with entropy-regularized optimal transport:

$$\mathbf{P}^* = \underset{\mathbf{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})}{\arg\max} \langle \mathbf{P}, \mathbf{A} \rangle + \varepsilon \mathcal{H}(\mathbf{P}), \qquad \boldsymbol{\mu} = \boldsymbol{\nu} = M^{-1}\mathbf{1}_M, \tag{4}$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is the transport polytope and $\mathcal{H}(\mathbf{P}) = -\sum_{m,n} P_{mn} \log P_{mn}$. A reliability gate then estimates how much each textual slot should contribute. If $\eta_{t,m}$ and $p_{t,m}$ are respectively the

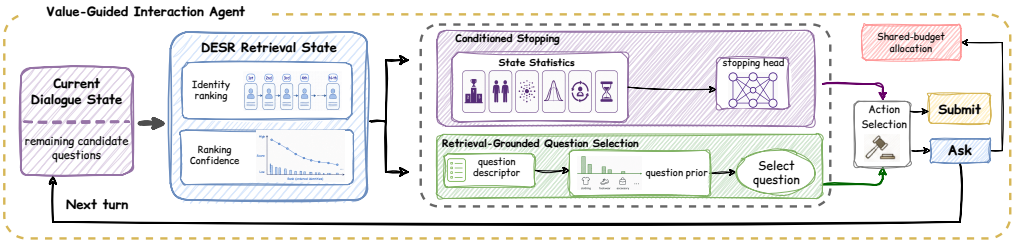

Figure 3: Overview of VGIA at inference. DESR identity scores form a target-blind state to rank candidate questions via a residual value head, trigger SUBMIT via a stopping head, and allocate shared budgets across dialogues.

entropy and peak mass of its attention map, then

$$\rho_{t,m} = \mathrm{sigmoid}(\mathrm{MLP}([\widehat{\mathbf{w}}_{t,m}; \mathbf{w}_t; \eta_{t,m}; p_{t,m}])) \,. \tag{5}$$

The gate is applied after transport. With row mass $u_{t,m} = \sum_n P^*_{mn}$ and normalized reliability $\bar{\rho}_{t,m} = \rho_{t,m} / \sum_{m'} \rho_{t,m'}$, the evidence-set score is

$$s_{\mathrm{set}}(\mathcal{X}_t, I) = \sum_{m=1}^{M} \bar{\rho}_{t,m} \sum_{n=1}^{M} \frac{P^*_{mn}}{u_{t,m}} A_{mn}. \tag{6}$$

This normalization makes the score a reliability-weighted average of the transported slot similarities; it also matches the uniform-marginal implementation, for which $u_{t,m} = 1/M$.

The global and combined retrieval scores are

$$\begin{aligned} s_g(\mathcal{X}_t, I) &= \mathbf{w}_t^\top \mathbf{v}, \\ s(\mathcal{X}_t, I) &= s_g(\mathcal{X}_t, I)/\tau_g + \alpha s_{\mathrm{set}}(\mathcal{X}_t, I)/\tau_{\mathrm{set}}, \end{aligned} \tag{7}$$

where τ_g and τ_{set} are temperatures and α balances global and evidence-set matching. The gallery-side image features, including visual slots, are cached once.

For training, let $\mathcal{L}(\cdot; \tau)$ denote a bidirectional multi-positive contrastive loss in which every dialogue–image pair sharing an identity label is a positive, in both the text-to-image and image-to-text directions. DESR uses the global score and the CAET score as separate supervision signals and regularizes slot attention:

$$\mathcal{L}_{\mathrm{global}} = \mathcal{L}(s_g; \tau_g), \qquad \mathcal{L}_{\mathrm{slot}} = \mathcal{L}(s_{\mathrm{set}}; \tau_{\mathrm{set}}), \tag{8}$$

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{M(M-1)} \sum_{m \neq m'} \left[\langle \mathbf{a}^{\mathrm{txt}}_m, \mathbf{a}^{\mathrm{txt}}_{m'} \rangle + \langle \mathbf{a}^{\mathrm{img}}_m, \mathbf{a}^{\mathrm{img}}_{m'} \rangle \right], \tag{9}$$

where the two $\mathbf{a}$ terms are normalized token- and patch-level attention maps. The complete retrieval objective is

$$\mathcal{L}_{\mathrm{ret}} = \mathcal{L}_{\mathrm{global}} + \beta \mathcal{L}_{\mathrm{slot}} + \gamma \mathcal{L}_{\mathrm{div}}. \tag{10}$$

The slot extractor and Sinkhorn solver are described in Appendix A.6.

3.3 Value-Guided Interaction Agent (VGIA)

VGIA controls additional evidence acquisition from the retrieval state produced by DESR. Gallery images in ChatPedes carry identity labels ℓ_j. These are benchmark metadata, used here, as in the evaluation metrics, only to group gallery images; in deployment the same grouping would correspond to track-level association of gallery images. The score of an identity is the maximum score among its gallery images:

$$c_t(y) = \max_{j:\ell_j = y} s(\mathcal{X}_t, I_j). \tag{11}$$

The online state is built from these identity scores and public progress, in three blocks. The score summary $\mathbf{s}_t \in \mathbb{R}^6$ contains the top-1 and top-2 scores, their margin, the top-1–top-6 gap, normalized

entropy, and the number of finite candidates. The history block $\mathbf{h}_t \in \mathbb{R}^{39}$ records how the ranking moved since the previous step: $\mathbf{s}_{t-1}$ and $\mathbf{s}_t - \mathbf{s}_{t-1}$, the centered current and previous top-six scores, per-candidate score changes, membership in the previous top-six panel, top-six Jaccard overlap, a top-1 flip indicator, and a flag for whether a previous step exists. The progress block $\mathbf{p}_t \in \mathbb{R}^3$ contains dialogue depth, the number of remaining legal questions, and the number of observed QA pairs. We compare three controller variants that differ only in the state they read:

$$\mathbf{z}_t^{\mathrm{Q}} = \varnothing, \qquad \mathbf{z}_t^{\mathrm{P}} = \mathbf{p}_t, \qquad \mathbf{z}_t^{\mathrm{F}} = [\mathbf{s}_t; \mathbf{h}_t; \mathbf{p}_t], \tag{12}$$

namely question-only, question+progress, and full state; $\mathbf{z}_t$ below denotes whichever of these a variant uses. No variant reads the target identity, its rank, an unrevealed answer, or a post-action score.

Each recorded question has a facet label $f(q)$ supplied by the QA annotation and a fixed descriptor $\mathbf{e}(q)$ built from its normalized text and $f(q)$. A hierarchical count model over global, facet, and question levels produces a smoothed distribution $\widehat{\mathbf{p}}(q) \in \Delta^4$ over the four transition classes defined below, and the prior logits

$$\mathbf{b}(q) = \log(\widehat{\mathbf{p}}(q) + \epsilon), \tag{13}$$

where $\epsilon > 0$ is a small numerical constant. Every variant receives $[\mathbf{e}(q); \mathbf{b}(q)]$, so the variants differ only in $\mathbf{z}_t$.

For a candidate question, define four transition classes $\mathcal{C} = \{00, 01, 10, 11\}$, whose two bits indicate whether the current top-1 prediction is correct before and after revealing that question's recorded answer. A residual controller predicts their probabilities:

$$\begin{aligned} \boldsymbol{\pi}_\psi(q \mid \mathbf{z}_t) &= \mathrm{softmax}(\mathbf{b}(q) + r_\psi(\mathbf{z}_t, \mathbf{e}(q), \mathbf{b}(q))), \\ v_\psi(q \mid \mathbf{z}_t) &= \pi_\psi^{01}(q \mid \mathbf{z}_t) - \pi_\psi^{10}(q \mid \mathbf{z}_t), \\ q_t^* &= \arg\max_{q \in \mathcal{Q}_t} v_\psi(q \mid \mathbf{z}_t). \end{aligned} \tag{14}$$

The value is a one-step expected correction benefit minus expected correction harm. At inference, the controller evaluates only the current state and the fixed candidate descriptors; the after-answer transition is used only to construct training targets.

Because v_ψ is a difference of two probabilities of the same events, it lies in $[-1, 1]$ in every state. The decisions that consume it nevertheless rely on different properties. For a fixed state and any $a(\mathbf{z}) > 0$ and $b(\mathbf{z})$,

$$\arg\max_{q \in \mathcal{Q}} v(q \mid \mathbf{z}) = \arg\max_{q \in \mathcal{Q}} \left[a(\mathbf{z})\, v(q \mid \mathbf{z}) + b(\mathbf{z})\right], \tag{15}$$

so within-dialogue ordering depends only on the order of values inside one state, whereas stopping and allocation compare values against a threshold or across states and therefore depend on cross-state calibration, which the training objective does not guarantee. We evaluate the three decisions as separate operating modes for this reason; the invariance is an elementary property and does not by itself explain any experimental outcome.

An independent stopping head estimates whether the current top-1 prediction is correct from a compact state $\bar{\mathbf{z}}_t$, consisting of the normalized entropy in $\mathbf{s}_t$ and the progress block $\mathbf{p}_t$:

$$\begin{aligned} p_\omega(\mathrm{correct} \mid \bar{\mathbf{z}}_t) &= \mathrm{sigmoid}(g_\omega(\bar{\mathbf{z}}_t)), \\ a_t &= \begin{cases} \texttt{SUBMIT}, & \mathcal{Q}_t = \varnothing \ \text{ or } t \geq K \ \text{ or } \ p_\omega(\mathrm{correct} \mid \bar{\mathbf{z}}_t) \geq \tau_{\mathrm{stop}}, \\ \texttt{ASK}(q_t^*), & \text{otherwise.} \end{cases} \end{aligned} \tag{16}$$

Equation equation 16 states the policy in general form. In the adaptive-stopping mode K is only an upper cap, set to the size of the legal pool, so `SUBMIT` is triggered by the stopping head or by exhausting the pool. The stopping experiments hold the selector fixed and vary only the stopping rule, so the `ASK` branch there follows the question-only selector rather than the full-state q_t^*; this keeps the stopping comparison free of differences in question order.

In the shared-budget mode the decision object is a dialogue rather than a question. The question order of every dialogue is fixed in advance by a single acquisition path, produced once by a trained full-state selector run to pool exhaustion and shared by all allocators. Each dialogue's next action is

Table 1: ChatPedes main results. (a) Fixed-prefix retrieval under identical recorded prefixes; * marks single-turn baselines whose input is unchanged by later QA pairs. (b) Fixed-cap interaction under matched additional-ASK budgets.

(a) Fixed-prefix retrieval

	R@1 (%)						Complete dialogue		
Method	r01	r02	r04	r06	r08	All	R@5	R@10	mAP
IRRA*	29.07	37.88	48.24	48.24	48.24	48.24	67.24	74.92	44.37
RaSa*	**31.99**	**41.48**	46.73	46.73	46.73	46.73	66.02	73.21	44.19
APTM*	31.54	40.01	42.71	42.71	42.71	42.71	60.02	67.34	39.41
TBPS-CLIP*	27.86	36.61	51.82	52.83	52.83	52.83	72.84	79.62	48.50
RDE*	31.20	40.48	52.83	52.81	52.81	52.81	70.40	77.33	47.85
OCDL*	30.09	39.74	51.30	51.29	51.29	51.29	69.47	76.97	48.15
UP-Person*	28.61	37.31	48.16	48.21	48.21	48.21	67.68	74.76	44.80
LoTLIP	17.73	23.50	35.38	41.10	41.36	41.36	63.78	72.82	35.47
DiaNA	27.23	35.34	52.59	64.88	70.92	72.43	88.76	93.05	64.61
EAVI retriever (DESR)	28.95	37.49	**54.96**	**67.19**	**73.03**	**74.63**	**90.41**	**94.72**	**66.12**

(b) Fixed-cap interaction: R@1 (%)

Additional ASK cap K	3	4	5	6	7
DESR + recorded order	55.02	61.25	65.38	68.29	69.69
DESR + question prior	**65.11**	**68.36**	**69.64**	**70.25**	**70.48**
EAVI (DESR + VGIA)	64.88	67.88	69.19	69.94	70.28

therefore already determined when the allocator runs; the allocator scores the active dialogues by the value of that next action and grants the interaction to the highest-scoring dialogue, without reordering questions within any dialogue. The deterministic scheduler is given in Appendix A.2. Fixed-cap ordering, stopping, and shared-budget allocation are evaluated as separate operating modes throughout.

3.4 Optimization Objective

We first train DESR and then freeze it while fitting VGIA. DESR is trained on dialogue prefixes whose number of retained turns is sampled at random, so that the retriever is exposed to every evidence depth it meets during interaction. For each state–question pair, replay across frozen DESR replicas yields an offline transition target $\mathbf{y}_{t,q} \in \Delta^4$ over $\mathcal{C}$. This target supervises the target-blind controller and is not an online input. With the replay weight $w_{t,q}$ defined in Appendix A.2, the value controller is trained with a four-class cross-entropy:

$$\mathcal{L}_{\text{ask}} = \mathbb{E}_{(t,q)} \left[-w_{t,q}\, \mathbf{y}_{t,q}^{\top} \log \boldsymbol{\pi}_{\psi}(q \mid \mathbf{z}_t) \right], \tag{17}$$

The stopping head is fitted independently with the pre-action top-1 correctness label:

$$\mathcal{L}_{\text{stop}} = \mathbb{E}_t \left[-y_t^{\text{top1}} \log p_{\omega,t} - (1 - y_t^{\text{top1}}) \log(1 - p_{\omega,t}) \right], \qquad p_{\omega,t} = p_{\omega}(\text{correct} \mid \bar{\mathbf{z}}_t). \tag{18}$$

Feature construction, replay weighting, prior fitting, split protocol, controller architecture, and the deterministic shared-budget scheduler are detailed in Appendix A.2.3.

4 Experiments

4.1 Experimental Setup

Protocol. We evaluate on ChatPedes Bai et al. (2025): 34,054 training images and 6,148 test queries from 1,000 identities, retrieved against the official 3,074-image gallery. Fixed-prefix evaluation gives every retriever the same first 1, 2, 4, 6, or 8 recorded QA pairs ($r01$–$r08$) or the complete dialogue (*All*), which isolates evidence matching from question selection. Interactive evaluation starts from one observed QA pair, counts only additional `ASK` actions, and covers the three operating modes of Section 3.3: fixed caps $K \in \{3, \ldots, 7\}$, shared budgets of 10–40% of all available additional actions, and adaptive stopping.

Table 2: Component and operating-mode analyses. (a) Separate ablation paths on validation. (b) Observable-state ablation under a fixed logged acquisition path (mean ± SD, $N = 3$). (c) Independent stopping evaluation; no-stopping exhausts the pool.

(a) Evidence-set alignment

Slot ext. + OT	Slot-level contrast.	Hard-OT superv.	R@1 (%)
–	–	–	80.54 ± 0.33
✓	–	–	80.72 ± 0.35
✓	–	✓	81.47 ± 0.28
✓	✓	–	81.47 ± 0.29

(c) Adaptive stopping

Setting	R@1 (%)	Mean ASK	ASK saved
No stop (full 15)	74.71	6.595	0.0%
Entropy rule	74.07	4.234	35.8%
Learned head	74.04	4.109	37.7%

(b) Shared-budget state inputs

State input	10%	20%	30%	40%	Mean
Question	48.64 ± 0.04	56.52 ± 0.11	62.74 ± 0.17	66.88 ± 0.16	58.70
Question + progress	47.77 ± 0.11	55.82 ± 0.36	62.07 ± 0.48	66.89 ± 0.42	58.14
Full state	**50.56 ± 0.26**	**61.73 ± 0.06**	**68.34 ± 0.06**	**71.96 ± 0.04**	**63.15**

Metrics. We report image-level R@1/R@5/R@10, mAP, and mean additional ASK. Fixed-prefix and fixed-cap results average three frozen DESR checkpoints, stopping thresholds are selected on development identities, and paired intervals use 2,000 identity-cluster bootstrap draws. Further protocol details and checkpoint provenance are given in Appendix A.4.

4.2 Main Results

Evidence matching. Table 1(a) compares retrieval under identical recorded prefixes. DESR improves R@1 over the released DiaNA checkpoint by 1.72–2.37 percentage points at every prefix, including 74.63 versus 72.43 on the complete dialogue. Its complete-dialogue mAP is 66.12 versus 64.61 for DiaNA. The advantage persists as QA evidence accumulates, whereas the single-turn TBPR baselines, which are competitive with one or two QA pairs, stop improving once the concatenated dialogue exceeds their native text length.

Within-dialogue question ordering. Table 1(b) evaluates question ordering at matched additional-ASK caps, with the realized number of asks identical across controllers, so the comparison varies question order alone. Three orderings are compared: the recorded order; the question prior, which ranks candidates by the value of Eq. equation 14 with $r_\psi \equiv 0$, i.e. from $\widehat{\mathbf{p}}(q)$ alone; and the full-state controller, which reads $\mathbf{z}_t^{\mathrm{F}}$. The recorded order obtains 55.02% R@1 at $K = 3$, whereas the question prior and the full-state controller obtain 65.11% and 64.88%. At $K = 7$ the three values are 69.69%, 70.48%, and 70.28%. Selecting questions by value is therefore worth roughly ten points over the recorded order at a small budget, but the retrieval state adds nothing measurable over a fitted dataset-level question prior. This is consistent with Eq. equation 15: ranking candidates inside one state does not use the cross-state calibration on which the remaining two decisions depend.

Shared-budget allocation. Table 2(b) and Fig. 4 evaluate the allocation decision. Every allocator replays the same frozen acquisition path produced by the full-state selector (Section 3.3), so each dialogue's question order is fixed in advance and the allocator only decides which dialogue advances next; the comparison therefore isolates allocation and says nothing about ordering. The learned allocators differ only in the state their value head reads, $\mathbf{z}_t^{\mathrm{Q}}$, $\mathbf{z}_t^{\mathrm{P}}$, or $\mathbf{z}_t^{\mathrm{F}}$. Full state improves over question+progress at every budget fraction, from 50.56 versus 47.77 R@1 at 10% to 71.96 versus 66.89 at 40%, a four-budget mean gain of 5.01 points (paired 95% CI [4.47, 5.55]). The gain

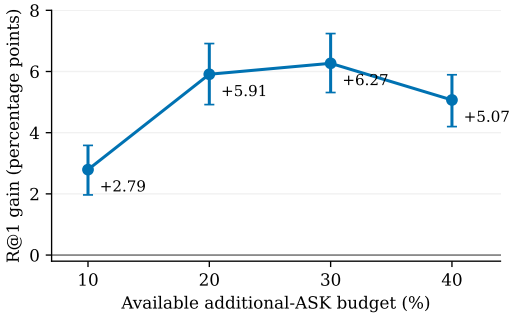

Figure 4: Retrieval feedback improves shared-budget allocation. Each point is full state minus question+progress on the same acquisition path.

Table 3: System-level efficiency comparison with DiaNA on a single NVIDIA A40 GPU under matched evaluation settings; see Appendix A.7 for detailed measurement boundaries.

Metric (resource unit)	**DESR**	DiaNA	**DESR advantage**
Loaded parameters	**442 M**	1,342 M	**3.03× fewer**
Dialogue-encoder parameters	**334 M**	1,237 M	**3.70× fewer**
Dialogue representation (s / 128-query batch)	**0.96–0.97**	6.58–6.77	**6.88–7.00× faster**
Full-test evaluation wall-clock (s)	**55.6**	331.4	**5.96× faster**
Evaluation peak allocated memory (GB)	**5.22**	12.33	**2.36× lower**

also holds against the strongest non-learning control, which advances the least confident dialogue first: +4.54 points ([3.93, 5.09]). Retrieval-state features thus matter for deciding which dialogue receives the next interaction, in the same setting where they did not matter for ordering.

Adaptive stopping. Table 2(c) evaluates stopping as a third operating mode. The selector is held fixed at the question-only controller for every row, and the threshold is chosen on development identities and then frozen, so the rows differ only in the stopping rule. Relative to exhausting the legal pool, the learned correctness head reduces mean additional ASK from 6.595 to 4.109, saving 37.7% of additional questions, while R@1 falls from 74.71% to 74.04% (-0.67 points, 95% CI $[-0.93, -0.40]$). An entropy rule, which submits once the normalized entropy in $\mathbf{s}_t$ falls below a development-calibrated threshold, reaches 74.07% at 4.234 ASK. Stopping is thus a cost–accuracy operating point rather than an accuracy improvement.

4.3 Component Analysis and Efficiency

Table 2(a) is a three-seed validation diagnostic of evidence-set supervision. Slot extraction with transport alone moves R@1 from 80.54% to 80.72%, whereas adding explicit slot-level supervision, either contrastive or Hard-OT, reaches 81.47%; most of the gain therefore appears once the alignment is supervised and does not depend on the form of supervision. The reliability gate is evaluated separately in Appendix A.6.

Table 3 compares efficiency on the same NVIDIA A40 GPU. With a smaller dialogue encoder (334 M versus 1,237 M parameters), DESR loads 3.03× fewer parameters, computes dialogue representations 6.88–7.00× faster, completes the full test evaluation 5.96× faster (55.6 s versus 331.4 s), and uses 2.36× less peak memory. Accounting boundaries are given in Appendix A.7.

4.4 Limitations

All results are on ChatPedes with its recorded QA pool, so the controller selects logged questions rather than generating them, and the DiaNA comparison uses its released checkpoint. Stopping thresholds also transfer imperfectly: under a stricter one-point tolerance, thresholds calibrated on development identities miss it on test, and a within-batch rank rule that meets it is sensitive to batch composition (Appendix A.8). We therefore report stopping as an operating point rather than as a risk guarantee.

5 Conclusion

We presented EAVI, which combines evidence-set retrieval with value-guided interaction for ChatPR. DESR improves R@1 over the released DiaNA checkpoint at every dialogue prefix with a smaller encoder, and VGIA reduces additional QA acquisitions by 37.7% at a small accuracy cost. Separating the interaction decisions shows where retrieval feedback matters: a fitted question prior already orders questions well within a dialogue, whereas retrieval state substantially improves shared-budget allocation across dialogues. Future work will study generated questions, live user responses, and other benchmarks.

AI USE STATEMENT

We used large language models (LLMs) for the following tasks. **Writing:** LLMs assisted in restructuring and editing the text of all sections and in checking that the reported numbers and protocol descriptions matched the recorded experimental outputs. **Code:** LLMs assisted in writing and debugging evaluation, efficiency-measurement, and plotting scripts. We did not use LLMs to produce research ideas, to design experiments, to generate citations or related-work claims, to create the overview figures, or to fabricate or alter experimental data.

All AI-assisted work was reviewed by the authors. Every number reported in the paper was rederived from the recorded result files, and the plotting scripts assert that each plotted value matches the corresponding table entry. LLM-assisted code was run and its outputs were checked by the authors against the logged artifacts. The authors take full responsibility for the content of this paper.

REFERENCES

Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3952–3962. IEEE, 2025.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. 2014.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2787–2797, 2023.

Nathan Kallus and Angela Zhou. Stateful offline contextual policy evaluation and learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 11169–11194. PMLR, 2022.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 791–809, 2024.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 61437–61449, 2023.

Dongsheng Li, Xinyuan Guo, Huijie Zhang, Pingting Hao, and Qiushi Xia. Quota-calibrated fine-grained alignment with context-aware marginals for text-based person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31103–31112, 2026.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025.

Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. An expected utility approach to active feature-value acquisition. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 4–pp. IEEE, 2005.

Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24245–24254. IEEE, 2025.

Jaewan Park, Solbee Cho, and Jay-Yoon Lee. Stop-rag: Value-based retrieval control for iterative rag. *arXiv preprint arXiv:2510.14337*, 2025.

Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak. PLOT: Text-based person search with part slot attention for corresponding part discovery. In *European Conference on Computer Vision (ECCV)*, pp. 474–490. Springer, 2024.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27197–27206, 2024.

Yuchao Qin, Mihaela van der Schaar, and Changhee Lee. Risk-averse active sensing for timely outcome prediction under cost pressure. *Advances in Neural Information Processing Systems*, 36: 6397–6411, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e5841df2166dd424a57127423d276bbe-Paper.pdf.

Wentao Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of MLLMs for transferable text-to-image person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17127–17137, 2024.

Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. Efficient learning by directed acyclic graph for resource constrained prediction. *Advances in neural information processing systems*, 28, 2015.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):2872–2893, 2021.

Bingqing Zhang, Zhuo Cao, Heming Du, Yang Li, Xue Li, Jiajun Liu, and Sen Wang. Quantifying and narrowing the unknown: Interactive text-to-video retrieval via uncertainty minimization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22120–22130. IEEE, 2025.

A Appendix

A.1 DESR Implementation Details

The dialogue encoder is XLM-RoBERTa-base and the visual encoder is BEiT-v2-base. Role and turn embeddings are added before slot extraction. The text and image slot extractors use shared prototype initialization but modality-specific attention updates; the released configuration uses $M = 8$ slots per modality. The transport plan is solved with Sinkhorn iterations and uniform marginals. Reliability is normalized over textual slots after transport, as in Eq. equation 6; it reweights the transported similarities without changing the transport plan. The multi-positive contrastive objectives use identity membership to define positive pairs in both directions. The attention-overlap term is applied to the text and image slot attention maps. Training uses Random Round Retaining to expose the retriever to dialogue prefixes with different numbers of observed turns: each training dialogue is kept whole with probability 0.35, and is otherwise truncated to a prefix whose length is drawn uniformly from one turn to its full length. Additional architecture and optimization settings follow the configuration reported with the released implementation.

A.2 VGIA Implementation Details

A.2.1 State and Candidate Features

The online state is computed only from information available before the next action. Gallery identity labels are used only to max-pool image scores into identity scores; the target identity and its rank are absent from every online feature. The full state $\mathbf{z}_t^{\mathrm{F}} = [\mathbf{s}_t; \mathbf{h}_t; \mathbf{p}_t]$ has 48 numeric features. The score summary $\mathbf{s}_t$ (6) contains the top-1 and top-2 scores, their margin, the top-1–top-6 gap, normalized entropy, and the number of finite candidate identities. The history block $\mathbf{h}_t$ (39) contains the previous score summary $\mathbf{s}_{t-1}$ (6) and the change $\mathbf{s}_t - \mathbf{s}_{t-1}$ (6); the centered current and previous top-six scores, per-candidate score changes, and previous-top-six membership (24); and top-six Jaccard overlap, a top-1 flip indicator, and a has-previous flag (3). At the first step, where no previous state exists, the previous quantities are copied from the current ones and the flag is zero. The progress block $\mathbf{p}_t$ (3) contains dialogue depth, the number of remaining legal questions, and the number of observed QA pairs; the implementation field named `remaining_actions` is this remaining-legal-question count.

Each candidate question is represented by a 64-dimensional signed-hash descriptor. It hashes normalized question tokens, adjacent word pairs, and a facet token, and then applies deterministic ℓ_2 normalization. A hierarchical prior is fitted on the fitting identities of the train-derived split only. It smooths global, facet, and question counts: global counts use an additive 0.001 term, and facet and question counts shrink toward the preceding level with strength 20. The resulting four class log probabilities are the prior logits $\mathbf{b}(q)$. Question-only, question+progress, and full-state controllers differ only in the state block $\mathbf{z}_t^{\mathrm{Q}}$, $\mathbf{z}_t^{\mathrm{P}}$, or $\mathbf{z}_t^{\mathrm{F}}$ of Eq. equation 12 appended to the descriptor and prior.

A.2.2 One-Step Transition Targets

For every retained replayed state–question pair, a frozen DESR replica computes the ranking before and after appending the candidate's recorded answer. Let y_t^{pre} and $y_{t,q}^{\mathrm{post}}$ denote the corresponding top-1 correctness indicators. The transition class is $c_{t,q} = (y_t^{\mathrm{pre}}, y_{t,q}^{\mathrm{post}}) \in \{00, 01, 10, 11\}$. The implementation replays each state–question pair with $J = 3$ frozen DESR replicas and forms the soft target

$$\mathbf{y}_{t,q} = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}\left[c_{t,q}^{(j)}\right] \in \Delta^4. \tag{19}$$

Here $\mathbf{1}[c]$ is the one-hot vector for class c. Replay paths define which observed prefix and legal question constitute a state; they are not an additional averaging axis after the state has been deduplicated. The value $y^{01} - y^{10}$ rewards an error-to-correct transition and penalizes a correct-to-error transition. Neither the post-answer ranking nor the future answer is passed to the controller at deployment.

A.2.3 Controller and Stopping-Head Details

The residual value controller concatenates the observable state features, the 64-dimensional question descriptor, and the four prior logits. Its input is mapped by two 128-unit ReLU layers with dropout 0.2 to four transition logits; the final residual layer is initialized to zero, so the prior is the initial prediction. Training uses only the replay-weighted four-class cross-entropy in Eq. equation 17. The replay weight gives every identity, every episode of an identity, and every depth of an episode equal mass, and splits that mass evenly over the replayed histories at a depth and then over the legal questions of each state. For a state at depth d of episode e of identity y,

$$w_{t,q} = \frac{1}{|\mathcal{Y}|\, E_y\, D_e\, H_{e,d}\, |\mathcal{Q}_t|}, \tag{20}$$

where $|\mathcal{Y}|$ is the number of fitting identities, E_y the number of episodes of y, D_e the number of replayed depths of e, and $H_{e,d}$ the number of replayed histories at depth d; the weights sum to one over the fitting partition. No pairwise ordering loss is used in this controller.

The stopping head is an independent classifier over the normalized entropy in $\mathbf{s}_t$ and the progress block $\mathbf{p}_t$, with the pre-action top-1 correctness indicator as its binary label. Its threshold is selected on the disjoint development identities by the rule in Appendix A.5 and then fixed for evaluation.

The prior and controller are fitted on train-derived fitting identities; model selection and stopping-threshold selection use disjoint development identities. Evaluation identities are excluded from all fitting and selection steps. All online state construction, prior features, replay weighting, and threshold selection exclude target labels, target ranks, unrevealed answers, and post-action scores from controller inputs. Top-1 correctness labels are used only for offline transition targets and stopping supervision.

A.2.4 SHARED-BUDGET EXECUTION

Every allocator replays one frozen acquisition path, produced once by the full-state selector run to pool exhaustion, so each active dialogue's question order is fixed before allocation begins and its next legal question is already determined. The allocator therefore ranks dialogues, not dialogue–question pairs: it scores each active dialogue by the one-step value of that dialogue's already-determined next action, computed by the value head of the allocator's state variant, and assigns the next interaction to the highest-scoring dialogue. The non-learning controls in Appendix A.3 substitute a different score in the same scheduler. A deterministic integer scheduler advances the selected dialogue by one QA and reinserts it with its next state, until the prescribed shared budget is exhausted. Ties are broken by the fixed dialogue order. No future answer or post-action score is exposed before the corresponding question is revealed.

Because the acquisition path is held fixed, this mode varies only which dialogue receives the next interaction; it does not reorder questions inside any dialogue, and its results are therefore not evidence about question ordering. Fixed-cap ordering, shared-budget allocation, and the independent stopping policy are evaluated as separate operating modes.

A.3 PROTOCOLS AND STATISTICAL AGGREGATION

Table 4: Clean fixed-K interaction results: R@1 (%), mean $\pm$ sample SD over three frozen DESR checkpoints. The gallery, recorded QA pool, and controller protocol are held fixed.

Controller	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
Recorded order	55.02 ± 0.15	61.25 ± 0.26	65.38 ± 0.25	68.29 ± 0.24	69.69 ± 0.25
Question prior	65.11 ± 0.20	68.36 ± 0.37	69.64 ± 0.31	70.25 ± 0.29	70.48 ± 0.23
Full state	64.88 ± 0.32	67.88 ± 0.26	69.19 ± 0.27	69.94 ± 0.22	70.28 ± 0.27

Fixed-prefix evaluation uses identical original QA content for every retriever. DESR results average three frozen retriever checkpoints. The clean fixed-K table (with per-checkpoint variation in Table 4) uses the same three checkpoints for each controller under the same gallery and recorded QA pool. Shared-budget allocation and stopping both use the first DESR checkpoint, but with different selectors: allocation replays the full-state selector path, whereas stopping runs the question-only selector. The main fixed-prefix mean of 74.626% and the single-checkpoint full-QA reference of 74.707% therefore have different aggregation units.

In fixed-cap interaction, K is an additional-ASK cap. Query i with N_i recorded pairs executes $\min(K, N_i - 1)$ additional actions, since the first QA is already observed. Four queries with no remaining action are retained. Answers are released only after selection; acquired QA pairs are serialized in original-position order. The recorded-order control reveals the remaining questions in their stored order. Controller states and correctness supervision use the identity scores of Eq. equation 11, whereas reported R@1 measures image retrieval.

A context-length audit over the 6,148 test episodes reports a maximum of 390 tokens for both the legal recorded pool and the complete original QA, against the 400-token model limit, so no episode is truncated in either condition. Differences between conditions therefore cannot be attributed to context truncation.

Shared-budget allocation uses, for every allocator, the path that the full-state selector produces when run to pool exhaustion. Each dialogue's question order is therefore fixed in advance, and at each step an allocator only chooses which dialogue advances by its already-determined next QA; no allocator reorders questions within a dialogue. Consequently the comparison isolates the allocation decision and cannot be read as evidence about question ordering. For budget fraction f, the total

additional-ASK budget is $\mathrm{round}(f \sum_i (N_i - 1))$. The learned allocators score a dialogue by the value v_ψ of its next action under the full-state, question-only, or question+progress value head, and each variant uses three independently fitted heads. The non-learning controls replace this score by the value implied by $\widehat{\mathbf{p}}(q)$ alone (question prior), by the current normalized entropy, advancing the least confident dialogue first (high entropy), or by the negative number of acquired QA pairs, advancing the least served dialogue first (uniform). The shared-budget mean is the arithmetic mean over $f \in \{0.1, 0.2, 0.3, 0.4\}$. These runs and the stopping configuration are separate operating modes.

For uncertainty estimation, identities are sampled with replacement, retaining all their queries, for 2,000 paired bootstrap draws. The shared-budget scheduler is rerun in every draw, charging every sampled copy separately. Intervals for the four individual budgets use the 0.00625 and 0.99375 quantiles to adjust for four comparisons. The interval [4.474,5.552] for the mean gain is an ordinary paired 95% interval, computed from the bootstrap distribution of the four-budget average. Value heads are fixed during resampling; their training-seed SD is reported separately. The development set was also used for controller model selection, and the official test set had prior experimental exposure; the follow-up controller analyses are exploratory.

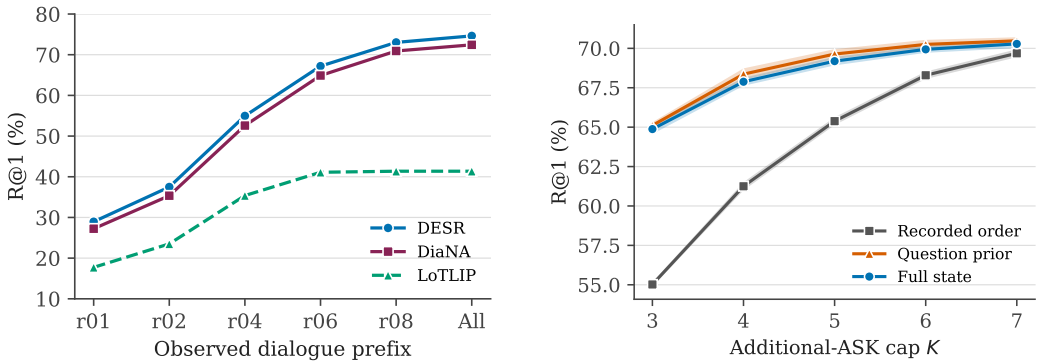

Figure 5: Supplementary retrieval and interaction curves on ChatPedes. (a) Fixed-prefix retrieval under identical observed dialogue prefixes. (b) Fixed-cap interaction at matched additional-ASK budgets, from the same three-checkpoint export as Table 1(b); lines are means and shaded bands are ± 1 sample SD over the three frozen DESR checkpoints.

A.4 Checkpoint Provenance

The cross-method table retains the initial manuscript's reproduced traditional TBPR results: IRRA, RaSa, APTM, TBPS-CLIP, RDE, OCDL, and UP-Person. The verified historical selection chains use test-selected checkpoints for IRRA (epoch 32) and RDE (epoch 36); OCDL's final run resumes an upstream test-selected epoch-7 checkpoint and uses the released auxiliary-mask/initialization configuration. These rows are marked in the main table. The other inherited TBPR rows are retained as historical reproduced results; a common development-only checkpoint-selection procedure has not been established for them.

LoTLIP uses the verified final epoch-60 checkpoint. Its training TSV matches official ChatPedes training descriptions, and its six evaluation prefixes match the original test content. DiaNA uses released weights. DESR uses the three existing retriever checkpoints for fixed-prefix comparisons, preserving their mean rather than selecting a test-best seed. LLaVA-ReID and ChatReID have no matched-protocol numerical results in this package.

A.5 Controller Comparisons and Stopping Calibration

Table 5 reports supplementary controller arms across $K = 3, \ldots, 7$. These exploratory rows use the first DESR checkpoint; the clean three-checkpoint fixed-K aggregates appear in the main table. In the notation of Eq. equation 12, question-only reads $\mathbf{z}_t^{\mathrm{Q}}$, question+progress reads $\mathbf{z}_t^{\mathrm{P}}$, and full state reads $\mathbf{z}_t^{\mathrm{F}}$. Two further controls modify the full-state controller: current-state only removes every

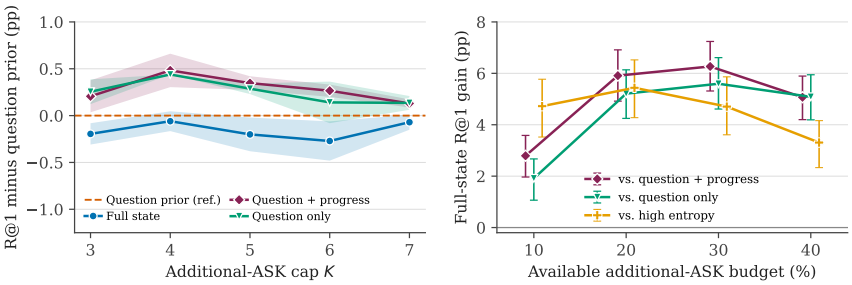

Figure 6: Additional controller comparisons. Left: within-dialogue ordering at fixed additional-ASK caps, shown as the R@1 difference of each learned controller from the question prior (dashed line at zero) within the single-checkpoint ablation of Table 5; shaded bands are ±1 sample SD over three controller seeds. Right: full-state shared-budget gains against three controls on the frozen acquisition path of Table 7; points at each budget are offset horizontally for legibility, and error bars are four-budget-adjusted 95% identity-cluster intervals. The main text reports the prespecified question+progress contrast.

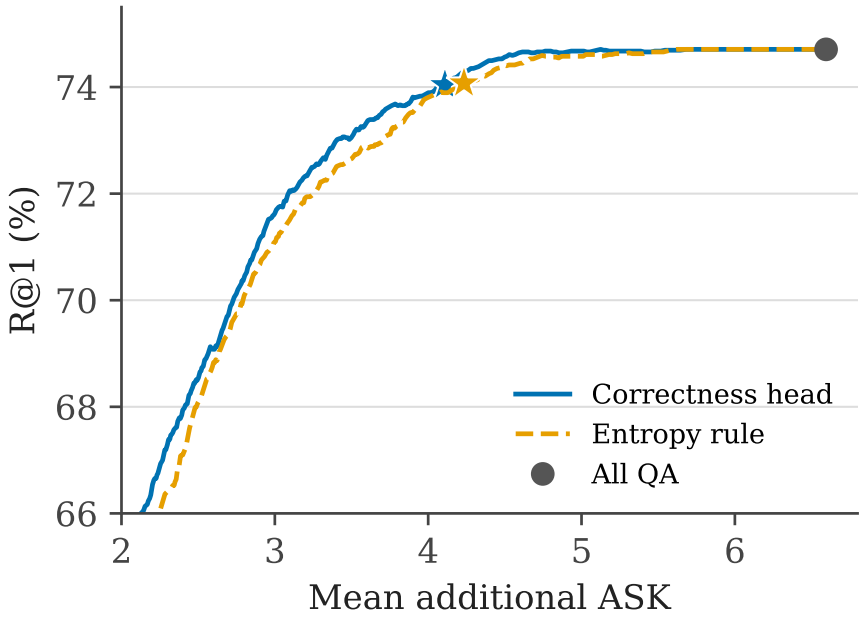

Figure 7: Stopping operating curves for the question-only selector. Each curve traces test R@1 against mean additional ASK as the threshold moves over the development grid; stars mark the configurations selected on development identities under the requirement that R@1 not fall below acquiring all QA, and the dot marks acquiring all QA.

Table 5: Within-dialogue state ablations on the first DESR checkpoint: R@1 (%), mean ± sample SD over three controller seeds for learned controllers; recorded order and question prior are single deterministic runs. The last row gives the realized ASK counts shared by all capped policies.

State input	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
Recorded order	54.81	61.89	67.23	70.98	73.39
Question prior	67.96	71.36	73.57	74.46	74.76
Question only	68.21 ± 0.13	71.80 ± 0.00	73.86 ± 0.06	74.60 ± 0.22	74.89 ± 0.07
Question + progress	68.16 ± 0.17	71.84 ± 0.18	73.92 ± 0.07	74.73 ± 0.07	74.89 ± 0.04
Full state	67.76 ± 0.11	71.30 ± 0.11	73.37 ± 0.18	74.19 ± 0.21	74.69 ± 0.08
Current state only	67.27 ± 0.57	71.08 ± 0.35	73.14 ± 0.33	74.19 ± 0.18	74.63 ± 0.13
Without prior	67.41 ± 0.21	70.89 ± 0.28	73.18 ± 0.29	74.21 ± 0.03	74.80 ± 0.05
Realized mean ASK	2.967	3.887	4.708	5.395	5.908

Table 6: Shared-budget R@1 (%), mean ± sample SD across three value-head seeds on one acquisition path.

State input	10%	20%	30%	40%
Question only	48.64 ± 0.04	56.52 ± 0.11	62.74 ± 0.17	66.88 ± 0.16
Question + progress	47.77 ± 0.11	55.82 ± 0.36	62.07 ± 0.48	66.89 ± 0.42
Full state	50.56 ± 0.26	61.73 ± 0.06	68.34 ± 0.06	71.96 ± 0.04

quantity that refers to the previous step, retaining $\mathbf{s}_t$, the centered current top-six scores, and $\mathbf{p}_t$; without prior sets $\mathbf{b}(q) = \mathbf{0}$ in Eq. equation 14. All arms retain the question descriptor $\mathbf{e}(q)$. The prespecified ordering controls are recorded order and question prior, the latter being Eq. equation 14 with $r_\psi \equiv 0$; a separate random-order run is not part of the locked evaluation package. Shared-budget absolute scores and the mean paired contrasts are reported in Tables 6 and 7; Figure 6 plots the learned ordering arms as differences from the question prior, together with the allocation gains.

The stopping candidate is chosen by requiring development R@1 to be at least its full-QA value, then minimizing mean additional ASK, with accuracy and deterministic ordering breaking ties. Correctness-head and absolute-entropy thresholds are searched on 1,001 empirical quantiles of development nonterminal states. The selected configuration is the first of the three question-only selectors with its correctness head, at threshold 0.872278; Figure 7 shows the full operating curves. Its test differences relative to full QA are -0.667 R@1 points (95% CI $[-0.928, -0.395]$) and -2.486 ASK ($[-2.590, -2.377]$). The 37.691% saving counts only additional questions; including the initially observed QA gives 32.729%.

For the same selector, averaging the three correctness heads reduces ASK by another 0.100 relative to finely calibrated entropy stopping, with a +0.016-point mean R@1 difference. This component comparison is distinct from the 37.7% saving relative to acquiring all QA. The preregistered primary configuration instead stopped once the largest remaining value $\max_{q \in \mathcal{Q}_t} v_\psi(q \mid \mathbf{z}_t)$ of the full-state controller fell below a development-calibrated threshold, under a one-point tolerance; it obtained 70.495% R@1 at 2.318 ASK and missed that tolerance on test. The development-selected correctness-head configuration reported in the main text is a subsequent analysis.

A.6 Evidence-Set Component Analyses

Table 2(a) is a three-seed validation diagnostic from the DESR development run, so its absolute R@1 values are not comparable with the test results. Its rows are separate supervision paths rather than a joint factorial design. Hard-OT supervision replaces the soft transport target with candidate-hard OT targets and is an alternative to the slot-level contrastive objective, not an addition to it.

The reliability gate is evaluated separately, because the development diagnostic above does not isolate it. Enabling and disabling the gate on otherwise identical models, paired over the same replayed states and resampled by identity, gives an enabled-minus-disabled identity-macro R@1 of $+0.254$ points (95% CI $[+0.163, +0.361]$) and a mean-reciprocal-rank difference of $+0.00215$ ($[+0.00161, +0.00275]$), across three training seeds. The measurement uses a replay population

Table 7: Full-state gains averaged over four shared budgets. Intervals are paired 95% intervals for each mean contrast, not simultaneous intervals over the control family.

Control	Mean gain (pp)	95% CI (pp)
Question + progress	+5.010	[4.474, 5.552]
Question only	+4.451	[3.916, 4.966]
Question prior	+4.613	[4.059, 5.192]
High entropy	+4.543	[3.932, 5.086]
Uniform	+7.492	[6.922, 8.050]

Table 8: Supplementary A40 efficiency quantities, reported separately because their accounting boundaries differ from the main table. Per-batch retrieval covers the same operation as the full-test wall-clock at a smaller scale; the dialogue-representation figure is a peak-allocation increment over the state already resident when that segment starts, not a standalone process peak.

Quantity (accounting boundary)	**DESR**	DiaNA	Ratio
Cached-gallery retrieval (s / 128-query batch)	**1.09–1.16**	6.90–7.12	6.12–6.33×
Dialogue-representation peak increment (GB)	**3.38**	6.79	2.01×
Text-tower-only GPU time ratio		–	7.04–7.16×

of 892 identities with identity-level R@1 of about 0.34, so its magnitude is not comparable with test R@1. It supports the gate as a useful post-transport aggregation component; the corresponding dilution-slope interaction is not significant, so it does not show that the gate mitigates evidence dilution.

A.7 Efficiency

All efficiency quantities are measured on the same NVIDIA A40 GPU with batch size 128 and sequence length 400, on the same evaluation data and gallery, and at each method's native precision.

Dialogue representation is GPU time recorded with CUDA events around the native text tower and the evidence-feature functions, excluding tokenization. It is reported per 128-query batch rather than per query, because the runtime pads a partial batch to 128 and a per-query figure would not describe single-query latency. For each prefix length the measurement performs two warm-up passes followed by five timed repetitions over the same 128 real test dialogues, giving 0.957/0.963/0.968 s for DESR and 6.583/6.671/6.771 s for DiaNA at $r01$/$r04$/complete dialogue. The text-tower-only ratio is 7.04–$7.16\times$.

The full-test wall-clock is a separate measurement over all 6,148 test dialogues against the 3,074-image gallery with a warm cached gallery, in 49 batches, timed after two warm-up batches. It includes serialization, tokenization, scoring, and CPU score transfer, and excludes model load, gallery cache construction, user response latency, and controller computation. The per-batch figures in Table 8 scaled to 49 batches predict 57 s and 349 s, against measured 55.6 s and 331.4 s, so the two measurements are consistent.

Evaluation peak allocated memory is the peak over that full-test pass. The dialogue-representation peak in Table 8 is a different quantity: it is the peak allocation during the dialogue-representation segment minus the allocation already resident when the segment begins, so gallery features remain resident and the value is an increment rather than a standalone process peak. Parameter counts cover all loaded modules, including modules unused on the current path.

A.8 Stopping Transfer and Batch Sensitivity

When a stopping rule is additionally required to stay within one R@1 point of full acquisition, absolute thresholds calibrated on development identities do not meet the tolerance on test. All twelve configurations, an entropy threshold and three independently trained correctness heads on each of three selectors, miss it by 3.17 to 3.74 points, and the same protocol applied to DiaNA misses by 11.08 points. Because the correctness heads and the entropy rule fail alike, the gap reflects a shift

between the calibration and deployment distributions rather than a defect of a particular head. A within-batch rank rule meets the tolerance (-0.10 to -0.11 points, and -0.73 for DiaNA) at the cost of conditioning on the batch, as measured next.

The within-batch rank rule discussed in Section 4.4 proceeds round by round: at each depth it submits the dialogues whose normalized entropy lies in the lowest locked fraction among the dialogues of the batch that are still running, so the pool never contains dialogues that have already submitted. On the full-state selector path, random batches of 32 give 74.480–74.593% R@1 and 4.856–4.857 ASK; batches of 128 give 74.528–74.642% and 5.054–5.061 ASK; batches of 512 give 74.545–74.577% and 5.165–5.178 ASK. Sorting by initial entropy before batching changes R@1 to 58.360%, 63.696%, and 66.493% for batch sizes 32, 128, and 512. This sensitivity concerns the batch-rank rule only; the correctness head reported in the main text operates per dialogue.