TRACE: TASK-ADAPTIVE REASONING AND REPRESENTATION LEARNING FOR UNIVERSAL MULTIMODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal multimodal retrieval must map both explicit matches and unresolved compositional intents into a shared embedding space. Intermediate reasoning can clarify a query's target semantics, but generating it for every input adds latency and may weaken retrieval. We introduce TRACE, a framework that learns retrieval representations through selective query-side reasoning. Its first autoregressive prediction selects immediate embedding or a retrieval-oriented reasoning sequence; both routes produce a vector from the hidden state predicting the same embedding marker. Joint generation and contrastive training couple intent resolution to representation learning, while directly encoded candidates remain indexable offline. We construct M-BEIR-CoT with 518K direct examples and 575K filtered reasoning examples to supervise both behaviors. TRACE achieves a 58.8 average score on M-BEIR, improving a same-backbone LamRA-Ret reproduction by 2.1 points, and improves over LamRA-Ret in 11 of 13 distribution-shifted settings. On separate MSCOCO and CIRR controls, it attains higher recall than independently trained direct-only and always-reasoning policies, with approximately $1.9\times$ the query throughput of always reasoning. CIRR interventions and ablations further support reasoning-enabled inference, joint representation learning, and query-only generation. Together, these results demonstrate selective reasoning within a shared, index-compatible multimodal retriever.

1 INTRODUCTION

Universal multimodal retrieval (UMR) seeks a single model that retrieves text, images, and image–text pairs from textual, visual, or interleaved queries (Wei et al., 2024; Lin et al., 2024; Liu et al., 2025b; Zhang et al., 2024c). The challenge extends beyond aligning modalities: a query may describe the target directly or specify how it differs from what is shown. For example, finding a similar garment with a different sleeve style requires separating the requested modification from the visual properties to preserve. Visual evidence retrieval likewise requires identifying what information a relevant item should contain. In these cases, constructing the target semantics is part of constructing the query representation (Vo et al., 2019; Wu et al., 2021; Baldrati et al., 2023; Vaze et al., 2023).

Multimodal large language models provide both representation and generation capabilities, but many retrievers use them as direct encoders (Jiang et al., 2024a;b; Lin et al., 2024; Liu et al., 2025b). Generated context and intermediate reasoning offer another way to resolve implicit intent before embedding (Cui et al., 2025; Liu et al., 2025a; Lan et al., 2025; Jiang et al., 2026). This raises two connected design questions: how should the generated sequence shape a discriminative representation, and when should the model generate it? An external reasoner can supply text to an encoder, but this separates generation from the state used for matching. Always generating with a shared model avoids that separation, yet adds sequential cost even when the input already specifies the target. Recent adaptive embedding methods use learned controllers or routing gates to address this cost (Wang et al., 2026; Zhang et al., 2026).

We introduce **TRACE**, which brings selective reasoning and representation learning into one autoregressive interface. Given a query and its retrieval instruction, the model predicts either the embedding marker <|emb|> immediately or a text token that starts a retrieval-oriented reasoning sequence.

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.

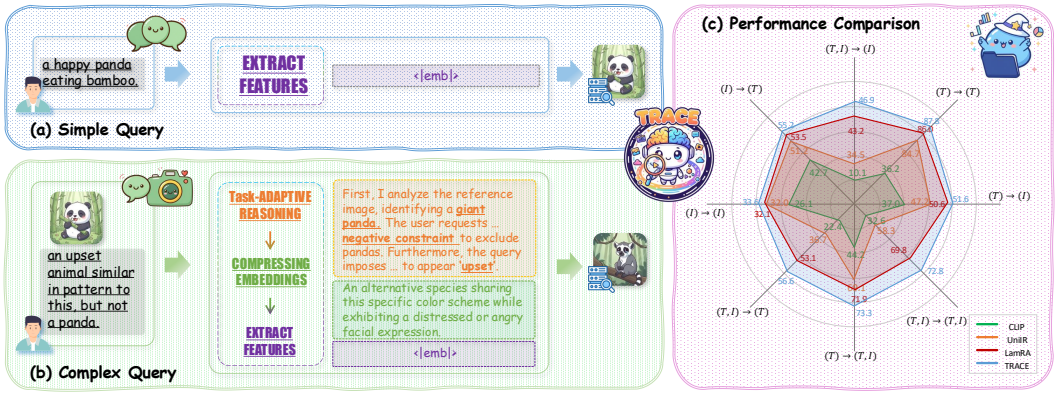

Figure 1: **TRACE overview.** Explicit matching queries use direct encoding; queries requiring intent resolution can generate a reasoning sequence before embedding. The examples illustrate these two behaviors. The right panel summarizes modality-level results for the selected baselines shown; Table 1 gives the setting-level benchmark comparison.

Both routes end at the same marker and use the hidden state that predicts it as the retrieval vector. Generation supervision teaches the output behavior, while a contrastive objective trains that state to distinguish relevant candidates. The generated sequence therefore forms part of the representation-learning process, rather than serving only as an independently produced rewrite. Route selection requires no additional classifier, separate reasoning adapter, or inference-time teacher.

TRACE also separates the computational roles of queries and candidates. A query expresses an information need that may require resolving a change, exclusion, or missing relation. A candidate is a content item against which many different queries are matched. We restrict variable computation to queries and encode candidates directly. This design lets the query representation incorporate generated context while retaining a fixed offline candidate index and the standard cosine-similarity search interface.

To learn both behaviors, we construct **M-BEIR-CoT**, a corpus of 1,093,753 examples derived from M-BEIR training pairs. Direct examples teach immediate embedding; filtered reasoning examples teach intent decomposition and target-state description before embedding. The teacher may access the positive candidate during offline trace construction, but TRACE receives only the query and retrieval instruction when encoding a query. A single training stage optimizes generation and retrieval on this mixture.

The experiments connect these design choices to retrieval performance. TRACE achieves 58.8 on M-BEIR versus 56.7 for a same-backbone LamRA-Ret reproduction, and improves over LamRA-Ret in 11 of 13 distribution-shifted settings. Separate MSCOCO and CIRR controls compare fixed and adaptive training policies. On CIRR, forcing a trained TRACE checkpoint to bypass reasoning reduces R@5 by 3.63 points, and an external supervised reasoner followed by the TRACE embedder performs below the end-to-end pipeline. Component and candidate-side ablations further examine how the reasoning sequence becomes a retrieval representation.

Our contributions are: (i) a shared UMR model that combines first-token route selection with a common retrieval state, enabling selective query-side reasoning and offline candidate indexing; (ii) M-BEIR-CoT and a joint generation–contrastive training scheme that supervise both immediate embedding and intent resolution; and (iii) an empirical study connecting broad retrieval performance and accuracy–throughput trade-offs to inference interventions, representation choices, and query–candidate asymmetry.

2 Related Work

Universal retrieval and embedding. CLIP and ALIGN learn shared image–text spaces through contrastive training (Radford et al., 2021; Jia et al., 2021). UniIR extends retrieval to eight query–candidate modality configurations and introduces M-BEIR (Wei et al., 2024). E5-V, VLM2Vec,

MM-Embed, GME, and LamRA adapt vision-language backbones to embedding tasks (Jiang et al., 2024a;b; Lin et al., 2024; Zhang et al., 2024c; Liu et al., 2025b); JFE, ReT-2, and PUMA study fusion, recurrent interaction, and efficient architectures (Huang et al., 2025; Caffagni et al., 2025; Lyu et al., 2025). U-MARVEL combines progressive training, hard-negative mining, and reranker distillation (Li et al., 2025). It reports a stronger M-BEIR local-pool average of 63.2 with a single distilled retriever. TRACE uses teacher-generated route and sequence supervision, but no reranker-score distillation. We discuss U-MARVEL separately from Table 1 because its supervision and training budget differ; its score remains relevant benchmark context, rather than a matched test of selective reasoning. MMEB-style universal multimodal embedding (UME) additionally casts classification, visual question answering, and grounding as ranking problems (Meng et al., 2025). TSEmbed studies task conflict in that broader setting (Wu et al., 2026b). TRACE focuses on retrieval of indexable content; its M-BEIR scores should not be compared directly with full MMEB averages.

Reasoning and adaptive computation. Think Then Embed generates embedding-oriented context before encoding, while Reasoning Guided Embeddings combines reasoning and embedding supervision (Cui et al., 2025; Liu et al., 2025a). UME-R1 and Embed-RL explore reasoning-driven embeddings and reinforcement learning (Lan et al., 2025; Jiang et al., 2026). UME-R1 also reports complementarity between discriminative and generative embeddings under oracle selection. More closely related, MMEmb-R1 combines pair-aware reasoning selection with reinforcement-learning control, and Think When Needed uses a routing gate with separate reasoning and embedding LoRA adapters (Wang et al., 2026; Zhang et al., 2026). PLUME explores latent reasoning, and RIME uses rewriting as an interface for generative embeddings (He et al., 2026; Wu et al., 2026a). TRACE learns first-token routing and a shared retrieval state with joint supervised training, without a separate gate, split adapters, or RL. Its query-only variable computation preserves independently indexed candidate embeddings.

Related mechanisms beyond multimodal embeddings. O1 Embedder combines thought generation with contrastive retrieval (Yan et al., 2025). Thinkless and AdaptThink learn when to use extended reasoning for language tasks (Fang et al., 2025; Zhang et al., 2025). These works establish related ingredients in text representation and adaptive generation. Retrv-R1 applies reasoning to candidate-aware multimodal retrieval (Zhu et al., 2025); its candidate interactions differ from TRACE's independently indexed dense representations. TRACE studies selective reasoning and representation learning together under the shared embedding interface of UMR.

3 METHOD

3.1 CONSTRUCTING M-BEIR-COT

Retrieval pairs identify a relevant target but do not specify the computation needed to encode a query. We construct M-BEIR-CoT from the M-BEIR training splits through route assignment, task-specific generation, and two-stage filtering (Figure 2). Validation and test examples are excluded from every construction stage.

Route assignment and trace generation. GPT-4.1 assigns a route to each query, rather than assigning one route to an entire dataset. Queries with sufficiently explicit target semantics receive direct supervision: their output is only the embedding marker $e = \texttt{<|emb|>}$. Queries requiring intent resolution receive a concise retrieval-oriented sequence before e. For composed retrieval, the teacher describes the source state, requested change, and context to preserve. For evidence retrieval, it identifies entities, evidence, or answer characteristics expected in a relevant item. The raw generation separates the reasoning and target description into `<reasoning>` and `<answer>` blocks. Route labels are teacher-provided supervision rather than measured counterfactual retrieval gains.

Quality control. Rule-based trace filtering requires exactly one reasoning block followed by one answer block, with 16–256 and 2–64 teacher-tokenizer tokens, respectively. It rejects malformed tags, repetition, refusals, generic boilerplate, and composed-retrieval answers that simply repeat the source caption. Normalized answers and query–answer pairs are deduplicated. Model-based verification then checks consistency with the query and positive target, retaining positive decisions with confidence above 0.7. For image targets, verification checks whether the requested change is

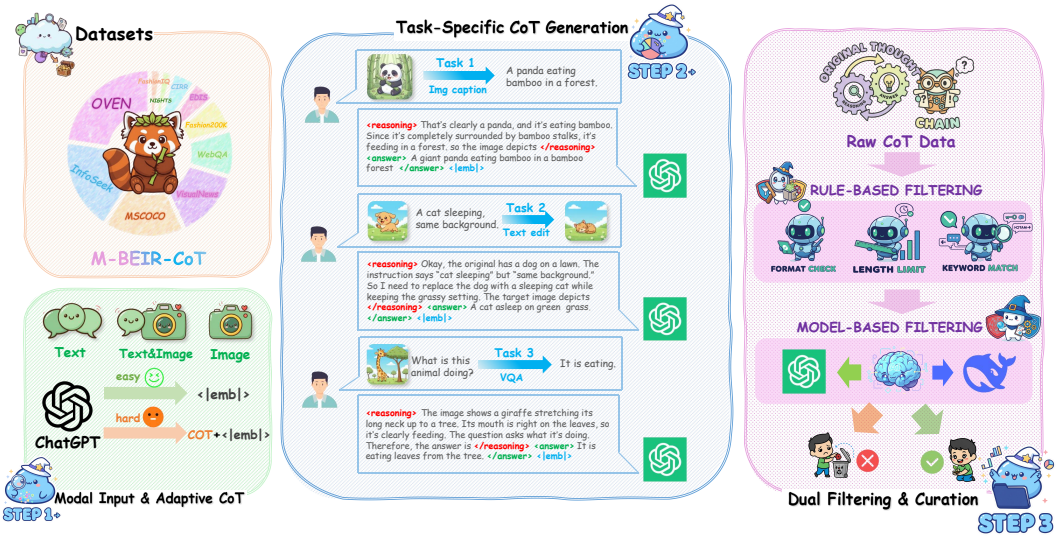

Figure 2: **M-BEIR-CoT construction.** A teacher assigns direct or reasoning supervision to training queries, task-specific prompts generate retrieval-oriented sequences, and rule-based and model-based filters check the outputs. The construction tags shown in the original illustration are used for parsing and removed before TRACE training. The examples illustrate the construction process.

applied while preserving invariant context; for text targets, it checks entailment against the target title or summary and rejects overly generic answers.

Training targets and target access. The positive candidate grounds offline generation and verification, but is excluded from TRACE's query context during both training and evaluation. We also filter explicit identifiers, URLs, complete target captions, and long copied spans to reduce leakage. Before training, construction-only tags are stripped from accepted traces; the natural-language reasoning and answer are followed by e. Direct examples retain only $[e]$. The final corpus contains 518,311 direct and 575,442 reasoning examples, giving the model explicit supervision for both immediate embedding and intent resolution. Appendix D provides the complete checks, and Appendix E documents the prompt interface. Section 4 evaluates a target-blind CIRR control to test the role of target access during trace construction.

3.2 Selective Reasoning and Shared Representations

Let Q be a multimodal query, $C \in \Omega$ a candidate item, and p_u^q, p_u^c the query and candidate instructions for retrieval relation u. TRACE maps both sides to one embedding space with a shared backbone f_θ (Figure 3). Let $x_{1:L}^q$ denote the tokenized query context, including its instruction and visual tokens. Greedy decoding selects the first output token:

$$\hat{y}_1 = \arg\max_{y \in \mathcal{V}} P_\theta(y \mid x_{1:L}^q), \qquad S_\theta(Q) = \begin{cases} [e], & \hat{y}_1 = e, \\ [r_1, \ldots, r_K, e], & \hat{y}_1 \neq e. \end{cases} \tag{1}$$

Predicting e selects direct encoding ($K = 0$); otherwise the model generates a retrieval-oriented sequence until it predicts e. The task instruction and query content can both influence this decision. There is no additional routing network or manually selected inference threshold.

The state predicting the embedding marker. In a causal decoder, the state at position j predicts the token at $j + 1$. If the query produces K reasoning tokens before e, we extract

$$\mathbf{e}_q = \mathbf{h}_{L+K}^q, \qquad \mathbf{e}_c = \mathbf{h}_{L_c}^c, \qquad s(Q, C) = \frac{\mathbf{e}_q^\top \mathbf{e}_c}{\|\mathbf{e}_q\|_2 \|\mathbf{e}_c\|_2}. \tag{2}$$

Here L_c is the candidate-context length; candidates always use the direct sequence $[e]$. For a direct query, the representation is the final prompt state $\mathbf{h}_L^q$. For a reasoning query, it is the state at the

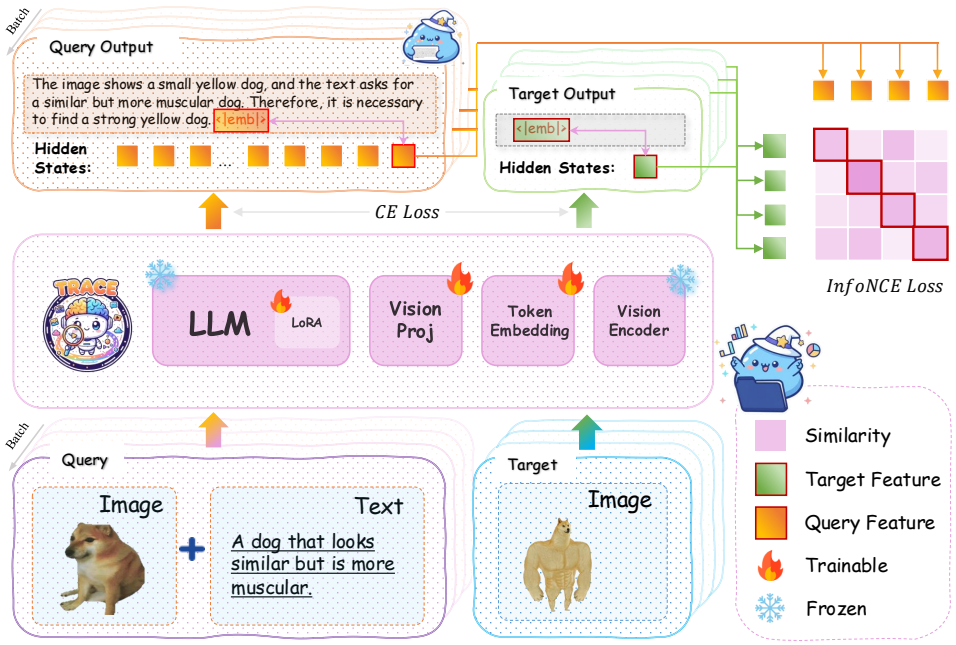

Figure 3: **TRACE architecture and joint training.** The reasoning route generates an intent-resolution sequence before the embedding marker; the direct query route omits that sequence. Query and candidate features are the hidden states predicting the marker. Query-output cross-entropy and query–candidate InfoNCE train the shared model, with a frozen vision encoder and trainable projector and LoRA parameters. Candidates use direct encoding and can be indexed offline.

final generated token before the marker. Through causal attention, this state combines the original multimodal context with the intermediate sequence and is directly optimized for retrieval. The extraction rule avoids an extra forward step to consume e. At the maximum decoding length, the implementation uses the final generated-token state if e has not appeared (Appendix B).

Query-side computation, candidate-side indexing. Query generation can resolve an information need before matching. Candidate representations instead describe reusable content items and are computed once for the index associated with a candidate instruction and retrieval setting. They do not depend on the current query or its generated trace. TRACE shares model parameters across both sides while assigning them different output behaviors: selective generation for queries and immediate embedding for candidates. Retrieval then ranks the fixed candidate vectors by cosine similarity, independently of how many reasoning tokens the query generated.

3.3 Joint Generation and Retrieval Learning

For a supervised output $Y_i = (y_{i,1}, \ldots, y_{i,T_i})$ ending in e, the generative objective is masked causal-language-model cross-entropy:

$$\mathcal{L}_{\text{gen}} = \text{CE}_{\text{masked}}\big(\{P_\theta(y_{i,t} \mid x_i^q, y_{i,<t})\}_{i,t}\big). \tag{3}$$

Only output positions are supervised. For a direct example, the single target token teaches immediate termination; for a reasoning example, the same objective supervises the initial route, trace, and final marker. Under teacher forcing, the pre-marker state is available for contrastive training without differentiating through an inference-time discrete routing decision.

For a contrastive batch of $B = 256$ positive pairs, normalized query and candidate representations are aligned using InfoNCE (Oord et al., 2018):

$$\mathcal{L}_{\text{ret}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s(Q_i, C_i)/\tau)}{\sum_{j=1}^{B} \exp(s(Q_i, C_j)/\tau)}, \qquad \mathcal{L} = \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{\text{ret}}\mathcal{L}_{\text{ret}}. \tag{4}$$

Table 1: **Selected M-BEIR local-pool comparisons across 16 settings.** t and i denote text and image. Metrics are R@5, except Fashion200K (F2K) and FashionIQ (FIQ), which use R@10. VN and InfoS denote VisualNews and InfoSeek. †: our Qwen2.5-VL-7B LamRA-Ret reproduction. Reranker-distilled U-MARVEL is discussed separately in Section 2. Avg. denotes the benchmark average. Bold and underline mark the best and second-best scores among the displayed methods.

	$t \to i$			$t \to t$	$t \to (i,t)$		$i \to t$			$i \to i$	$(i,t) \to t$		$(i,t) \to i$		$(i,t) \to (i,t)$		
Method	VN R@5	COCO R@5	F2K R@10	WebQA R@5	EDIS R@5	WebQA R@5	VN R@5	COCO R@5	F2K R@10	NIGHTS R@5	OVEN R@5	InfoS R@5	FIQ R@10	CIRR R@5	OVEN R@5	InfoS R@5	Avg.
UniIR-CLIP	**42.6**	81.1	18.0	84.7	59.4	78.7	**43.1**	92.3	18.3	32.0	45.5	27.9	24.4	44.6	67.6	48.9	50.6
JFE	34.6	78.5	37.2	**88.7**	54.3	82.4	33.1	90.0	36.9	27.8	46.0	35.6	31.8	54.0	72.7	61.1	54.0
PUMA	34.4	79.9	25.4	86.1	57.2	78.2	34.2	90.5	27.9	32.0	53.0	47.3	30.4	49.1	74.1	65.0	54.1
ReT-2	38.9	**84.8**	**50.0**	76.3	53.7	78.4	42.0	**95.0**	**52.2**	31.5	54.1	32.3	35.3	57.1	72.1	48.3	56.4
LamRA-Ret	41.6	81.5	28.7	86.0	62.6	81.2	39.6	90.6	30.4	32.1	54.1	52.1	33.2	53.1	76.2	63.3	56.6
LamRA-Ret†	41.3	81.8	29.0	85.7	62.9	80.9	40.2	90.8	30.1	32.4	53.8	52.4	32.9	53.4	75.9	63.6	56.7
TRACE	42.1	82.3	30.5	87.8	**64.1**	**82.5**	41.2	91.3	33.2	**33.6**	**57.4**	**55.8**	**36.4**	**57.3**	**78.5**	**67.1**	**58.8**

We use $\tau = 0.07$ and $\lambda_{\text{gen}} = \lambda_{\text{ret}} = 1$. Generation teaches route execution, while retrieval supervision shapes the final state to distinguish the positive candidate from in-batch negatives. Their shared parameters couple the output behavior to the retrieval representation during training. Thus, both an immediate output and a longer generated sequence terminate in a state trained for the same matching objective.

4 Experiments

We evaluate three aspects of TRACE: retrieval quality across modalities and distribution shifts, the accuracy and cost of selective computation, and the representation choices that connect reasoning to retrieval.

4.1 Setup and Comparison Protocols

Benchmarks. TRACE is trained on M-BEIR-CoT using only M-BEIR training pairs. We evaluate the M-BEIR local-pool protocol: 16 settings across eight modality configurations and ten datasets, with R@5 except Fashion200K and FashionIQ (R@10). The benchmark average weights settings equally. Transfer evaluation covers 13 settings across ten datasets: 11 retrieval settings and two image–text matching diagnostics (CC-Neg and Sugar-Crepe). Some reuse M-BEIR image sources; Table 2 and Appendix F specify metrics and overlap.

Baselines and implementation. We compare UniIR, JFE, PUMA, ReT-2, and LamRA-Ret, including our Qwen2.5-VL-7B LamRA-Ret reproduction. Transfer baselines also include CLIP, Long-CLIP, E5-V, MagicLens, and EVA-CLIP (Radford et al., 2021; Zhang et al., 2024a; Jiang et al., 2024a; Zhang et al., 2024b; Sun et al., 2024). TRACE uses Qwen2.5-VL-7B (Bai et al., 2025), a frozen vision encoder, trainable projector, and LoRA (Hu et al., 2021) (rank 128, scale 256). We train for one epoch on 16 H20 GPUs with AdamW at 2×10^{-4}, cosine decay, and task-aware sampling. Four accumulation steps over 256 contrastive pairs give optimizer batch size 1024. Inference uses greedy decoding with a KV cache (Appendix B).

Separate experimental questions. The main benchmark uses the full training mixture; MSCOCO and CIRR policy controls are trained separately with matched backbones, pairs, schedules, and update counts. Hence CIRR's main-table 57.3 and control 57.03 belong to different settings. Forced Direct instead intervenes on a trained TRACE checkpoint. CIRR ablations report means over three distinct random seeds without standard deviations; small differences are interpreted descriptively.

Table 2: **Distribution-shifted evaluation: 13 settings across ten datasets.** R@1 is used except CIRCO (MAP@5), MT-FIQ (R@5), and ITM (accuracy). Share4V, Urban, VisD, VIST, MT-FIQ, and Sugar denote ShareGPT4V, Urban-1K, Visual Dialog, Visual Storytelling, multi-turn FashionIQ, and Sugar-Crepe. Image-source overlap is documented in Appendix F; these settings are not uniformly data-disjoint.

	$t \to i$			$i \to t$			Composed		Dialog	Interleaved		ITM	
Method	Share4V	Urban	Flickr	Share4V	Urban	Flickr	CIRCO	GeneCIS	VisD	VIST	MT-FIQ	CC-Neg	Sugar
CLIP-L	84.0	52.8	67.3	81.8	68.7	87.2	4.0	13.3	23.7	0.6	17.7	66.7	73.0
Long-CLIP-L	**95.6**	86.1	76.1	**95.8**	82.7	89.3	5.7	16.3	37.9	1.1	18.5	76.3	80.9
UniIR-CLIP	85.8	75.0	78.7	84.1	78.4	94.2	12.5	16.8	26.8	0.6	39.4	79.9	80.3
E5-V	86.7	84.0	79.5	84.0	82.4	88.2	24.8	18.5	54.6	10.0	19.2	83.2	84.7
MagicLens-L	85.5	59.3	72.5	60.9	24.2	84.6	29.6	16.3	28.0	3.3	22.6	62.7	75.9
EVA-CLIP-8B	91.2	77.8	80.8	93.1	80.4	95.6	6.0	13.1	23.2	1.2	22.1	59.4	81.7
EVA-CLIP-18B	92.1	81.7	83.3	94.0	83.3	**96.7**	6.1	13.6	24.7	1.0	21.9	63.8	83.1
LamRA-Ret	93.3	**95.1**	82.8	88.1	**94.3**	92.7	33.2	18.9	62.8	23.1	60.9	79.6	85.8
TRACE	94.9	94.8	**84.5**	89.1	94.1	94.5	**34.8**	**20.5**	**65.4**	**25.8**	**63.2**	**84.1**	**87.5**

Table 3: **Independently trained policy controls.** Models use matched backbones, training pairs, candidate pools, schedules, and update counts within each dataset. QPS measures online query encoding on one H20 at batch size 1, excluding candidate indexing.

	MSCOCO (text $\to$ image)		CIRR	
Training policy	R@5 ↑	QPS ↑	R@5 ↑	QPS ↑
Direct-Trained	87.40	**15.68**	53.06	**12.15**
Always-Reasoning-Trained	63.90	4.45	54.73	3.42
TRACE	**89.10**	8.25	**57.03**	6.48

4.2 Retrieval Performance and Transfer

M-BEIR. TRACE achieves the highest average among the selected baselines in Table 1: 58.8 versus 56.7 for same-backbone LamRA-Ret, with improvements in all 16 settings. Gains include 3.6 and 3.4 points on OVEN and InfoSeek text retrieval, and 3.5 and 3.9 on FashionIQ and CIRR. ReT-2 also leads on Fashion200K and MSCOCO image-to-text retrieval. The backbone-matched comparison controls model family but does not separate routing from additional synthetic supervision.

Distribution shifts. TRACE improves over the original Qwen2-VL-7B LamRA-Ret in 11 of 13 settings. Gains include 4.5 points on CC-Neg, 2.7 on Visual Storytelling, 2.6 on Visual Dialog, and 2.3 on multi-turn FashionIQ. CIRCO and GeneCIS each improve by 1.6 points. The two Urban-1K directions decrease by 0.3 and 0.2 points. Long-CLIP and EVA-CLIP remain stronger in some caption-retrieval settings. The gains span exclusion constraints, compositional changes, and multi-turn context, showing that the trained representation transfers beyond the original query formats.

4.3 Selective Computation: Accuracy and Cost

The fixed-policy controls are separately trained models. Direct-Trained learns immediate e; Always-Reasoning-Trained learns a trace followed by e for every query, using additional supervision from the same teacher and filters on MSCOCO and CIRR. TRACE retains its learned adaptive behavior; the controls are not inference modes of that checkpoint.

On MSCOCO, always reasoning gives 63.90 R@5 versus 87.40 for direct training; TRACE reaches 89.10 at $8.25/4.45 = 1.85\times$ the throughput of always reasoning (Table 3). On CIRR, TRACE improves on direct training from 53.06 to 57.03, with $6.48/3.42 = 1.89\times$ the throughput of always reasoning. Direct encoding remains faster on both tasks: TRACE trades extra computation for recall while reducing cost relative to unconditional reasoning.

Table 4: **CIRR controls under the M-BEIR local pool.** Forced Direct changes only inference in TRACE; Direct Training is trained separately. External SFT-CoT is a two-stage pipeline, not a route-only intervention.

Configuration	R@5
Direct Training	53.06
TRACE, Forced Direct	53.40
External SFT-CoT + TRACE Embedder	55.00
TRACE, End-to-End Adaptive	**57.03**

Table 5: **Reasoning-to-representation ablations on CIRR.** Each block compares separately trained variants; scores average three runs in the M-BEIR local pool. Candidate ablations keep query supervision and adaptive inference fixed. Complete metrics and settings appear in Appendices I–J.

Design choice	Variant	R@5
Query supervision	Answer only	53.67
	Reasoning only	53.41
	Reasoning + answer	**57.03**
Extraction state	After <|emb|>	52.25
	Structured </think> state	53.06
	State predicting <|emb|>	**57.03**
Candidate encoding	Full reasoning sequence	18.90
	Answer only	19.78
	Direct encoding	**57.03**

Observed routing behavior. On MSCOCO, 96% of queries take the direct route; on CIRR, 62% take the reasoning route. In a separate 1,000-query analysis, reasoning activates for 93% of queries containing "not" or "without" and 74% of scenes with more than three detected entities. Short single-object descriptions and short factual queries use direct encoding in 96% of cases. These frequencies characterize behavior; the following intervention tests the benefit of executing reasoning.

4.4 Reasoning and Joint Representation Learning

Reasoning with parameters fixed. Forcing the trained TRACE checkpoint to predict e immediately reduces CIRR R@5 from 57.03 to 53.40 (Table 4). This 3.63-point difference isolates the benefit of allowing its reasoning-enabled policy at inference, complementing the separately trained policy comparisons.

End-to-end states versus an external textual handoff. An external reasoner trained with the same reasoning supervision and followed by the TRACE embedder reaches 55.00 R@5 versus 57.03 for end-to-end TRACE. This supports joint generation and representation learning in the tested system. Trace conditioning also differs, so the 2.03-point gap cannot be attributed solely to a textual bottleneck.

Target access during trace construction. A target-blind CIRR model is trained from scratch with matched pairs, route labels, and optimization budget, hiding positive candidates from trace generation and verification. It reaches 56.98 R@5 versus 57.03 with grounded traces and 53.06 with direct training (Appendix H.2), retaining 98.7% of the gain. Thus the CIRR improvement persists without target-visible traces. Route labels remain fixed, and the 0.05-point gap does not establish a significant grounding benefit.

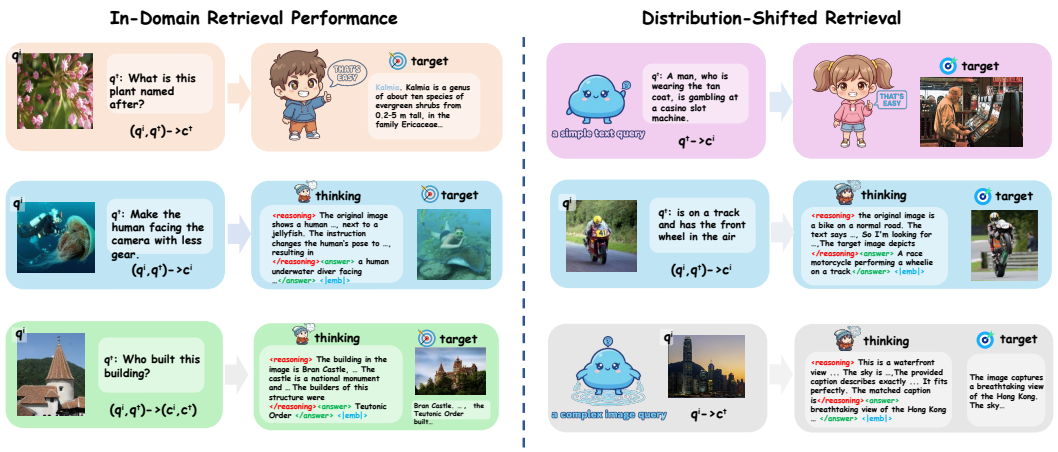

Figure 4: **Qualitative examples of direct and reasoning-enhanced retrieval.** Left: in-domain examples; right: distribution-shifted examples. Text is abbreviated and copy-edited for readability. The <reasoning> and <answer> tags are display annotations, not literal model-output delimiters.

4.5 Which Representation Choices Matter?

Supervision and extraction. Table 5 links the generated sequence to the resulting retrieval state. Full reasoning-and-answer supervision yields 57.03 R@5, compared with 53.67 for answer-only and 53.41 for reasoning-only supervision. The combination of intent decomposition and a target-state description is more effective than either part alone in these controls. Using the state predicting e yields 57.03, compared with 52.25 for the state after e and 53.06 for the alternative structured </think> extraction. This supports the pre-marker representation used by both TRACE routes. Appendix I provides the full R@1/5/10 results and extraction definitions.

Query–candidate asymmetry. With query supervision and adaptive inference fixed, candidate-side full traces and answer-only outputs yield 18.90 and 19.78 R@5, versus 57.03 for direct candidates. These separately trained variants change only candidate-side supervision and encoding. The results favor direct candidate indexing; they do not isolate content from positional effects (Appendix J).

Optimization and backbone transfer. Fixing $\lambda_{\text{ret}} = 1$, generation-loss weights of 0, 0.5, 1, and 2 yield CIRR R@5 of 53.06, 55.92, 57.03, and 55.74. The zero-weight experiment disables only the generation-loss contribution. Equal weighting and LoRA rank 128 with scale 256 give the best tested results. TRACE reaches M-BEIR averages of 50.2, 55.6, and 58.8 on Qwen2-VL-2B, Qwen2-VL-7B, and Qwen2.5-VL-7B, respectively (Appendix K).

4.6 Qualitative Visualization

Figure 4 presents queries, generated sequences, and targets. The diver and motorcycle cases illustrate requested visual changes; the text-to-image example illustrates direct matching. For the motorcycle, the trace changes the road to a track and adds the requested wheelie while retaining the subject. Additional ranked examples appear in Appendix M.

5 Conclusion

TRACE learns multimodal retrieval representations through selective query-side reasoning. A first-token decision chooses immediate embedding or intent resolution, and both routes terminate in a shared representation trained with generation and contrastive objectives. M-BEIR-CoT supplies supervision for these complementary behaviors. Broad retrieval evaluations, fixed-policy comparisons, and CIRR analyses support this combination of variable query computation and direct candidate indexing.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, assist in drafting the Method section and its mathematical exposition, and support the development of experimental code. Generative models also provided route supervision, synthetic reasoning traces, and consistency verification for M-BEIR-CoT (Section 3.1 and Appendix D). The research idea, method design, and experimental conception originated with the authors, who conducted the experiments and produced and interpreted the reported results. AI assistance supported the authors' writing and implementation; the reported numerical results were obtained from the authors' experiments. The authors reviewed the AI-assisted text and code and take full responsibility for the final manuscript, technical claims, implementation, and results.

ETHICS STATEMENT

This work studies selective reasoning and representation learning for universal multimodal retrieval using established benchmarks and supervision derived from their training subsets. Our experiments involve no new human-subject study or collection of user data. We recognize that retrieval models and synthetic traces can inherit biases and privacy risks from pretrained models and source data, and that broader access to retrieval systems requires responsible use. Released materials will respect the licenses and usage terms of the underlying models and datasets. We encourage downstream deployments to assess retrieval quality across user groups and input modalities and to apply appropriate access controls when indexing sensitive content.

REPRODUCIBILITY STATEMENT

We will publicly release the TRACE implementation, M-BEIR-CoT, and the configurations for training and evaluation. Section 3 specifies data construction, route selection, representation extraction, and training objectives. Appendix B provides implementation details and hyperparameters; Appendices C–E document corpus composition, trace generation and filtering, and prompt interfaces. Evaluation protocols are detailed in Appendix F, while Appendices G–K provide efficiency analyses, controlled comparisons, representation ablations, and backbone sensitivity results. These materials document the experimental pipeline and support independent reproduction and further development.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Recurrence meets transformers for universal multimodal retrieval. *arXiv preprint arXiv:2509.08897*, 2025.

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: LLM learns when to think. *arXiv preprint arXiv:2505.13379*, 2025.

Chenwei He, Xiangzhao Hao, Tianyu Yang, Yuxiang Ma, Yuheng Jia, Lingxiang Wu, Chaoyang Zhao, Haiyun Guo, and Jinqiao Wang. PLUME: Latent reasoning based universal multimodal embedding. *arXiv preprint arXiv:2604.02073*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Lang Huang, Qiyu Wu, Zhongtao Miao, and Toshihiko Yamasaki. Joint fusion and encoding: Advancing multimodal retrieval from the ground up. *arXiv preprint arXiv:2502.20008*, 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-RL: Reinforcement learning for reasoning-driven multimodal embeddings. *arXiv preprint arXiv:2602.13823*, 2026.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024a.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024b.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring reasoning-driven generative multimodal embeddings. *arXiv preprint arXiv:2511.00405*, 2025.

Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. U-MARVEL: Unveiling key factors for universal multimodal retrieval via embedding learning with MLLMs. *arXiv preprint arXiv:2507.14902*, 2025.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. *arXiv preprint arXiv:2411.02571*, 2024.

Chunxu Liu, Jiyuan Yang, Ruopeng Gao, Yuhan Zhu, Feng Zhu, Rui Zhao, and Limin Wang. Reasoning guided embeddings: Leveraging MLLM reasoning for improved multimodal retrieval. *arXiv preprint arXiv:2511.16150*, 2025a.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. LamRA: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4015–4025, 2025b.

Yibo Lyu, Rui Shao, Gongwei Chen, Yijie Zhu, Weili Guan, and Liqiang Nie. PUMA: Layer-pruned language model for efficient unified multimodal retrieval with modality-adaptive learning. *arXiv preprint arXiv:2507.08064*, 2025.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents, 2025. URL https://arxiv.org/abs/2507.04590.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. EVA-CLIP-18B: Scaling CLIP to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024.

Sagar Vaze, Nicolas Carion, and Ishan Misra. GeneCIS: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6439–6448, 2019.

Yuchi Wang, Dingkang Yang, Haiyang Yu, Weikang Bian, Jiefeng Long, Xiao Liang, Chao Feng, and Hongsheng Li. MMEmb-R1: Reasoning-enhanced multimodal embedding with pair-aware selection and adaptive control. *arXiv preprint arXiv:2604.06156*, 2026.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 2024.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11307–11317, 2021.

Peixi Wu, Ke Mei, Feipeng Ma, Bosong Chai, Zhibin Lan, Chenxi Zhao, Shannan Yan, Jie Chen, Zhangchi Hu, Yansong Peng, Bo Lin, Junjie Zhou, Dacheng Yin, Tianyi Wang, Fengyun Rao, Jing Lyu, Hebei Li, and Xiaoyan Sun. Beyond chain-of-thought: Rewrite as a universal interface for generative multimodal embeddings. *arXiv preprint arXiv:2604.22280*, 2026a.

Yebo Wu, Feng Liu, Ziwei Xie, Zhiyuan Liu, Changwang Zhang, Jun Wang, and Li Li. TSEmbed: Unlocking task scaling in universal multimodal embeddings. *arXiv preprint arXiv:2603.04772*, 2026b.

Ruiran Yan, Zheng Liu, and Defu Lian. O1 Embedder: Let retrievers think before action. *arXiv preprint arXiv:2502.07555*, 2025.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *European Conference on Computer Vision*, 2024a.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. AdaptThink: Reasoning models can learn when to think. *arXiv preprint arXiv:2505.13417*, 2025.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *International Conference on Machine Learning*, 2024b.

Longxiang Zhang, Weilong Dai, Guanghao Zhang, Hao Jiang, and Pipei Huang. Think when needed: Adaptive reasoning-driven multimodal embeddings with a dual-LoRA architecture. *arXiv preprint arXiv:2605.14448*, 2026.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. GME: Improving universal multimodal retrieval by multimodal LLMs. *arXiv preprint arXiv:2412.16855*, 2024c.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, and Shiqi Wang. Retrv-R1: A reasoning-driven MLLM framework for universal and efficient multimodal retrieval. *arXiv preprint arXiv:2510.02745*, 2025.

Table 6: **Guide to the supplementary evidence.**

Question	Supporting section
How is TRACE trained and decoded?	Sec. B
How is M-BEIR-CoT constructed and filtered?	Secs. C–D
What prompts and instructions are used?	Sec. E
What are the evaluation sources and shifts?	Sec. F
How much computation does routing save?	Sec. G
Does reasoning help with the embedder fixed?	Sec. H
Which reasoning and pooling choices matter?	Sec. I
Why is reasoning restricted to queries?	Sec. J
How sensitive is TRACE to the backbone and objectives?	Sec. K
Why do we evaluate UMR rather than broad UME?	Sec. L

A Supplementary Overview

This appendix provides implementation details, dataset statistics, prompt templates, extended evaluations, and controlled analyses that support the main paper. The main paper remains self-contained. Section B describes training and inference. Sections C–F document M-BEIR-CoT and the evaluation protocols. Sections G–K present the extended efficiency, routing, pipeline, representation, candidate-side, and optimization analyses. Section L discusses the distinction between retrieval-native UMR and broad UME training. Sections M and N provide additional qualitative observations and limitations. Quantitative tables accompany the corresponding discussion sections. Table 6 maps the main experimental questions to the corresponding evidence.

B Experimental and Reproducibility Details

Direct-Trained, Always-Reasoning-Trained, and TRACE are independently optimized models rather than inference-time variants of the same checkpoint. Direct-Trained is supervised to emit <|emb|> for every query, whereas TRACE is trained with the mixed direct and reasoning-enhanced outputs in M-BEIR-CoT.

For the fixed-policy comparisons on MSCOCO and CIRR, we separately construct always-reasoning supervision using the same teacher model, task-specific prompting strategy, and filtering pipeline. Always-Reasoning-Trained generates a retrieval-oriented sequence before <|emb|> for every query. This additional always-reasoning corpus is constructed only for MSCOCO and CIRR, not for the complete M-BEIR training mixture.

B.1 Model Architecture and Trainable Parameters

TRACE uses Qwen2.5-VL-7B as its backbone. The vision encoder is frozen, the multimodal projector remains trainable, and the language model is adapted using LoRA with rank $r = 128$ and scale $\alpha = 256$. LoRA is applied to the key and value projection modules with dropout 0.1. The newly introduced <|emb|> token is added to the vocabulary and its embedding is trainable. Query and candidate representations are extracted from the hidden state immediately preceding <|emb|>. Candidate items always use direct encoding and can therefore be embedded and indexed offline.

B.2 Optimization and Training Configuration

We train for one epoch on 16 NVIDIA H20 GPUs with a global batch size of 1024. The per-device batch size is 16, and gradient accumulation uses four steps. Optimization uses AdamW with a peak learning rate of 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01, and gradient clipping at 2.0. The learning rate follows a cosine schedule with warm-up. Training uses bf16 precision, and the complete training process requires approximately 16 hours.

Task-aware sampling balances the heterogeneous M-BEIR training tasks. Let K denote the number of sampling units used by the training mixture. Each unit is sampled uniformly,

$$p_k = \frac{1}{K}, \tag{5}$$

followed by sampling an example from the selected unit. TRACE and the backbone-matched LamRA-Ret reproduction use the same optimization budget on the full M-BEIR training mixture. For the dataset-specific MSCOCO and CIRR controls, Direct-Trained, Always-Reasoning-Trained, and TRACE use matched optimization budgets.

B.3 Synthetic Data Construction

Route assignment and reasoning-trace generation use GPT-4.1, while model-based verification uses GPT-4o. Teacher generation uses temperature 1.0, top-$p = 1.0$, and a maximum of 4096 generated tokens. The construction pipeline makes approximately 1M teacher calls and 0.6M verifier calls.

The positive candidate is visible only to the teacher and verifier during data construction. It is never included in the input context used by TRACE to construct a query embedding. The target provides semantic grounding and supports consistency verification. To reduce target leakage, generated content is rejected if it contains image identifiers, file paths, URLs, the complete target caption, or long verbatim spans copied from the target text.

B.4 Generation and Contrastive Objectives

Generation and retrieval supervision operate at different granularities. Let $Y_i = (y_{i,1}, \ldots, y_{i,T_i})$ denote the supervised query-side output sequence for query Q_i, including the final token $<|\texttt{emb}|>$. For a direct example, Y_i contains only $<|\texttt{emb}|>$. For a reasoning-enhanced example, Y_i contains the intermediate retrieval-oriented tokens followed by $<|\texttt{emb}|>$.

We apply masked causal-language-model cross-entropy to the complete supervised query-side output sequence:

$$\mathcal{L}_{\text{gen}} = \text{CE}_{\text{masked}}\big(\{P_\theta(y_{i,t} \mid Q_i, p^q_{u_i}, y_{i,<t})\}_{i,t}\big). \tag{6}$$

Input-prompt and padding positions are masked and do not contribute to $\mathcal{L}_{\text{gen}}$. Thus, a direct example contributes the cross-entropy term that supervises immediate prediction of $<|\texttt{emb}|>$, whereas a reasoning-enhanced example contributes cross-entropy terms for route selection, retrieval-oriented reasoning, and termination with $<|\texttt{emb}|>$. The loss supervises the same output positions as standard masked causal-language-model training.

Retrieval supervision is applied once per query–candidate pair. Let t^e_i and $t^e_{c_i}$ denote the positions at which the final $<|\texttt{emb}|>$ token is predicted for the query and candidate, respectively. We extract the query and candidate representations from the hidden states immediately preceding this token:

$$\mathbf{e}_{q_i} = \mathbf{h}^{q_i}_{t^e_i - 1}, \qquad \mathbf{e}_{c_i} = \mathbf{h}^{c_i}_{t^e_{c_i} - 1}. \tag{7}$$

The representations are L2-normalized before contrastive learning. Using cosine similarity with fixed temperature $\tau = 0.07$, we optimize

$$\mathcal{L}_{\text{ret}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(\mathbf{e}_{q_i}, \mathbf{e}_{c_i})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\mathbf{e}_{q_i}, \mathbf{e}_{c_j})/\tau\right)}. \tag{8}$$

Candidate embeddings are gathered across data-parallel workers. Each contrastive denominator contains 256 candidates (16 examples per device across 16 devices), including the positive candidate. Four gradient-accumulation steps give an effective optimizer batch size of 1024; candidates are not pooled across those steps. No contrastive loss is applied separately to the intermediate reasoning tokens or their hidden states.

The final training objective is

$$\mathcal{L} = \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{\text{ret}}\mathcal{L}_{\text{ret}}, \tag{9}$$

where $\lambda_{\text{gen}} = \lambda_{\text{ret}} = 1$.

B.5 Input Processing and Autoregressive Decoding

The maximum multimodal context length is 4096 tokens, and the maximum generated reasoning length is 4096 tokens. At inference, TRACE uses greedy decoding with KV caching enabled. Decoding stops when <|emb|> is emitted. If this token is not generated within the maximum output length, the final generated-token state is used as a fallback representation.

Direct-Trained and Always-Reasoning-Trained use the same decoding settings as TRACE. Always-Reasoning-Trained is independently trained to produce an intermediate sequence for every query; it is not obtained by masking <|emb|> in an adaptive TRACE checkpoint. The separate "TRACE, Forced Direct" intervention changes only the first inference token of a trained TRACE checkpoint and is used solely for the matched-model analysis.

B.6 Evaluation and Efficiency Measurement

The primary evaluation follows the M-BEIR local-pool protocol, which contains 16 settings across eight query–candidate configurations and ten datasets. Controlled CIRR experiments use the M-BEIR candidate pool of approximately 21K images. CC-Neg and Sugar-Crepe are included as auxiliary image–text compatibility diagnostics and are not counted as core UMR retrieval tasks. Throughput measurements use one query at a time (batch size 1) on one NVIDIA H20 and report online query encoding, excluding offline candidate indexing. Per-route token counts and a complete timing breakdown are not reported, so aggregate QPS alone cannot identify the source of overhead.

B.7 Software, Randomness, and Model Selection

Experiments use Ubuntu 24.04, Python 3.11, PyTorch 2.3.1, Transformers 4.54.0, and CUDA 12.4. The main-table single-run experiments are reported with random seed 42. The main M-BEIR and distribution-shifted tables report one run. CIRR ablations are means over three runs with distinct random seeds. We report these means without standard deviations and interpret small differences descriptively, rather than as statistically established improvements.

Hyperparameters are selected on the official M-BEIR validation splits. The M-BEIR test sets and the final CIRR reporting pool are evaluated only after fixing the model and hyperparameters. The CIRR component studies characterize LoRA rank, loss weights, and representation-extraction positions under the local-pool protocol; their scores are reported separately from the full-mixture benchmark.

B.8 Availability

We will publicly release the TRACE implementation, M-BEIR-CoT route labels and reasoning traces, data-construction scripts, and training and evaluation configurations. We also plan to release model checkpoints upon publication.

C M-BEIR-CoT Composition

M-BEIR-CoT uses GPT-4.1 to assign routes at the query level within the M-BEIR training splits. For each query, the teacher determines whether its target semantics are sufficiently explicit for direct encoding or whether additional intent resolution would be useful. Queries assigned to the direct route are supervised to emit <|emb|> immediately, whereas queries assigned to the reasoning route receive a target-grounded intermediate sequence before <|emb|>. The route labels serve as teacher-provided supervision, not as measured counterfactual retrieval utility.

The resulting corpus contains 518,311 direct examples and 575,442 filtered reasoning examples. Their distributions reflect differences in query complexity across datasets: datasets dominated by explicit semantic matching contribute more direct examples, while datasets involving interleaved modalities, compositional modification, or knowledge-intensive retrieval contribute more reasoning examples. Aggregate counts alone do not quantify the correlation between route assignment and task identity. Only M-BEIR training examples are used during construction; validation and test examples are never provided to the teacher or used to train TRACE.

Table 7: **Summary of the TRACE training and evaluation configuration.**

Category	Item	Setting
Model	Backbone	Qwen2.5-VL-7B
Model	Vision encoder / projector	Frozen / trainable
Model	LoRA	$r = 128,\ \alpha = 256$; key/value projections; dropout 0.1
Model	Retrieval state	Hidden state immediately preceding <|emb|>
Optimization	Epochs / global batch	1 / 1024
Optimization	Per-device batch / accumulation	16 / 4 steps
Optimization	Hardware / training time	16 NVIDIA H20 GPUs / approximately 16 hours
Optimization	Optimizer / learning rate	AdamW / 2×10^{-4}
Optimization	Betas / weight decay / clipping	$(0.9, 0.95)$ / 0.01 / 2.0
Optimization	Schedule / precision	Cosine decay with warm-up / bf16
Loss	Contrastive candidates	256 candidates per contrastive step; no cross-accumulation candidate cache
Loss	Similarity / temperature	L2-normalized representations; cosine similarity / 0.07
Loss	Generation / retrieval weights	$\lambda_{\text{gen}} = \lambda_{\text{ret}} = 1$
Data construction	Teacher / verifier	GPT-4.1 / GPT-4o
Decoding	Strategy / maximum length	Greedy with KV cache / 4096
Decoding	Termination / fallback	Stop at <|emb|>; otherwise use the final generated-token state
Software	Python / PyTorch / CUDA	3.11 / 2.3.1 / 12.4
Randomness	Main-table seed	42

D Data Generation and Filtering

Figure 2 summarizes the construction pipeline; this section expands its route assignment, generation, and filtering steps.

D.1 Route Assignment

GPT-4.1 assigns route-specific supervision to individual queries during data construction. The assignment is a distillation signal rather than an intrinsic or model-independent definition of difficulty. For direct examples, the target output is only <|emb|>. For reasoning examples, the teacher produces a concise sequence describing the expected target semantics before <|emb|>. The positive target is visible only to the teacher and verifier during data construction. TRACE never accesses the positive target while encoding a test query.

D.2 Task-Specific Trace Generation

For composed retrieval, the generation prompt asks the teacher to identify the source state, the requested change, and the context that should remain unchanged. For visual information-seeking retrieval, the prompt asks for the entities, evidence, or answer characteristics that should appear in a relevant item. The generated sequence is intended to guide retrieval rather than to provide an unconstrained explanation.

D.3 Rule-Based Filtering

The first stage filters generated reasoning traces using the following checks; direct examples retain their single-marker target:

- **Format.** The raw output must contain exactly one <reasoning> block and one <answer> block, in that order. Nested, missing, or malformed tags are rejected.
- **Length.** The reasoning span must contain 16–256 teacher-tokenizer tokens, and the answer span must contain 2–64 tokens.
- **Language and repetition.** Non-English outputs, excessive repeated n-grams, obvious prompt leakage, and meta tokens such as "assistant:" or "system:" are removed.

- **Refusal and boilerplate.** Refusals and generic statements such as "As an AI" or "I cannot determine" are discarded.
- **Semantic sanity.** The answer must contain at least one salient term from the query instruction or target description after stopword removal.
- **Composed-retrieval degeneration.** An answer that normalizes to the source caption is removed because it fails to apply the requested modification.
- **Deduplication.** Normalized answers and query–answer pairs are deduplicated to reduce templated supervision.

D.4 Model-Based Verification

The remaining examples are checked by a strong verifier that returns a consistency decision and confidence in $[0, 1]$. We retain an example only when the decision is positive and confidence exceeds 0.7. For image-target tasks, the verifier receives the reference query, modification text, target image, and generated answer. It checks that the modification is applied and invariant context is preserved. For text-target tasks, it checks entailment against the target title or summary and rejects vague descriptions that could match many candidates.

The `<reasoning>` and `<answer>` tags are used only for parsing during construction. They are removed before TRACE training, leaving the natural-language sequence followed by `<|emb|>`.

D.5 Leakage Controls

M-BEIR-CoT is constructed only from M-BEIR training pairs. Validation and test splits are excluded from route assignment, generation, filtering, and optimization. During M-BEIR evaluation, TRACE follows the benchmark task instructions used by the LamRA protocol. TRACE additionally uses route and output-format instructions for its generation interface; this protocol alignment does not imply identical complete prompts or isolate prompting from the effects of training.

E Prompt Templates and Retrieval Instructions

E.1 Generic Query and Candidate Templates

The following templates illustrate the interface shared across retrieval configurations.

Query-side template.

```
You are a multimodal retrieval encoder for [TASK]. Analyze
the query and produce a retrieval representation.  For a
query whose target semantics are explicit, output only
<|emb|>.  When additional intent resolution is useful, first
produce a concise description of the expected target and
then emit <|emb|>.  Output nothing after <|emb|>.
```

Candidate-side template.

```
You are encoding a [CANDIDATE MODALITY] candidate for [TASK].
Represent its content and distinctive attributes for
matching with queries.  Your output must be exactly <|emb|>.
Do not generate reasoning or any other text.
```

E.2 Representative Task-Specific Guidance

The following shortened templates summarize how the generic interface is specialized. They are included to clarify the intended retrieval behavior; task names and modality placeholders are instantiated by the data pipeline.

Table 8: **Representative M-BEIR evaluation instructions.** Multiple paraphrases are used in the original benchmark; one example per dataset and direction is shown.

Configuration	Dataset	Representative instruction
$q^t \to c^i$	VisualNews	Identify the news image that best matches the described event.
	MSCOCO	Find an everyday image that matches the caption.
	Fashion200K	Retrieve the fashion image that matches the product description.
$q^t \to c^t$	WebQA	Retrieve a Wikipedia passage that provides evidence for the question.
$q^t \to (c^i, c^t)$	EDIS	Find the news image–text item that matches the caption.
	WebQA	Retrieve a Wikipedia image–text item that answers the question.
$q^i \to c^t$	VisualNews	Retrieve the news caption that matches the image.
	MSCOCO	Find the caption that best describes the everyday image.
	Fashion200K	Retrieve the product description for the fashion image.
$q^i \to c^i$	NIGHTS	Find the everyday image most similar to the reference image.
$(q^i, q^t) \to c^t$	OVEN	Retrieve a Wikipedia passage that identifies or explains the visual entity.
	InfoSeek	Retrieve a Wikipedia passage that answers the question about the image.
$(q^i, q^t) \to c^i$	FashionIQ	Find a fashion image that satisfies the modification applied to the reference.
	CIRR	Find an everyday image similar to the reference but with the requested changes.
$(q^i, q^t) \to (c^i, c^t)$	OVEN	Retrieve a Wikipedia image–description pair that identifies the visual entity.
	InfoSeek	Retrieve a Wikipedia image–description pair that provides evidence for the question.

Text-to-image retrieval.

```
Identify the visual concepts, attributes, composition, and
spatial relations stated by the text query.  Map abstract
terms to observable image properties.  For an explicit
description, output <|emb|> directly.  When the description
contains interacting constraints or unresolved visual
intent, summarize the expected target image and then output
<|emb|>.
```

Text-to-text evidence retrieval.

```
Determine the information need, the type of passage that
would answer it, and the entities or relations that the
passage should contain.  For a short factual query, output
<|emb|> directly.  When the information need requires
intermediate resolution, briefly characterize the expected
evidence and then output <|emb|>.
```

Image–text-to-image retrieval.

```
Inspect the reference image, apply the requested change, and
preserve all context that is not modified.  Describe the
expected target state concisely, then output <|emb|>.  If
the instruction is already explicit and requires no
additional resolution, output <|emb|> directly.
```

Image–text-to-text or image–text-pair retrieval.

```
Use the image and question to infer the entity or evidence
that a relevant Wikipedia item should contain.  Describe the
expected evidence only when additional resolution is useful,
then output <|emb|>.
```

E.3 Representative Evaluation Instructions

Table 8 lists one representative test-time instruction for every M-BEIR dataset and retrieval direction. M-BEIR provides several paraphrases per setting; we follow the same instruction protocol used by the benchmark and LamRA. The table also clarifies that VQA-style inputs in InfoSeek and OVEN retrieve indexable evidence items rather than answer labels.

Table 9: **Distribution-shifted evaluation settings.** Some image sources overlap with broad public corpora used by M-BEIR, while the query style, interaction structure, or retrieval dynamics differ.

Dataset	Image source	Configuration	Query format	Candidate format
ShareGPT4V	SA-1B	$q^t \to c^i$	Long text	Image
		$q^i \to c^t$	Image	Long text
Urban-1K	Visual Genome	$q^t \to c^i$	Long text	Image
		$q^i \to c^t$	Image	Long text
Flickr30K	Flickr30K	$q^t \to c^i$	Short text	Image
		$q^i \to c^t$	Image	Short text
CIRCO	MSCOCO unlabeled	$(q^i, q^t) \to c^i$	Reference image + relative caption	Image
GeneCIS	MSCOCO	$(q^i, q^t) \to c^i$	Reference image + relative caption	Image
Visual Dialog	MSCOCO	$q^{\text{dialog}} \to c^i$	Multi-round question–answer history	Image
Visual Storytelling	Flickr	Interleaved $\to c^i$	Text–image story prefix	Image
MT-FIQ	FashionIQ	Interleaved $\to c^i$	Multi-round image + modification history	Image
CC-Neg	CC3M	ITM	Image + text	Match label
Sugar-Crepe	MSCOCO	ITM	Image + text	Match label

F Evaluation-Set Details

Table 9 summarizes the image sources, query formats, and candidate formats used in the 13 distribution-shifted evaluation settings.

These settings introduce distribution shifts but are not uniformly image-disjoint. Urban-1K pairs Visual Genome images with long descriptions (Zhang et al., 2024a); a different dataset name alone does not establish image-level disjointness from pretraining or retrieval-training sources. CIRCO and GeneCIS introduce reference-image-plus-relative-caption queries. Visual Dialog and MT-FIQ alter the interaction structure. CC-Neg and Sugar-Crepe target difficult semantic negatives.

G Extended Efficiency and Routing Analysis

G.1 Aggregate Query Throughput

Table 11(a) reports online query-encoding throughput on the M-BEIR test mixture. The separate MSCOCO policy comparison in the main paper evaluates text-to-image retrieval. Adaptive TRACE is approximately 18% lower in throughput than direct encoding over the mixed workload. Direct routes avoid reasoning-token generation, but the aggregate measurement does not attribute overhead to individual processing stages. This statistic and the task-specific table in the main paper answer different questions: the former reflects the natural benchmark mixture, whereas the latter isolates simple and compositional settings.

G.2 Generated-Token Budget

Direct routes generate no reasoning tokens, while active routes decode a variable-length sequence before the embedding marker. The 4096-token maximum is a decoding limit, not an observed average. Aggregate QPS summarizes the measured workload but does not establish route-conditioned length distributions or tail latency.

G.3 Observed Route Frequencies

Table 11(b) summarizes route frequencies on MSCOCO, CIRR, and a separate 1,000-query cue-analysis subset. The results show correlations between route activation and interpretable query properties. They are not a ground-truth routing-accuracy measure because a unique optimal route is not annotated for every test query.

H Matched-Model and Pipeline Comparisons

Table 11(c) compares end-to-end TRACE with external reasoning-to-embedding pipelines under the same CIRR candidate pool. The third and fourth rows use the TRACE embedder and the same

Table 10: **Target-access ablation on CIRR.** All models are trained from scratch with matched training pairs, architecture, and optimization budget. The two trace-based variants also share route labels. Target-Blind Traces prevents the trace generator and verifier from accessing the positive candidate.

Supervision	R@1	R@5	R@10
Direct Training	24.52	53.06	63.10
Target-Blind Traces	<u>27.91</u>	<u>56.98</u>	<u>67.79</u>
Target-Grounded Traces	**28.37**	**57.03**	**68.30**

reasoning supervision. This comparison evaluates the two implemented pipelines; differences in trace generation and conditioning may also contribute to the gap. The end-to-end model improves R@5 by 2.03 points over the matched external-handoff pipeline.

H.1 Inference Intervention with Fixed Parameters

We additionally force the trained TRACE model to predict <|emb|> as its first output on CIRR. This intervention changes only inference behavior, not model parameters or the retrieval head. R@5 drops from 57.03 to 53.40, yielding a 3.63-point gain from allowing reasoning relative to this forced-direct intervention. It does not compare adaptive routing with an always-reasoning policy on the same checkpoint.

H.2 Target-Blind Trace Control

To examine whether the CIRR improvement depends on access to the positive target during synthetic trace construction, we train a matched target-blind variant from scratch. This variant uses the same training pairs, route labels, model architecture, initialization, optimization schedule, and number of updates as the target-grounded model. During trace generation and verification, the positive candidate is hidden from the teacher and verifier; they receive only the reference image, modification instruction, and task instruction. The positive candidate remains available as the positive item in the standard InfoNCE retrieval objective.

Table 10 reports the results. Target-Blind Traces reaches 27.91 R@1, 56.98 R@5, and 67.79 R@10, improving Direct Training by 3.39, 3.92, and 4.69 points, respectively. Its remaining gaps relative to Target-Grounded Traces are only 0.46, 0.05, and 0.51 points. In particular, the target-blind model retains 98.7% of the target-grounded R@5 improvement over Direct Training. These results indicate that access to positive-target content during trace generation and verification is not necessary for most of the observed CIRR gain. The 0.05-point R@5 gap should not be interpreted as statistically significant without uncertainty estimates.

I Reasoning and Representation Ablations

I.1 Reasoning Supervision Components

Table 12(a) separates the effects of target rewriting, intermediate decomposition, and the full supervised sequence. Answer-only supervision approximates a target rewrite, while reasoning-only supervision removes the explicit target conclusion. Neither component alone approaches the full sequence. The result supports combining intent decomposition with target-state summarization.

I.2 Representation Extraction Position

Table 12(b) compares three extraction positions. Each variant is trained end to end with its designated representation state. The pre-token state is responsible for predicting <|emb|> under causal attention and has access to the query and all preceding reasoning tokens. The state after consuming <|emb|> includes an additional marker token and occupies a different causal position. Since the supervised output ends at the marker, this variant does not receive a supervised next-token target beyond it. The alternative structured baseline extracts the closing </think> tag state. The pre-marker and final

Table 11: **Efficiency, routing, and pipeline analyses.** (a) Aggregate online query throughput on the M-BEIR test mixture. (b) Observed route frequencies. (c) End-to-end and two-stage pipelines on CIRR under the M-BEIR candidate pool.

(a) Aggregate throughput

Policy	QPS	Rel. cost
Direct encoding	13.9	1.00×
TRACE adaptive	11.4	1.22×

(b) Observed routing behavior

Query subset or cue	Activation
MSCOCO queries choosing direct	96%
CIRR queries choosing reasoning	62%
Queries containing "not" or "without" choosing reasoning	93%
Scenes with more than three detected entities choosing reasoning	74%
Short single-object descriptions or short factual queries choosing direct	96%

(c) Pipeline comparison

Pipeline	R@5
LamRA + zero-shot Qwen2.5-VL CoT	50.10
LamRA + SFT Qwen2.5-VL CoT	50.40
TRACE embedder + external SFT CoT	55.00
TRACE end-to-end adaptive	**57.03**

reasoning-token positions otherwise coincide when no tag intervenes, and therefore are the same pooling position in that sequence. These comparisons concern the tested training and formatting choices, not a universal optimality claim.

J Candidate-Side Asymmetry

Table 12(c) varies candidate-side supervision and encoding while retaining the query-side TRACE setup in all three rows. Query supervision includes the full reasoning-and-answer sequence for reasoning-route examples and immediate <|emb|> for direct examples; test queries still choose their route adaptively. In this block, full-sequence candidate encoding means generating complete candidate-side reasoning and answer text; it does not force full query generation. The direct-candidate row reports adaptive TRACE (57.03 R@5). Always-Reasoning-Trained in Table 3 is a different model trained with reasoning sequences for every query (54.73 R@5). Candidate-side generation lowers retrieval recall in this setting. The main interpretation is functional asymmetry: a query may contain unresolved intent and can benefit from target-state construction, whereas a candidate is already the item to be indexed and matched, so generation can shift it away from a stable content anchor.

J.1 A Mechanistic Hypothesis

Candidate-side generation changes both the content available to the final representation and the positions of the tokens it attends to. Under RoPE, inserting a variable-length sequence changes the relative distances between image tokens and the pre-token extraction state. Generated descriptions can also alter which attributes of an item are emphasized. These changes offer possible explanations for the observed recall reduction, but the experiment does not isolate them. Because the variants are trained end to end with their respective candidate outputs, the scores alone do not establish that a positional mismatch causes the decline or that such configurations are absent from training.

Table 12: **Reasoning, representation, and candidate-side analyses on CIRR.** (a) Reasoning-supervision components. (b) Representation-extraction position. (c) Candidate-side generation with query supervision and adaptive inference fixed.

(a) Reasoning supervision

Supervision	R@1	R@5	R@10
Direct training	24.52	53.06	63.10
Answer only	25.99	53.67	64.48
Reasoning only	26.38	53.41	64.60
Full sequence	**28.37**	**57.03**	**68.30**

(b) Extraction position

State	R@1	R@5	R@10
<|emb|> state	24.46	52.25	63.81
Structure </think>	24.36	53.06	63.93
Pre-token state	**28.37**	**57.03**	**68.30**

(c) Candidate-side generation

Query inference	Candidate	R@1	R@5	R@10
Adaptive	Full sequence	7.75	18.90	26.50
Adaptive	Answer only	8.61	19.78	26.88
Adaptive	Direct	**28.37**	**57.03**	**68.30**

K Optimization and Backbone Sensitivity

K.1 LoRA Hyperparameters

Table 13(a) reports sensitivity to LoRA rank and scale. The rank trend is not monotonic: R@5 is 56.31 at rank 16 and 56.19 at rank 32. Rank 128 with scale 256 provides the best measured result without the additional parameter cost of rank 256.

K.2 Generation-Loss Weight

Table 13(b) varies the generation-loss weight while fixing the retrieval-loss weight to one. The zero-weight variant belongs to this loss-weight experiment: it disables the generation-loss contribution while retaining the retrieval objective. Raising the weight from zero to one improves R@5 from 53.06 to 57.03, while weights of 0.5 and 2 give 55.92 and 55.74. Equal weighting is the best tested setting; the experiment alone does not identify the optimization mechanism behind the differences.

K.3 Backbone Scaling

Table 13(c) evaluates TRACE on three backbones from the Qwen-VL family. The trend shows that TRACE benefits from stronger multimodal backbones. It does not by itself establish that the gain is caused only by reasoning ability because the backbones also differ in perception, instruction following, and pretraining.

L Why We Focus on UMR Rather Than Broad UME

MMEB-style UME casts classification, visual question answering, visual grounding, and retrieval into a shared ranking interface. This is useful for assessing broad embedding functionality, but the four meta-tasks differ in target semantics, candidate construction, and representation granularity. The resulting cross-task optimization problem is distinct from the cross-modal retrieval problem studied by TRACE.

Our experiments do not establish performance on the full MMEB task mixture. TSEmbed (Wu et al., 2026b) studies task conflict in that broader setting. Holding the objective family fixed here makes

Table 13: **Optimization and backbone sensitivity.** (a) LoRA rank and scale. (b) Generation-loss weight with $\lambda_{\text{ret}} = 1$. (c) Backbone scaling.

(a) LoRA hyperparameters

Rank	Scale	R@1	R@5	R@10
16	32	26.67	56.31	66.45
32	64	26.90	56.19	67.29
128	256	**28.37**	**57.03**	**68.30**
128	512	27.48	56.74	67.84
256	512	28.19	56.81	68.02

(b) Generation-loss weight

λ_{gen}	R@1	R@5	R@10
0.0	24.52	53.06	63.10
0.5	26.89	55.92	66.80
1.0	**28.37**	**57.03**	**68.30**
2.0	27.01	55.74	66.91

(c) Backbone scaling

Backbone	M-BEIR Avg.	CIRR R@5
Qwen2-VL-2B	50.2	51.9
Qwen2-VL-7B	55.6	54.2
Qwen2.5-VL-7B	**58.8**	**57.3**

the scope explicit, but does not eliminate the need to compare adaptive methods under a matched retrieval protocol.

M Additional Qualitative Examples and Error Analysis

Figure 5 presents seven ranked-retrieval examples from the original visualization. It shows immediate embedding for direct queries and intermediate text for reasoning queries, together with the displayed candidate rankings. The highlighted candidates and abbreviated text are preserved from the original figure. The paper's quantitative comparisons do not depend on these individual examples.

Generated text may fail by inventing entities, substituting relations, losing a negation, or adding unnecessary attributes. These are distinct from ranking errors caused by an otherwise faithful representation. The aggregate recall and route-frequency results do not provide a measured breakdown of these error types. A route-conditioned human audit should examine the original query, generated sequence, both routes' rankings, and relevance annotations together; isolated examples cannot establish the faithfulness of the model's reasoning.

N Limitations and Scope

Granularity of adaptation. TRACE conditions its routing decision on both the retrieval instruction and the query content. This design is intentional in UMR, where the task instruction specifies the retrieval relation and is part of the model input. The matched, from-scratch MSCOCO and CIRR experiments show that direct and reasoning-enhanced inference are preferable in different retrieval regimes under the same backbone, training pairs, candidate pools, optimization schedule, and number of updates, while the cue-based analysis further shows that routing behavior varies with query properties such as negation, compositional constraints, and entity density. However, because the benchmarks do not provide oracle route annotations, the observed routing decisions should not be interpreted as a globally optimal policy for every individual query. The current analysis does not measure both routes for every query and therefore cannot quantify the gap to an oracle choice.



Figure 5: **Additional ranked retrieval examples.** Queries and generated text appear on the left, with ranked candidates on the right. Red highlights reproduce the original annotations.

Autoregressive query-side cost. TRACE preserves the standard dense-retrieval interface because candidate representations remain fixed and can be computed and indexed offline. On the aggregate M-BEIR workload, adaptive routing incurs approximately $1.22\times$ the query-side cost of direct encoding, while providing about $1.9\times$ the throughput of always-on reasoning on the representative direct and compositional tasks studied in the main paper. Nevertheless, queries that activate the reasoning route still require sequential autoregressive decoding. The measured throughput gains are relative to always-on reasoning, while direct encoding remains faster on the representative tasks.

Teacher-grounded offline supervision. The positive candidate is used by the teacher and verifier during the main offline trace-construction pipeline, but it is not included in the query-side input to TRACE. Construction is restricted to the M-BEIR training splits, and the filtering pipeline removes explicit identifiers, URLs, complete target captions, and long verbatim spans. We additionally conduct a matched target-blind control on CIRR that preserves the training pairs, route labels, architecture, and optimization budget while hiding the positive candidate from the trace generator and verifier. The target-blind model reaches 56.98 R@5 and retains 98.7% of the improvement achieved by the target-grounded model over Direct Training, substantially reducing the concern that the main CIRR gain is explained by target-specific trace content. However, the target-blind comparison is currently limited to CIRR and has not been repeated over the complete M-BEIR training mixture or other retrieval configurations. The evidence therefore supports the tested CIRR setting and does not establish the same conclusion for the full training mixture.

O ORIGINAL PROMPT ILLUSTRATIONS

The following PDF figures reproduce task-specific query and candidate prompt examples from the original experimental materials. They supplement the shortened interface descriptions in Appendix E; they do not replace a complete executable prompt configuration for every setting.

Prompt for task 1 (text → image q^t)

You are an advanced multimodal retrieval encoder specialized for text-to-image retrieval. Your goal is to analyze the input and generate a high-quality embedding for retrieval.

--- Task-Specific Strategy ---
You are processing a TEXT QUERY to retrieve matching IMAGES. Your reasoning should focus on:
1. **Visual Concepts:** Extract all visual elements described in the text (objects, scenes, colors, compositions, spatial relationships).
2. **Visual Attributes:** Identify key visual attributes that distinguish the target image (style, lighting, perspective, mood).
3. **Semantic Mapping:** Map abstract concepts in text to concrete visual features.

For simple queries (e.g., 'a red ball'), output the embedding directly.
For complex queries requiring visual reasoning, first output a concise visual description of the target image, then emit <|emb|>.

--- Output Rules --- -
- For simple queries: Output only <|emb|>
- For complex queries: Output [reasoning summary] <|emb|>
- After <|emb|>, output nothing else.

Prompt for task 1 (text → image c^i)

You are a multimodal retrieval encoder specialized for encoding CANDIDATE items.Your goal is to fully represent this candidate for matching with queries.

Candidate Modality: IMAGE

--- Task-Specific Strategy ---

You are encoding an IMAGE CANDIDATE for text-to-image retrieval.
Your embedding must capture:
1. **Visual Content:** All objects, scenes, and elements in the image.
2. **Visual Attributes:** Colors, textures, lighting, composition, style.
3. **Semantic Meaning:** The high-level semantic content and relationships.
4. **Distinctive Features:** Unique characteristics that distinguish this image.

--- Output Rules ---
- Your output MUST and ONLY be: <|emb|>
- No reasoning, no text, no whitespace, no punctuation.
- The <|emb|> must be the complete crystallization of all candidate features.

Prompt for task 2 (text → text q^t)

You are an advanced multimodal retrieval encoder specialized for text-to-text retrieval.Your goal is to analyze the input and generate a high-quality embedding for retrieval.

--- Task-Specific Strategy ---
You are processing a TEXT QUERY to retrieve relevant TEXT PASSAGES.Your reasoning should focus on:
1. **Semantic Understanding:** Deeply understand the question's intent and required information.
2. **Answer Characteristics:** Infer what type of text would answer this query (definition, explanation, fact, narrative).
3. **Key Concepts:** Identify critical keywords and semantic relationships that the answer must contain.

For simple factual queries, output the embedding directly.
For complex queries, first output a brief description of the expected answer's content, then emit <|emb|>.

--- Output Rules ---
- For simple queries: Output only <|emb|>
- For complex queries: Output [reasoning summary]<|emb|>
- After <|emb|>, output nothing else.

Prompt for task 2 (text → text c^t)

You are a multimodal retrieval encoder specialized for encoding CANDIDATE items.Your goal is to fully represent this candidate for matching with queries.

Candidate Modality: TEXT

--- Task-Specific Strategy ---
You are encoding a TEXT CANDIDATE for text-to-text retrieval.
Your embedding must capture:
1. **Semantic Content:** The core meaning and information in this text.
2. **Key Concepts:** Important entities, facts, and concepts mentioned.
3. **Text Type:** Whether it's a definition, explanation, narrative, or fact.
4. **Information Density:** What questions this text can answer.

--- Output Rules ----
- Your output MUST and ONLY be: <|emb|>
- No reasoning, no text, no whitespace, no punctuation.
- The <|emb|> must be the complete crystallization of all candidate features.

Prompt for task 3 (text → image,text q^t)

You are an advanced multimodal retrieval encoder specialized for text-to-image-text pair retrieval.Your goal is to analyze the input and generate a high-quality embedding for retrieval.

--- Task-Specific Strategy ---
You are processing a TEXT QUERY to retrieve IMAGE-TEXT PAIRS.Your reasoning should focus on:
1. **Multimodal Understanding:** The query seeks both visual content and textual information.
2. **Visual Elements:** What images would be relevant to this query.
3. **Textual Content:** What accompanying text (captions, descriptions) should match.
4. **Coherence:** The image-text pair must be semantically aligned.

For simple queries, output the embedding directly.
For complex queries, first describe the characteristics of the target image-text pair, then emit <|emb|>.

--- Output Rules ---
- For simple queries: Output only <|emb|>
- For complex queries: Output [reasoning summary]<|emb|>
- After <|emb|>, output nothing else.

Prompt for task 3 (text → image,text c^i,c^t)

You are a multimodal retrieval encoder specialized for encoding CANDIDATE items.Your goal is to fully represent this candidate for matching with queries.

Candidate Modality: IMAGE-TEXT PAIR

--- Task-Specific Strategy ---
You are encoding an IMAGE-TEXT PAIR CANDIDATE.Your embedding must capture:
1. **Visual Content:** Complete visual information from the image.
2. **Textual Content:** Semantic information from the accompanying text.
3. **Multimodal Coherence:** How image and text relate and complement each other.
4. **Joint Semantics:** The combined meaning conveyed by both modalities.

--- Output Rules ---
- Your output MUST and ONLY be: <|emb|>
- No reasoning, no text, no whitespace, no punctuation.
- The <|emb|> must be the complete crystallization of all candidate features.

Figure 6: Original query (left) and candidate (right) prompt examples. From top to bottom: text-to-image, text-to-text, and text-to-image–text-pair retrieval.