LANGUAGE-CONDITIONED MEAN-FIELD UNFOLDING: A VARIATIONAL INFERENCE PARADIGM FOR REFERRING EXPRESSION SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep algorithm unfolding bridges iterative optimization and differentiable neural architectures, yet its formulation has largely centered on recovering latent coefficients under predefined structural priors. In this work, we extend this paradigm to language-conditioned semantic inference, where the object to be unfolded is not an image-reconstruction coefficient, but a semantic posterior over a query-conditioned probability simplex. We propose **Language-Conditioned Unfolding of Mean-Field Inference Networks (LUMIN)**, which formulates referring expression segmentation (RES) as variational mean-field inference in a text-conditioned energy model. LUMIN dynamically constructs a semantic state space from textual hypotheses together with an adaptive background state. A text-derived Gram compatibility kernel is coupled with a symmetric visual affinity graph to propagate language-conditioned evidence through simplex-constrained Softmax updates. Unrolling this inference process yields a compact and interpretable network in which each stage corresponds to a principled surrogate step on the variational free energy. A background-relative differential readout further calibrates marginal log-odds by suppressing encoder-induced common-mode drift. Experiments on RefCOCO/+/g demonstrate competitive segmentation performance under a zero-shot transfer setting, and transparent step-wise posterior refinement with only three unfolding stages.

1 INTRODUCTION

Referring expression segmentation (RES) requires a model to produce a pixel-level mask for the object described by a natural-language expression (Hu et al., 2016; Yu et al., 2016). Unlike closed-vocabulary semantic segmentation, RES operates on a dynamic prediction target: the semantic identity of the mask is determined on the fly by the input expression rather than by a predefined label ontology. Recent approaches predominantly improve cross-modal fusion via recurrent interactions (Li et al., 2018a), cross-modal attention (Ye et al., 2019), vision-language Transformers (Ding et al., 2021; Yang et al., 2022), and CLIP-driven alignment (Radford et al., 2021; Wang et al., 2022). Despite their varied architectures, these methods typically adhere to a feed-forward heuristic: visual and textual representations are fused, and a decoder predicts the target mask in an unconstrained, feed-forward pass without an explicit inference objective.

This feed-forward paradigm leaves an essential inference process implicit. Given a visual scene and multiple textual queries, the model must assign each visual location to the background or a query-conditioned semantic state while propagating contextual evidence without losing spatial boundaries. These interactions naturally suggest viewing RES as a structured posterior inference problem, rather than unconstrained feature fusion.

We propose **Language-Conditioned Unfolding of Mean-Field Inference Networks (LUMIN)**, which reformulates RES as variational mean-field inference within a text-conditioned energy model. As conceptualized in Figure 1, LUMIN constructs a query-conditioned semantic state space together with an adaptive background, and explicitly models interactions among visual locations and language-derived semantic states. The resulting inference process is unfolded into a compact three-stage network, where each stage iteratively refines the semantic posterior according to visual and

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.

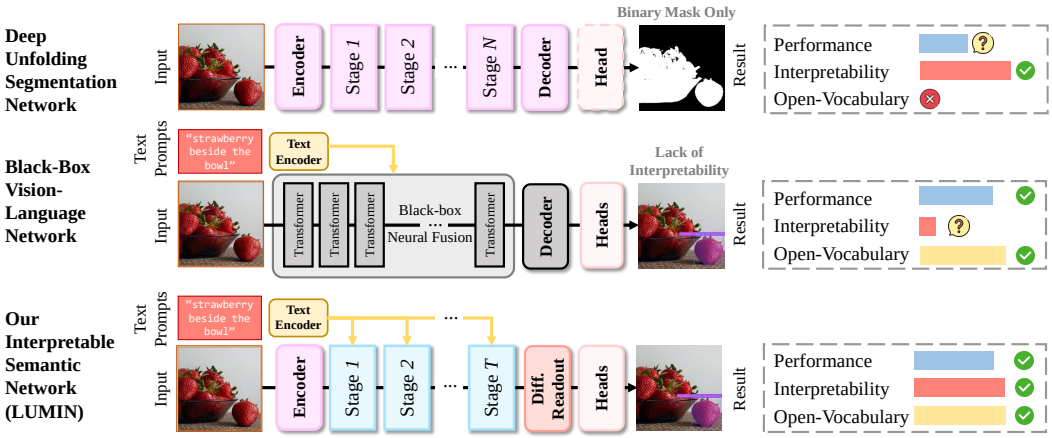


Figure 1: **Comparison of segmentation paradigms.** While classical deep unfolding networks lack language-driven multimodal conditioning (top), modern vision-language models rely on unconstrained black-box Transformer decoders without optimization semantics (middle). LUMIN (bottom) bridges this gap by unfolding language-conditioned variational mean-field inference, achieving both rigorous theoretical interpretability and competitive multimodal performance.

language-conditioned evidence. A background-relative differential readout then converts the refined posterior into the final spatial prediction.

LUMIN establishes a conceptual bridge between vision-language alignment, conditional random fields (CRFs), and deep algorithm unfolding. While pioneering works such as CRF-RNN (Zheng et al., 2015) demonstrate the feasibility of recurrent mean-field inference, they typically operate over a fixed label space. In parallel, classical deep unfolding frameworks (Gregor & LeCun, 2010; Monga et al., 2021) have predominantly focused on recovering latent coefficients under predefined structural priors or dictionaries. **LUMIN shifts this perspective from recovering task-specific latent coefficients to iteratively inferring language-conditioned semantic posteriors over a query-conditioned probability simplex.** This formulation provides an interpretable, optimization-based inference mechanism in which visual interactions are explicitly modulated by language semantics, while the iterative inference process is transformed into a compact learnable architecture.

The contributions of this work are as follows.

- **A Language-Conditioned Unfolding Paradigm:** We formulate dense semantic inference as the estimation of language-conditioned semantic posteriors over a query-conditioned simplex, and unfold the resulting mean-field inference process to a learnable multi-stage architecture.
- **Principled Variational Inference:** We construct a text-derived Gram compatibility kernel coupled with a symmetric visual affinity graph, and formulate their interaction through a continuous variational relaxation. We further show that each unfolding stage corresponds to a convex majorization step of the resulting variational free energy.
- **Parameter-Efficient Posterior Refinement:** We instantiate the proposed paradigm as a compact three-stage architecture for referring expression segmentation, where a small set of stage-wise parameters controls iterative posterior refinement. Experiments on RefCOCO/+/g demonstrate its effectiveness and zero-shot transfer capability.

2 Related Work

2.1 Referring Expression Segmentation

Referring expression segmentation (RES) extends natural-language grounding from bounding boxes to pixel-level masks (Hu et al., 2016; Yu et al., 2016). Early approaches rely on recurrent architectures and convolutional backbones for cross-modal feature fusion (Liu et al., 2017; Li et al., 2018a). Subsequent paradigms exploit cross-modal attention (Ye et al., 2019; Hu et al., 2020),

multi-scale Transformer interactions (Ding et al., 2021; Yang et al., 2022), and aligned representations from vision-language models like CLIP (Radford et al., 2021; Wang et al., 2022). Recent trends have further explored polygon-based generation (Liu et al., 2023c), generalized multi-target formulations (Liu et al., 2023a), foundation model / SAM adaptation (Tang et al., 2025), prototype binding (Li et al., 2025), and zero-shot referring segmentation (Yu et al., 2023).

Despite their architectural diversity, existing RES methods largely treat mask generation as a feed-forward decoding process without an explicit inference objective. LUMIN instead separates representation from inference, using frozen vision-language embeddings as initial evidence and performing iterative posterior refinement through an explicit energy model. The resulting inference process is realized as a compact and modular unfolding architecture.

2.2 Mean-Field CRFs and Differentiable Structured Inference

Conditional random fields (CRFs) provide an established probabilistic framework for dense structured prediction (Lafferty et al., 2001). Dense Gaussian CRFs efficiently refine spatial object boundaries by solving mean-field approximations (Krähenbühl & Koltun, 2011). Seminal works such as CRF-RNN (Zheng et al., 2015) and Gaussian CRF Networks (Vemulapalli et al., 2016) demonstrate that iterative mean-field inference can be unrolled into recurrent or differentiable network layers and trained end-to-end (Arnab et al., 2016).

However, these formulations typically operate over a fixed label space with a predefined label-compatibility structure. This setting differs from RES, where the semantic state space and its interactions should be conditioned on the input language. LUMIN addresses this gap by constructing a query-conditioned semantic state space and deriving language-dependent interactions directly from the input text, enabling mean-field inference to adapt to arbitrary textual queries.

2.3 Deep Algorithm Unrolling

Algorithm unrolling bridges classical optimization algorithms and deep neural networks by mapping finite iterations into interpretable, feed-forward architectures (Monga et al., 2021). While initially established for sparse coding via LISTA (Gregor & LeCun, 2010), unrolling has achieved success in both low-level vision tasks (e.g., compressive sensing (Zhang & Ghanem, 2018; Yang et al., 2018; Zhang et al., 2020; Shen & Gan, 2025; Wang & Gan, 2026)) and high-level vision tasks (e.g., segmentation (Van Luong et al., 2021; Wu et al., 2024; Zhang et al., 2025; Wu et al., 2026)).

Nevertheless, conventional unrolling methods optimize continuous variables in Euclidean space under ℓ_1 or nuclear-norm penalties, implementing element-wise soft-thresholding or shrinkage operators (Zhang & Ghanem, 2018; Wu et al., 2024). Such formulations cannot model categorical probability distributions or multi-hypothesis competition on a probability simplex. **LUMIN fundamentally shifts the unrolling object from sparse visual coefficients under fixed dictionaries to language-conditioned semantic posteriors on a dynamic probability simplex.** The resulting layer updates transition from Euclidean proximal shrinkage to entropy-regularized simplex projections, backed by stage-wise convex majorization guarantees.

3 Methodology

3.1 Problem Formulation and Query-Induced Semantic State Space

Given an input image I and an arbitrary set of M open-vocabulary textual queries $\mathcal{T} = \{T_1, \ldots, T_M\}$ (subsuming canonical single-referent RES when $M = 1$), the objective is to predict binary masks $\{Y_m\}_{m=1}^{M} \in \{0,1\}^{M\times H\times W}$ for the targeted concepts. A vision backbone maps I into an ℓ_2-normalized feature grid $E = [e_1, \ldots, e_N]^\top \in \mathbb{R}^{N\times d}$ at spatial resolution $N = H'W'$. Concurrently, a text encoder embeds each query T_m into a d-dimensional vector, forming the query representation $T = [t_1, \ldots, t_M]^\top \in \mathbb{R}^{M\times d}$, where M strictly indexes distinct semantic queries rather than intra-query tokens.

Unlike classical dense CRFs that operate over a fixed, closed-world label ontology, LUMIN derives a query-conditioned discrete state space directly from the given textual queries. Specifically, we

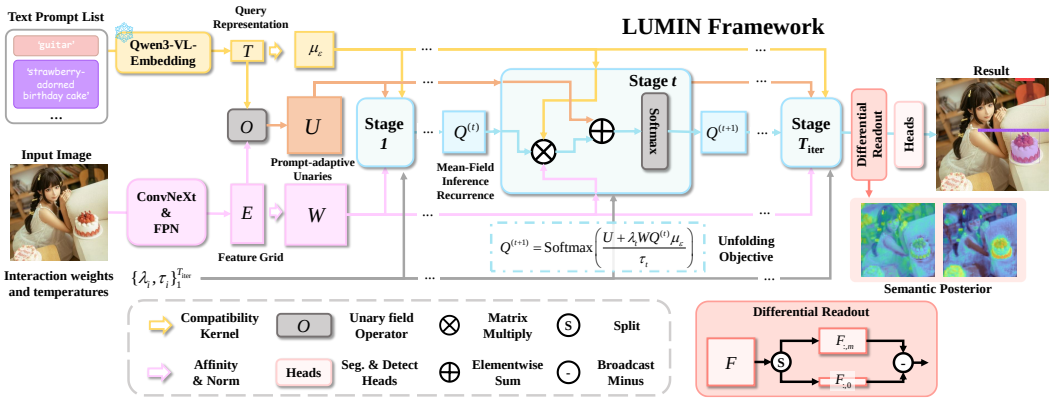


Figure 2: **Overview of the LUMIN framework.** Given an input image and a referring expression with constituent phrases, frozen multimodal encoders extract visual features E and text embeddings T. LUMIN constructs a continuous variational relaxation composed of prompt-adaptive unaries U, a PSD-completed visual affinity graph W, and a language Gram compatibility kernel μ_ϵ. This formulation is unfolded into a parameter-light multi-stage Mean-Field inference recurrence. Finally, a background-referenced differential energy field is extracted and fed into the task head to synthesize the refined target mask without probability saturation. The specific calculation procedure is described in Algorithm 1 of Appendix D.

introduce one contextual background state and one semantic state for each textual query:

$$z_i \in \{0, 1, \ldots, M\}, \qquad K = M + 1, \tag{1}$$

where $i \in \{1, \ldots, N\}$ indexes the visual nodes, state 0 denotes the adaptive contextual background, and each semantic state $m \in \{1, \ldots, M\}$ corresponds one-to-one to the m-th textual query embedding t_m.

To infer the latent assignment $Z = \{z_i\}_{i=1}^N$, we introduce a fully factorized variational posterior

$$\mathcal{Q}(Z) = \prod_{i=1}^{N} Q_i(z_i), \tag{2}$$

where each q_i is a categorical distribution over the $K = M + 1$ semantic states. We collect these marginal probabilities into

$$\mathbf{Q} \in \mathbb{R}^{N \times K}, \qquad Q_{ik} = q_i(z_i = k), \qquad s.t. \sum_{k=0}^{M} Q_{ik} = 1, \qquad Q_{ik} \geq 0. \tag{3}$$

3.2 Text-Conditioned Energy and Semantic Compatibility

We formulate dense multimodal prediction as minimizing an explicit continuous energy functional that couples text-conditioned unary observations with spatial and semantic pairwise interactions.

Text-Conditioned Unary Field. The foreground unary energies are directly provided by the cross-modal affinity $A = ET^\top \in \mathbb{R}^{N \times M}$ without extra projection. Because the background lacks an explicit textual prompt, we synthesize its complementary unary energy conditioned on local visual cues and the peak foreground competition:

$$U_{i0} = g_\theta([e_i; \max_{1 \leq m \leq M} A_{im}]), \quad U = [U_{\cdot 0}, A] \in \mathbb{R}^{N \times K}, \tag{4}$$

where $g_\theta : \mathbb{R}^{d+1} \to \mathbb{R}$ is a lightweight MLP, yielding the prior distribution $\mathbf{Q}^{(0)} = \mathrm{Softmax}(U)$ over the K competing states. where g_θ consists of a two-layer MLP. The unary contribution to the variational energy is expressed as $-\mathrm{Tr}(\mathbf{Q}^\top U)$.

Language-Induced Compatibility Kernel. Standard CRF compatibility matrices assume fixed inter-class transition penalties. In contrast, LUMIN derives semantic compatibility directly from the Gram structure of the normalized query embeddings T:

$$\mu(T) = \begin{bmatrix} \mu_{\text{bg}} & \mathbf{0}^\top \\ \mathbf{0} & TT^\top \end{bmatrix} \in \mathbb{R}^{K\times K}, \qquad \mu_\epsilon = \mu(T) + \epsilon I_K, \quad \epsilon > 0, \tag{5}$$

where $\mu_{\text{bg}} \geq 0$ is a learnable background self-compatibility scalar, $\mu_{\text{text}} = TT^\top \succeq 0$ is positive semidefinite (PSD) by construction, and ϵI_K guarantees strict positive definiteness $\mu_\epsilon \succ 0$. Crucially, the Gram structure is not introduced merely as a semantic heuristic; its PSD property couples linguistic geometry with optimization requirements, enabling the tight majorization construction used by the subsequent unfolding.

Symmetric Visual Affinity Graph. Spatial coherence is governed by normalized projected visual features $\tilde{e}_i = Pe_i/\|Pe_i\|_2$. Denoting the cosine similarity as $S_{ij} = \langle \tilde{e}_i, \tilde{e}_j \rangle$, we define the raw visual affinity using an exponential kernel and apply symmetric degree normalization:

$$W_{ij}^{\text{raw}} = \exp\left(\frac{S_{ij}}{\sigma^2}\right), \quad W = D^{-1/2} W^{\text{raw}} D^{-1/2}, \tag{6}$$

where $D_{ii} = \sum_j W_{ij}^{\text{raw}}$ and $\sigma > 0$ is a learnable bandwidth parameterized via Softplus. By the Schur product theorem, expanding the exponential as a power series of the Gram matrix $S \succeq 0$ confirms that $W \succeq 0$.

Diagonal PSD Completion and Continuous Relaxation. In discrete graphical models, the pairwise potential relies on a hollow affinity matrix $W_{\text{off}} = W - \text{diag}(W)$. However, $\text{Tr}(W_{\text{off}}) = 0$ inherently implies negative eigenvalues ($\lambda_{\min} < 0$), violating the positive semidefiniteness required for convex majorization. To construct a majorizable continuous geometry, LUMIN introduces a *diagonal PSD completion* by retaining non-negative self-affinities: $W = W_{\text{off}} + \text{diag}(W) \succeq 0$. This yields the continuous pairwise energy $-\frac{\lambda}{2}\text{Tr}(\mathbf{Q}^\top W \mathbf{Q} \mu_\epsilon)$, which expands into the Kronecker form:

$$\text{Tr}(\mathbf{Q}^\top W \mathbf{Q} \mu_\epsilon) = \text{vec}_r(\mathbf{Q})^\top (W \otimes \mu_\epsilon)\, \text{vec}_r(\mathbf{Q}), \tag{7}$$

where $B = W \otimes \mu_\epsilon \succeq 0$ since the Kronecker product of two PSD matrices preserves positive semidefiniteness. As proven in Appendix A, this surrogate maintains vertex consistency with the discrete energy while guaranteeing monotonic descent under Majorization-Minimization.

3.3 Continuous Variational Relaxation on the Simplex

Coupling the text-conditioned unary observations with the PSD-completed pairwise interaction and entropic regularization, we formulate inference as minimizing the continuous variational free energy $\mathcal{F}(Q)$ over the product probability simplex Δ^K:

$$\mathcal{F}(\mathbf{Q}) = -\text{Tr}(\mathbf{Q}^\top U) - \frac{\lambda}{2}\text{Tr}(\mathbf{Q}^\top W \mathbf{Q} \mu_\epsilon) + \tau \sum_{i=1}^{N}\sum_{k=0}^{M} Q_{ik} \log Q_{ik}, \tag{8}$$

where $\tau > 0$ denotes the variational temperature controlling entropic dispersion. Rather than viewing Eq. 8 as the literal expectation of an unregularized discrete CRF, we define it as a continuous variational relaxation whose stationarity admits exact, closed-form Majorization-Minimization updates.

Setting the first-order Lagrangian variation to zero under the simplex constraints $\sum_k Q_{ik} = 1$ analytically yields the mean-field fixed-point recursion (derivation in Appendix A):

$$\mathbf{Q}^* = \text{Softmax}\left(\frac{U + \lambda W \mathbf{Q}^* \mu_\epsilon}{\tau}\right), \tag{9}$$

where the Softmax operator acts independently across the label dimension for each visual node.

3.4 Parameter-Efficient Mean-Field Unfolding

Rather than running Eq. 9 to asymptotic convergence, LUMIN unrolls the inference into an explicit sequence of $T_{\text{iter}} - 1$ normalized mean-field updates followed by a final unnormalized energy readout stage (see Figure 2).

Unrolled Update Sequence. Starting from the unary prior $Q^{(0)} = \mathrm{Softmax}(U)$, the intermediate posterior states are updated recursively:

$$\mathbf{Q}^{(t+1)} = \mathrm{Softmax}\left(\frac{U + \lambda_t W \mathbf{Q}^{(t)} \mu_\epsilon}{\tau_t}\right), \quad t = 0, \ldots, T_{\text{iter}} - 2, \tag{10}$$

where the stage-wise interaction weights and temperatures are parameterized as $\lambda_t = \mathrm{Softplus}(\theta_{\lambda,t})$ and $\tau_t = \mathrm{Softplus}(\theta_{\tau,t}) + \epsilon_\tau$ to enforce strict nonnegativity. We adopt $T_{\text{iter}} = 3$ as the default configuration and empirically study the effect of inference depth in Section 4.

Parameter Efficiency. Beyond the compact background head g_θ and projection P, LUMIN introduces no stage-specific weight tensors and is governed entirely by stage-wise scalar parameters $\{\lambda_i, \tau_i\}_{i=1}^{T_{\text{iter}}}$.

Majorization-Minimization Semantics. By linearizing the quadratic interaction around $\mathbf{Q}^{(t)}$, we construct a convex surrogate $\mathcal{A}_t(\mathbf{Q}, \mathbf{Q}^{(t)})$ that globally majorizes the stage free energy $\mathcal{F}_t(\mathbf{Q})$ via $W \otimes \mu_\epsilon \succeq 0$. Each unrolled stage executes $\mathbf{Q}^{(t+1)} = \arg\min_{\mathbf{Q} \in \Delta^K} \mathcal{A}_t(\mathbf{Q}, \mathbf{Q}^{(t)})$, strictly guaranteeing stage-wise monotonic descent:

$$\mathcal{F}_t(\mathbf{Q}^{(t+1)}) \leq \mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)}) \leq \mathcal{F}_t(\mathbf{Q}^{(t)}). \tag{11}$$

The learned scalars (λ_t, τ_t) thus realize an exact continuation path over progressively annealed landscapes (proofs in Appendix C).

Why Unfold Mean-Field Inference Rather Than Heuristic Refinement? While LUMIN bears a superficial structural resemblance to multi-layer graph neural networks (GNNs) or cross-attention decoders, their algorithmic nature is fundamentally distinct. Standard message-passing layers implement unconstrained, heuristic feature mixing:

$$\mathbf{Q}^{(t+1)} = \sigma(W \mathbf{Q}^{(t)} \Theta_t),$$

governed by high-dimensional channel-mixing matrices $\Theta_t \in \mathbb{R}^{K \times K}$ without explicit optimization meaning. In contrast, every stage in LUMIN constitutes an exact minimizer of a tight convex surrogate over the probability simplex, parameterized by only two scalar hyperparameters (λ_t, τ_t). The architecture of LUMIN is therefore not an ad-hoc neural network with an empirical recipe, but the physical execution of a stage-wise Majorization-Minimization algorithm.

3.5 Background-Referenced Differential Readout

Following the $T_{\text{iter}} - 1$ normalized updates, the final network stage executes an energy-field readout rather than computing another saturated probability distribution. We compute the unnormalized energy field across all K states:

$$F = U + \lambda_{T_{\text{iter}}-1} W \mathbf{Q}^{(T_{\text{iter}}-1)} \mu_\epsilon \in \mathbb{R}^{N \times K}, \tag{12}$$

where $\lambda_{T_{\text{iter}}-1} \geq 0$ is a dedicated readout-stage interaction strength.

For any specific textual query of interest $T_m \in \mathcal{T}$, its binary segmentation logit is obtained via a background-referenced differential readout that contrasts its semantic field $F_{:,m}$ against the contextual background field $F_{:0}$:

$$\ell_m = s(F_{:,m} - F_{:,0}) + b \in \mathbb{R}^{N \times 1}, \ m \in \{1, \cdots, M\} \tag{13}$$

where $s = \exp(\theta_s) > 0$ is a learnable positive scale and b is a learnable bias.

Mask Realization and Training Objective. A Prototype Generation Network (ProtoNet) produces P spatial mask prototypes from the stride-8 visual features, while lightweight TextFiLM layers modulate these prototypes according to the target query embedding. Prototype routing then combines the text-conditioned prototypes under the guidance of T to synthesize the final full-resolution mask. The model is trained end-to-end with a multi-task objective

$$\mathcal{L}_{\text{total}} = 7.5\mathcal{L}_{\text{mask}} + 7.5\mathcal{L}_{\text{box}} + 1.5\mathcal{L}_{\text{DFL}}, \tag{14}$$

where

$$\mathcal{L}_{\text{mask}} = 0.5\mathcal{L}_{\text{BCEWithLogits}} + 0.5\mathcal{L}_{\text{Dice}}, \tag{15}$$

with the mask loss applied to the target-query prediction, while $\mathcal{L}_{\text{box}}$ and $\mathcal{L}_{\text{DFL}}$ supervise the bounding-box regression and distributional localization branches, respectively.

Table 1: Referring expression segmentation performance (mIoU, %) across all eight standard splits of RefCOCO, RefCOCO+, and RefCOCOg. Trainable visual parameter counts are provided for comparison.

	RefCOCO			RefCOCO+			RefCOCOg	
Method	**val**	**testA**	**testB**	**val**	**testA**	**testB**	**val**	**test**
VLT (Ding et al., 2021)	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
LAVT (Yang et al., 2022)	74.53	76.94	70.89	65.84	71.03	59.20	63.34	62.96
CRIS (Wang et al., 2022)	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
M3Att (Liu et al., 2023b)	73.60	76.23	70.36	65.34	70.50	56.98	64.92	67.92
PolyFormer-L (Liu et al., 2023c)	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17
CM-MaskSD (Wang et al., 2024)	74.85	77.50	71.28	67.51	71.78	59.93	66.49	66.61
LUMIN (Ours)	74.37	75.69	72.36	62.15	68.20	56.88	65.64	66.23

4 EXPERIMENTS

4.1 DATASETS

Multi-Task Pretraining Corpus. LUMIN is pretrained on a unified multi-task grounding corpus comprising approximately 3.37M samples from 2.32M images and spanning multiple grounding-related task domains. This pretraining stage provides the general visual-language grounding capability used by the model. Table 5 in Appendix D, enumerates the publicly available datasets leveraged for LUMIN's training across various subtasks.

Evaluation Benchmarks and Protocols. After pretraining, we directly evaluate LUMIN across three complementary protocols covering standard, boundary-level, and generalized referring expression segmentation without using their training splits for model optimization or task-specific fine-tuning. For standard RES, we report mean Intersection-over-Union (**mIoU**) across all eight canonical splits of RefCOCO, RefCOCO+, and RefCOCOg (Hu et al., 2016; Yu et al., 2016). Finally, we assess generalization to multi-target and null-target expressions on the val, testA, and testB splits of gRefCOCO (Liu et al., 2023a), reporting cumulative IoU (**cIoU**), generalized IoU (**gIoU**), and no-target accuracy (**N-acc**).

4.2 IMPLEMENTATION DETAILS

LUMIN employs a compact visual pipeline totaling 49M trainable parameters, integrating a DINOv3-pretrained ConvNeXt-Tiny backbone (Siméoni et al., 2025), an FPN+PAN neck, and a unified task head, with variational mean-field unrolling operating directly on the stride-8 feature grid ($T_{\text{iter}} = 3$ by default). Text representations are extracted using a frozen Qwen3-VL-Embedding-2B (Li et al., 2026) model, whose native features are precomputed offline. All models are trained on 640×640 letterboxed inputs for 30 epochs with an effective batch size of 256 using fused AdamW (base learning rate 1×10^{-4}). Complete architectural breakdowns, loss weighting schedules, and hardware configurations are detailed in Appendix D.

4.3 MAIN RESULTS

Standard RES Benchmarks. Table 1 summarizes results across the eight canonical splits. With only 49M trainable visual parameters and an offline frozen text embedding space, LUMIN achieves competitive performance under the zero-shot transfer setting. On RefCOCO+, where spatial cues are removed, LUMIN remains competitive, suggesting that query-conditioned semantic interactions can provide useful contextual information beyond explicit spatial cues.

Generalized Referring Expression Segmentation. Table 2 evaluates LUMIN on gRefCOCO, where expressions may target multiple entities or have no valid target. LUMIN achieves strong no-target accuracy (N-acc), indicating effective suppression of false positives through the contextual background state. Target rejection is therefore incorporated into the posterior inference process

Table 2: Performance on generalized referring segmentation (gRefCOCO). Metrics include cumulative IoU (cIoU, %), generalized IoU (gIoU, %), and no-target classification accuracy (N-acc, %).

	gRefCOCO val			gRefCOCO testA			gRefCOCO testB		
Method	**cIoU**	**gIoU**	**N-acc**	**cIoU**	**gIoU**	**N-acc**	**cIoU**	**gIoU**	**N-acc**
LAVT (Yang et al., 2022)	57.6	58.4	49.3	65.3	65.9	49.3	55.0	55.8	48.5
ReLA (Liu et al., 2023a)	62.4	63.6	56.4	69.3	70.0	59.0	59.9	61.0	58.4
LUMIN (Ours)	56.5	58.8	60.3	52.3	53.4	55.2	50.1	50.2	53.2

Table 3: Inference-depth and update-operator ablations on RefCOCO validation. The first four rows evaluate the unfolding trajectory, while the last row tests a heuristic residual propagation rule using the same W and μ_ϵ.

Configuration	**Update Rule**	**mIoU (%)**	**oIoU (%)**
Unary ($T_{\text{iter}} = 0$)	$\mathbf{Q}^{(0)} = \text{Softmax}(U)$	69.5	68.3
1-Stage Mean-Field	$\mathbf{Q}^{(1)} = \text{Softmax}\Big(\frac{U+\lambda_0 W\mathbf{Q}^{(0)}\mu_\epsilon}{\tau_0}\Big)$	73.7	70.7
2-Stage Unfolding	Two-step unrolled inference	74.3	71.2
3-Stage LUMIN	$\arg\min_{\mathbf{Q}\in\Delta^K} \mathcal{A}_t(\mathbf{Q}, \mathbf{Q}^{(t)})$	**74.4**	**71.9**
4-Stage LUMIN	–	**74.4**	**71.9**
Heuristic Propagation	$\mathbf{Q}^{(t+1)} = \text{Softmax}\Big(\mathbf{Q}^{(t)} + \lambda_t W\mathbf{Q}^{(t)}\mu_\epsilon\Big)$	0.00	0.00

rather than handled by post-hoc confidence thresholding. Although the absolute performance figures are not high, it achieved competitive results in a zero-shot, low-parameter setting.

4.4 Ablation Studies

Inference Depth. Starting from the unary baseline (68.3% oIoU), a single mean-field update markedly improves performance to 70.7%, followed by sustained improvements at two and three stages (71.2% and 71.9%). The diminishing margins indicate that a shallow three-stage unfolding is sufficient to capture the bulk of structured variational gains without unnecessary computational overhead.

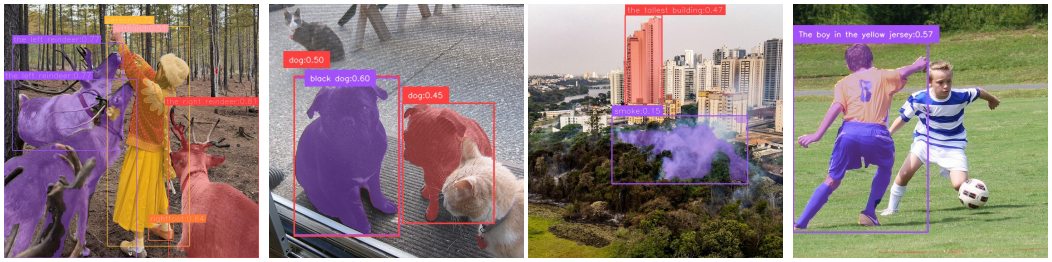
Variational Update vs. Heuristic Propagation. We additionally test an unconstrained residual propagation rule using the exact same visual graph W, text kernel μ_ϵ, initialization, and stage count. Because this heuristic rule lacks explicit unary re-anchoring to U and fails to optimize a valid variational free energy, the propagation quickly diverges and saturates the logits, collapsing to empty mask predictions (0.00% oIoU). We treat this outcome as diagnostic confirmation that the structured Majorization-Minimization update operator is indispensable, demonstrating that the empirical success does not arise merely from generic graph smoothing.

Semantic Compatibility Kernel. The semantic-kernel ablation reveals a substantial advantage from our text-induced geometry. Disabling cross-state compatibility ($\mu_{\text{text}} = I_M$) drops oIoU from 71.9% to 67.6%, as semantic queries cannot coordinate to resolve visual ambiguities. An unconstrained learned matrix achieves only 67.0% and forfeits the positive semidefiniteness required by our convex surrogate bound. In contrast, the regularized Text Gram kernel provides both superior empirical grounding and rigorous alignment with the optimization guarantees of our derivation.

Differential Readout and Interface Dynamics. Directly reading out mask predictions from the normalized posterior field Q_i degenerates completely (0.0% oIoU); because competing textual queries share the simplex denominator, the evaluated channel suffers severe probability mass dilution. Predicting from isolated target energy ($sF_{:,1:} + b$) restores performance to 71.5%, while introducing the background-referenced contrast ($s(F_{:,1:} - F_0) + b$) yields the top performance of 71.9%. This consistent gain validates the role of the contextual background field as an adaptive spatial baseline that rejects scene-wide feature drift.

Table 4: Ablation of the semantic compatibility kernel and mask readout interface on RefCOCO validation. Kernel variants modify strictly the interactions among the M language-derived semantic states. For the readout mechanism, F_i denotes the energy field of the evaluated query state and F_0 denotes the contextual background field.

Axis	Variant	oIoU (%)
Semantic Kernel	Identity: $\mu_{\text{text}} = I_M$	67.6
	Learned: $\mu_{\text{text}} \in \mathbb{R}^{M \times M}$	67.0
	Text Gram: $\mu_{\text{text}} = TT^\top + \epsilon I_M$	**71.9**
Readout Mechanism	Posterior Probability: Q_i	0.00
	Target Energy: $sF_{:,1:} + b$	71.5
	Differential Energy: $s(F_{:,1:} - F_0) + b$	**71.9**



(a) the right reindeer. the left reindeer. person. left hand. right foot. (b) black dog. dog. (c) smoke. the tallest building. (d) the boy in the yellow jersey.

Figure 3: **Qualitative results across diverse referring challenges.** LUMIN qualitatively demonstrates query-dependent grounding under spatial ambiguity and fine-grained part references (*the right/left reindeer*, *left hand*, *right foot*), attribute-based instance disambiguation (*black dog* vs. *dog*), contextual and non-rigid grounding (*the tallest building*, *smoke*), and appearance-based grounding under interaction and occlusion (*the boy in the yellow jersey*). Different textual queries are represented by distinct semantic states, yielding query-specific masks and boxes; masks and boxes of the same query are color-matched.

4.5 Visualizations

Figure 3 further illustrates how LUMIN refines spatial predictions according to different textual queries, covering spatial, attribute, contextual, and fine-grained part-level references. More visualization results can be found in Appendix E.

5 Conclusion

We presented **LUMIN**, a language-conditioned mean-field unfolding framework that reformulates dense vision-language grounding as variational inference over a query-conditioned semantic posterior space. By coupling a text-derived Gram compatibility kernel with a symmetric visual affinity graph and deriving the inference process through majorization-minimization, LUMIN transforms structured posterior refinement into a compact and interpretable three-stage network. A background-relative differential readout further provides an explicit mechanism for separating target evidence from contextual background. Experiments on RefCOCO/+/g, and gRefCOCO demonstrate competitive performance under zero-shot transfer, while the ablation studies support the role of the proposed semantic interactions and iterative refinement. More broadly, LUMIN illustrates how deep unfolding can move beyond task-specific latent coefficient recovery toward language-conditioned semantic inference, providing a general formulation for structured dense grounding and a basis for extending this paradigm to other open-vocabulary vision-language prediction tasks.

AI USE STATEMENT

In this work, we used generative AI tools (specifically, large language models) strictly for grammatical editing, text refinement, and stylistic polishing of the manuscript, as well as assisting in formatting visualization scripts. We have not used generative AI tools for conceiving research hypotheses, formulating mathematical derivations, designing the core algorithmic framework, or analyzing experimental results. All text and code assisted by AI were thoroughly reviewed, verified, and validated by the authors. The authors take full responsibility for the entirety of this work, including all claims, theoretical derivations, empirical findings, and artifacts presented.

ETHICS STATEMENT

This work introduces a foundational framework for open-vocabulary referring expression segmentation and evaluates exclusively on widely adopted, publicly available academic benchmarks (e.g., RefCOCO, RefCOCO+, and G-Ref). Our research does not involve human subject trials, proprietary biometric tracking, crowdsourced personal data collection, or the deployment of surveillance systems. We acknowledge that multimodal vision-language representations may inherit subtle societal, demographic, or cultural biases present in broad web-scale pretraining corpora. We encourage practitioners to exercise due diligence and audit downstream predictions when deploying vision-language models in sensitive or user-facing domains.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of our theoretical and empirical claims, we have documented all components of this work as follows:

- **Theoretical Foundations:** Complete mathematical derivations and rigorous proofs regarding the diagonal PSD completion, Kronecker interaction kernel, and stage-wise Majorization-Minimization monotonic descent are provided in Appendices A and C.
- **Architectural & Training Details:** Comprehensive specifications of the model architecture (including the background head g_θ, feature projection P, and stage-wise scalars $\{\lambda_t, \tau_t\}$), alongside full optimization hyperparameters, loss configurations, and training schedules, are detailed in Section 4 and Appendix D.

REFERENCES

Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip H. S. Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pp. 524–540, 2016.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978, 2014.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16321–16330, 2021.

Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5332–5340. IEEE, 2019.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 399–406, 2010.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pp. 128–145. Springer, 2022.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Junyi Hu, Tian Bai, Fengyi Wu, Yian Huang, Wei Wen, Zaoli Li, Junli Lin, Xingchen Li, Zhenming Peng, and Yi Zhang. Scaling representation diversity: Modulated attention and reconstructive regularization for visual grounding. *arXiv preprint arXiv:2608.12748*, 2026.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pp. 108–124, 2016.

Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4424–4433, 2020.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58 (1):30–37, 2004.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2018a.

Weize Li, Zhicheng Zhao, Haochen Bai, and Fei Su. Bring adaptive binding prototypes to generalized referring expression segmentation. *IEEE Transactions on Multimedia*, 27:6059–6069, 2025.

Xiaojie Li, Lu Yang, Qing Song, and Fuqiang Zhou. Detector-in-detector: Multi-level analysis for human-parts. In *Asian Conference on Computer Vision*, pp. 228–240. Springer, 2018b.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23592–23601, 2023a.

Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. Multi-modal mutual attention and iterative interaction for referring image segmentation. *IEEE Transactions on Image Processing*, 32:3054–3065, 2023b.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1271–1280, 2017.

Jiang Liu, Henghui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R. Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023c.

Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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, 2021.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7141–7151. IEEE, 2023.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438. IEEE, 2019.

Feiyang Shen and Hongping Gan. Hunet: homotopy unfolding network for image compressive sensing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12799–12808. IEEE, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Wei Tang, Xuejing Liu, Yanpeng Sun, and Zechao Li. Ssp-sam: Sam with semantic-spatial prompt for referring expression segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Huynh Van Luong, Boris Joukovsky, Yonina C Eldar, and Nikos Deligiannis. A deep-unfolded reference-based rpca network for video foreground-background separation. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1432–1436. IEEE, 2021.

Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellappa. Gaussian conditional random field network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3224–3233, 2016.

Martin J Wainwright, Michael Irwin Jordan, et al. Graphical models, exponential families, and variational inference. 2008.

Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19787–19797. IEEE, 2023.

Wenxuan Wang, Xingjian He, Yisi Zhang, Longteng Guo, Jiachen Shen, Jiangyun Li, and Jing Liu. Cm-masksd: Cross-modality masked self-distillation for referring image segmentation. *IEEE Transactions on Multimedia*, 26:6906–6916, 2024.

Xiaoyang Wang and Hongping Gan. Spectrally adaptive channel-aware unrolling network for compressed sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10190–10198, 2026.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Fengyi Wu, Tianfang Zhang, Lei Li, Yian Huang, and Zhenming Peng. Rpcanet: Deep unfolding rpca based infrared small target detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4797–4806. IEEE, 2024.

Fengyi Wu, Yimian Dai, Tianfang Zhang, Yixuan Ding, Jian Yang, Ming-Ming Cheng, and Zhenming Peng. Rpcanet^{++}: Deep interpretable robust pca for sparse object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):12420–12440, 2026. doi: 10.1109/TPAMI.2026.3697815.

Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Admm-csnet: A deep learning approach for image compressive sensing. *IEEE transactions on pattern analysis and machine intelligence*, 42 (3):521–538, 2018.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10502–10511, 2019.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85, 2016.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19456–19465, 2023.

Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1828–1837. IEEE, 2018.

Yi Zhang, Jichang Guo, Huihui Yue, Sida Zheng, and Guanhua An. Edulis: Edge-aware deep unfolding network for low-light instance segmentation. *Information Sciences*, pp. 122878, 2025.

Zhonghao Zhang, Yipeng Liu, Jiani Liu, Fei Wen, and Ce Zhu. Amp-net: Denoising-based deep unfolding for compressive image sensing. *IEEE Transactions on Image Processing*, 30:1487–1500, 2020.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1529–1537, 2015.

APPENDIX

This appendix provides theoretical derivations, mathematical proofs, detailed implementation specifications, and extended qualitative evaluations supporting the main text:

- **Appendix A:** Derivation of the variational free energy and the mean-field fixed-point recursion from first principles via finite-dimensional Lagrangian stationarity on the product simplex.
- **Appendix B:** Rigorous proofs of positive semidefiniteness (PSD) for the visual affinity kernel W, the language compatibility kernel μ_ϵ, and their compound Kronecker product $W \otimes \mu_\epsilon$.
- **Appendix C:** Complete Majorization-Minimization (MM) proof establishing the tight convex surrogate bound and stage-wise monotonic free-energy descent.
- **Appendix D:** Comprehensive architectural configurations, optimization hyperparameters, loss formulations, and dataset preprocessing protocols.
- **Appendix E:** Further results and visualizations of the internal unfolding dynamics provide better insight into how LUMIN operates.
- **Appendix F:** Analysis and discussion regarding the limitations of this work and potential future directions.

A CONTINUOUS VARIATIONAL RELAXATION AND FIXED-POINT DERIVATION

This section provides the rigorous foundation connecting the discrete conditional random field (CRF) to our continuous variational relaxation, formalizes the necessity of diagonal PSD completion, and proves the closed-form mean-field fixed-point recursion.

A.1 DISCRETE CRF FORMULATION AND THE INDEFINITENESS OF HOLLOW INTERACTIONS

Let $Z = \{z_1, \dots, z_N\}$ denote a discrete joint label configuration over the N visual nodes, where each latent variable takes a state from the query-induced label space $\mathcal{L}(\mathcal{T}) = \{0, 1, \dots, M\}$ (K states). In a discrete CRF, the authentic joint posterior distribution over latent assignments conditioned on the image I and text prompt $\mathcal{T}$ follows the Gibbs distribution:

$$P(Z \mid I, \mathcal{T}) = \frac{1}{\mathcal{Z}(I, \mathcal{T})} \exp\left(-E(Z \mid I, \mathcal{T})\right), \tag{16}$$

where $\mathcal{Z}(I, \mathcal{T}) = \sum_{Z \in \mathcal{L}(\mathcal{T})^N} \exp\left(-E(Z \mid I, \mathcal{T})\right)$ is the partition function (Wainwright et al., 2008). Computing $\mathcal{Z}(I, \mathcal{T})$ requires summing over $(M+1)^N$ combinatorial configurations, which is NP-hard.

In an authentic discrete random field, the total Gibbs energy consists of prompt-conditioned unary potentials and pairwise spatial regularization acting strictly across distinct node pairs ($i \neq j$):

$$E(Z; I, \mathcal{T}) = -\sum_{i=1}^{N} U_{i,z_i} + E_{\text{pair}}^{\text{off}}(Z), \qquad \text{with} \quad E_{\text{pair}}^{\text{off}}(Z) = -\frac{\lambda}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} W_{ij} (\mu_\epsilon)_{z_i, z_j}. \tag{17}$$

To approximate the intractable posterior $P(Z \mid I, \mathcal{T})$, we define a fully factorized continuous mean-field variational family:

$$\mathcal{Q}(Z) = \prod_{i=1}^{N} Q_i(z_i), \tag{18}$$

where each q_i is a categorical distribution over the $K = M + 1$ semantic states. We represent Q_i by a probability vector

$$\mathbf{q}_i = \left[Q_i(z_i = 0), \dots, Q_i(z_i = M)\right]^\top \in \Delta^K, \tag{19}$$

where

$$\Delta^K = \left\{ \mathbf{q} \in [0,1]^K \;\middle|\; \sum_{k=0}^{M} Q_k = 1 \right\}. \tag{20}$$

Then we have:

$$\mathbf{Q} = \left[\mathbf{q}_1, \cdots, \mathbf{q}_N\right]^\top \in \mathbb{R}^{N \times K}, \qquad Q_{ik} = Q_i(z_i = k). \tag{21}$$

The optimal variational posterior is obtained by minimizing the Kullback-Leibler (KL) divergence from $\mathcal{Q}(Z)$ to the true Gibbs distribution:

$$\begin{aligned} \mathrm{KL}(\mathcal{Q} \parallel P) &= \sum_{Z} \mathcal{Q}(Z) \log \frac{\mathcal{Q}(Z)}{P(Z \mid I, \mathcal{T})} \\ &= \sum_{Z} \mathcal{Q}(Z) \log \mathcal{Q}(Z) + \sum_{Z} \mathcal{Q}(Z) E(Z \mid I, \mathcal{T}) + \log \mathcal{Z}(I, \mathcal{T}) \\ &= \mathbb{E}_{\mathcal{Q}}[E(Z \mid I, \mathcal{T})] - H(\mathcal{Q}) + \log \mathcal{Z}(I, \mathcal{T}), \end{aligned} \tag{22}$$

where $H(\mathcal{Q}) = -\sum_{i=1}^{N} \sum_{k=0}^{M} Q_{ik} \log Q_{ik}$ denotes the Shannon entropy of the factorized distribution. Because the log-partition function $\log \mathcal{Z}(I, \mathcal{T})$ is strictly independent of $\mathcal{Q}$, minimizing $\mathrm{KL}(\mathcal{Q} \parallel P)$ is mathematically equivalent to minimizing the uncompleted variational free energy $\mathcal{F}_{\text{off}}(\mathcal{Q}) \triangleq \mathbb{E}_{\mathcal{Q}}[E(Z)] - H(\mathcal{Q})$.

Evaluating the expectations under the mean-field factorized distribution $\mathcal{Q}(Z)$:

1. **Unary Expectation:** Since $\mathbb{E}_{\mathcal{Q}}[\mathbb{I}(z_i = k)] = Q_{ik}$, the expected unary energy evaluates to:

$$\mathbb{E}_{\mathcal{Q}}\left[-\sum_{i=1}^{N} U_{i,z_i}\right] = -\sum_{i=1}^{N}\sum_{k=0}^{M} U_{ik}Q_{ik} = -\operatorname{Tr}(\mathbf{Q}^\top U). \tag{23}$$

2. **Pairwise Off-Diagonal Expectation:** Under the mean-field independence assumption for distinct nodes ($i \neq j$), indicator expectations decouple strictly: $\mathbb{E}_Q[\mathbb{I}(z_i = k)\mathbb{I}(z_j = l)] = Q_{ik}Q_{jl}$ ($\forall i \neq j$). Therefore:

$$\mathbb{E}_{\mathcal{Q}}\left[E^{\text{off}}_{\text{pair}}(Z)\right] = -\frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{k=0}^{M}\sum_{l=0}^{M} W_{ij}Q_{ik}Q_{jl}(\mu_\epsilon)_{kl} = -\frac{\lambda}{2}\operatorname{Tr}\left(\mathbf{Q}^\top W_{\text{off}}\mathbf{Q}\mu_\epsilon\right), \tag{24}$$

   where $W_{\text{off}} = W - \operatorname{diag}(W)$ is a symmetric hollow matrix with zero diagonal entries: $(W_{\text{off}})_{ii} = 0$.

Proposition 1 (Inherent Indefiniteness of Hollow Graph Kernels)**.** *Let $W_{off} \in \mathbb{R}^{N\times N}$ be a non-zero symmetric hollow matrix ($(W_{off})_{ii} = 0$). Then W_{off} is inherently indefinite; that is, its extremal eigenvalues satisfy $\lambda_{\max}(W_{off}) > 0$ and $\lambda_{\min}(W_{off}) < 0$.*

Proof. Because W_{off} is real and symmetric, its eigenvalues $\{\gamma_i\}_{i=1}^N$ are real numbers. The trace of W_{off} equals both the sum of its diagonal entries and the sum of its eigenvalues:

$$\operatorname{Tr}(W_{\text{off}}) = \sum_{i=1}^{N}(W_{\text{off}})_{ii} = 0 \implies \sum_{i=1}^{N}\gamma_i = 0. \tag{25}$$

Since $W_{\text{off}} \neq \mathbf{0}$, there must exist at least one non-zero eigenvalue. For a set of non-zero real numbers to sum to zero, there must exist at least one strictly positive eigenvalue $\gamma_p > 0$ and at least one strictly negative eigenvalue $\gamma_q < 0$. Therefore, $\lambda_{\max}(W_{\text{off}}) > 0$ and $\lambda_{\min}(W_{\text{off}}) < 0$, which proves that W_{off} is indefinite. □

The Majorization Dilemma: Because W_{off} possesses strictly negative eigenvalues, the compound interaction kernel $W_{\text{off}} \otimes \mu_\epsilon$ is indefinite. Consequently, the quadratic residual difference $\frac{\lambda_t}{2}\operatorname{vec}_r(\mathbf{Q} - R)^\top(W_{\text{off}} \otimes \mu_\epsilon)\operatorname{vec}_r(\mathbf{Q} - R)$ can take negative values, which invalidates the construction of a globally tight convex majorizer for unrolled optimization.

A.2 Diagonal PSD Completion and Vertex Consistency

To establish a mathematically rigorous Majorization-Minimization framework without perturbing off-diagonal spatial message passing, we define our objective via a continuous variational relaxation enhanced by a *diagonal positive semidefinite (PSD) completion*:

$$W = W_{\text{off}} + \operatorname{diag}(W) \succeq 0, \tag{26}$$

where $\operatorname{diag}(W)_{ii} = W_{ii} = W^{\text{raw}}_{ii}/D_{ii} > 0$. This completion forms the continuous pairwise interaction functional:

$$\mathcal{E}_{\text{pair}}(\mathbf{Q}) \triangleq -\frac{\lambda}{2}\operatorname{Tr}(\mathbf{Q}^\top W\mathbf{Q}\mu_\epsilon) = -\frac{\lambda}{2}\operatorname{Tr}(\mathbf{Q}^\top W_{\text{off}}\mathbf{Q}\mu_\epsilon) + \mathcal{R}_{\text{self}}(\mathbf{Q}), \tag{27}$$

where $\mathcal{R}_{\text{self}}(\mathbf{Q}) \triangleq -\frac{\lambda}{2}\operatorname{Tr}(\mathbf{Q}^\top \operatorname{diag}(W)\mathbf{Q}\mu_\epsilon) = -\frac{\lambda}{2}\sum_{i=1}^{N} W_{ii}(Q_i^\top \mu_\epsilon Q_i)$ acts as a local quadratic self-consistency potential.

Proposition 2 (Vertex Consistency)**.** *The continuous quadratic potential $\mathcal{R}_{self}(Q)$ is strictly consistent with the discrete self-interaction energy on all extremal vertices of the probability simplex Δ^K.*

Proof. Consider an arbitrary discrete configuration where every node posterior is an indicator basis vector $Q_i = e_{z_i}$ for $z_i \in \{0, 1, \ldots, M\}$. The inner product evaluates to:

$$Q_i^\top \mu_\epsilon Q_i = e_{z_i}^\top \mu_\epsilon e_{z_i} = (\mu_\epsilon)_{z_i,z_i}. \tag{28}$$

Substituting Eq. 28 into $\mathcal{R}_{\text{self}}(\mathbf{Q})$:

$$\mathcal{R}_{\text{self}}(\mathbf{Q}) = -\frac{\lambda}{2}\sum_{i=1}^{N} W_{ii}(\mu_\epsilon)_{z_i,z_i} \equiv E_{\text{self}}(Z). \tag{29}$$

Thus, on the discrete vertices of the simplex, $\mathcal{R}_{\text{self}}(\mathbf{Q})$ precisely matches the discrete self-energy. Across the continuous interior of the simplex, $Q_i^\top \mu_\epsilon Q_i = \sum_{k,l} Q_{ik}Q_{il}(\mu_\epsilon)_{kl}$ provides a smooth interpolation whose positive semidefiniteness restores the spectral condition $W \otimes \mu_\epsilon \succeq 0$. □

A.3 Lagrangian Stationarity of the Relaxed Objective

Coupling the linear unary observation, the PSD-completed continuous pairwise interaction, and the entropic penalty, the continuous relaxed free energy is defined on the product simplex Δ^K:

$$\mathcal{F}(\mathbf{Q}) = -\operatorname{Tr}(\mathbf{Q}^\top U) - \frac{\lambda}{2}\operatorname{Tr}(\mathbf{Q}^\top W \mathbf{Q}\mu_\epsilon) + \tau\sum_{i=1}^{N}\sum_{k=0}^{M} Q_{ik}\log Q_{ik}, \tag{30}$$

where $\tau > 0$ denotes the variational temperature. We enforce the probability simplex constraints $\sum_{k=0}^{M} Q_{ik} = 1$ for all $i \in \{1,\dots,N\}$ via the Lagrangian functional with multipliers $\boldsymbol{\alpha} = [\alpha_1,\dots,\alpha_N]^\top \in \mathbb{R}^N$:

$$\begin{aligned}\mathcal{L}(\mathbf{Q},\boldsymbol{\alpha}) = &-\sum_{i=1}^{N}\sum_{k=0}^{M} U_{ik}Q_{ik} - \frac{\lambda}{2}\sum_{i,j=1}^{N}\sum_{k,l=0}^{M} W_{ij}Q_{ik}Q_{jl}(\mu_\epsilon)_{kl} \\ &+ \tau\sum_{i=1}^{N}\sum_{k=0}^{M} Q_{ik}\log Q_{ik} + \sum_{i=1}^{N}\alpha_i\left(\sum_{k=0}^{M} Q_{ik} - 1\right).\end{aligned} \tag{31}$$

We compute the partial derivative of $\mathcal{L}$ with respect to an arbitrary component Q_{ik}:

$$\frac{\partial\mathcal{L}}{\partial Q_{ik}} = -U_{ik} - \frac{\lambda}{2}\frac{\partial}{\partial Q_{ik}}\left(\sum_{m,n=1}^{N}\sum_{c,d=0}^{M} W_{mn}Q_{mc}Q_{nd}(\mu_\epsilon)_{cd}\right) + \tau(\log Q_{ik} + 1) + \alpha_i. \tag{32}$$

Applying the product rule to the quadratic summation yields non-zero terms under two disjoint index matchings:

1. When $m = i$ and $c = k$:

$$\Delta_1 \triangleq \sum_{n=1}^{N}\sum_{d=0}^{M} W_{in}Q_{nd}(\mu_\epsilon)_{kd} = (WQ\mu_\epsilon)_{ik}. \tag{33}$$

2. When $n = i$ and $d = k$:

$$\Delta_2 \triangleq \sum_{m=1}^{N}\sum_{c=0}^{M} W_{mi}Q_{mc}(\mu_\epsilon)_{ck}. \tag{34}$$

Because the completed spatial affinity is strictly symmetric ($W_{mi} = W_{im}$) and the text compatibility kernel is strictly symmetric ($(\mu_\epsilon)_{ck} = (\mu_\epsilon)_{kc}$), we substitute these identities into Δ_2 and rename the dummy summation indices ($m \to n, c \to d$):

$$\Delta_2 = \sum_{n=1}^{N}\sum_{d=0}^{M} W_{in}Q_{nd}(\mu_\epsilon)_{kd} \equiv \Delta_1 = (WQ\mu_\epsilon)_{ik}. \tag{35}$$

Therefore, the variation of the quadratic interaction evaluates to:

$$\frac{\partial}{\partial Q_{ik}}\left[\frac{1}{2}\operatorname{Tr}(\mathbf{Q}^\top W\mathbf{Q}\mu_\epsilon)\right] = \frac{1}{2}(\Delta_1 + \Delta_2) = \frac{1}{2}\cdot 2(W\mathbf{Q}\mu_\epsilon)_{ik} = (W\mathbf{Q}\mu_\epsilon)_{ik}. \tag{36}$$

The factor of 2 generated by the structural symmetries of W and μ_ϵ exactly cancels the scalar $\frac{1}{2}$. Setting the first-order variation to zero:

$$\frac{\partial\mathcal{L}}{\partial Q_{ik}} = -U_{ik} - \lambda(W\mathbf{Q}^*\mu_\epsilon)_{ik} + \tau(\log Q^*_{ik} + 1) + \alpha_i = 0. \tag{37}$$

A.4 Simplex Normalization and Closed-Form Fixed Point

Isolating the log-posterior term in Eq. 37:

$$\tau \log Q^*_{ik} = U_{ik} + \lambda (W Q^* \mu_\epsilon)_{ik} - \tau - \alpha_i. \tag{38}$$

Dividing by $\tau > 0$ and exponentiating both sides yields:

$$Q^*_{ik} = \exp\left(\frac{U_{ik} + \lambda (W \mathbf{Q}^* \mu_\epsilon)_{ik}}{\tau}\right) \cdot \exp\left(-1 - \frac{\alpha_i}{\tau}\right). \tag{39}$$

The scalar factor $C_i \triangleq \exp\left(-1 - \frac{\alpha_i}{\tau}\right)$ is strictly positive and depends only on node i, independent of label k. Substituting Eq. 39 into the simplex normalization condition $\sum_{k'=0}^{M} Q^*_{ik'} = 1$:

$$C_i \sum_{k'=0}^{M} \exp\left(\frac{U_{ik'} + \lambda (W \mathbf{Q}^* \mu_\epsilon)_{ik'}}{\tau}\right) = 1 \implies C_i = \left[\sum_{k'=0}^{M} \exp\left(\frac{U_{ik'} + \lambda (W \mathbf{Q}^* \mu_\epsilon)_{ik'}}{\tau}\right)\right]^{-1}. \tag{40}$$

Substituting C_i back into Eq. 39 yields the explicit coordinate-wise fixed-point formula:

$$Q^*_{ik} = \frac{\exp\left(\frac{1}{\tau}[U_{ik} + \lambda (W \mathbf{Q}^* \mu_\epsilon)_{ik}]\right)}{\sum_{k'=0}^{M} \exp\left(\frac{1}{\tau}[U_{ik'} + \lambda (W \mathbf{Q}^* \mu_\epsilon)_{ik'}]\right)}. \tag{41}$$

When expressed across all N visual nodes simultaneously, the row-wise exponential normalization corresponds to the row-wise Softmax operator:

$$\mathbf{Q}^* = \mathrm{Softmax}\left(\frac{U + \lambda W \mathbf{Q}^* \mu_\epsilon}{\tau}\right). \tag{42}$$

This completes the derivation of the mean-field-style fixed-point equation for the continuous relaxed objective in Eq. 9. ■

B Positive Semidefiniteness of Geometric Kernels

Here we establish the spectral properties of the constituent matrices and their compound Kronecker product, which guarantee the validity of the Majorization-Minimization surrogate.

Lemma 1 (Positive Semidefiniteness of Visual Affinity W)**.** *The symmetric degree-normalized visual affinity matrix W defined in Eq. 6 satisfies $W \succeq 0$.*

Proof. Let $\tilde{E} \in \mathbb{R}^{N \times d}$ denote the ℓ_2-normalized projected visual features. The cosine similarity matrix $S = \tilde{E}\tilde{E}^\top$ is a Gram matrix, which is positive semidefinite ($S \succeq 0$). The raw affinity matrix is constructed via the element-wise exponential kernel:

$$W^{\text{raw}} = \exp\left(\frac{S}{\sigma^2}\right) = \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{S}{\sigma^2}\right)^{\circ r}, \tag{43}$$

where $A^{\circ r}$ denotes the r-th Hadamard (element-wise) power of matrix A. By the Schur product theorem (Horn & Johnson, 2012), the Hadamard product of two PSD matrices is PSD. Because $S \succeq 0$ and $\sigma^2 > 0$, by induction, $(S/\sigma^2)^{\circ r} \succeq 0$ for all integers $r \geq 0$. Since the cone of PSD matrices is closed under non-negative linear combinations, the infinite sum converges to a PSD matrix, establishing $W^{\text{raw}} \succeq 0$.

Next, degree normalization is defined as $W = D^{-1/2} W^{\text{raw}} D^{-1/2}$, where $D_{ii} = \sum_j W^{\text{raw}}_{ij} > 0$ because all entries of W^{raw} are strictly positive under the exponential kernel. Thus, $D^{-1/2}$ is a well-defined, positive diagonal matrix. For any non-zero vector $v \in \mathbb{R}^N$:

$$v^\top W v = v^\top (D^{-1/2} W^{\text{raw}} D^{-1/2}) v = (D^{-1/2} v)^\top W^{\text{raw}} (D^{-1/2} v) \geq 0, \tag{44}$$

which follows directly from $W^{\text{raw}} \succeq 0$. Therefore, $W \succeq 0$. □

Lemma 2 (Strict Positive Definiteness of Compatibility Kernel μ_ϵ)**.** *The regularized text compatibility kernel μ_ϵ defined in Eq. 5 satisfies $\mu_\epsilon \succ 0$.*

Proof. The text embedding matrix $T \in \mathbb{R}^{M\times d}$ defines the text Gram block $\mu_{\text{text}} = TT^\top \in \mathbb{R}^{M\times M}$, which satisfies $\mu_{\text{text}} \succeq 0$. In Eq. 5, the background state is assigned a non-negative self-compatibility parameter $\mu_{\text{bg}} \geq 0$ with zero cross-compatibility to text queries:

$$\mu(T) = \begin{bmatrix} \mu_{\text{bg}} & \mathbf{0}^\top \\ \mathbf{0} & TT^\top \end{bmatrix}. \tag{45}$$

For any $u = [u_0, u_{\text{text}}^\top]^\top \in \mathbb{R}^K$:

$$u^\top \mu(T) u = \mu_{\text{bg}} u_0^2 + u_{\text{text}}^\top (TT^\top) u_{\text{text}} \geq 0, \tag{46}$$

confirming that $\mu(T) \succeq 0$. Adding the regularization term ϵI_K with $\epsilon > 0$ shifts the entire spectrum:

$$\lambda_{\min}(\mu_\epsilon) = \lambda_{\min}(\mu(T)) + \epsilon \geq 0 + \epsilon = \epsilon > 0. \tag{47}$$

Hence, μ_ϵ is strictly positive definite ($\mu_\epsilon \succ 0$). □

Lemma 3 (Positive Semidefiniteness of the Compound Kronecker Kernel)**.** *Under row-major vectorization* $\text{vec}_r(\cdot)$*, the compound interaction kernel* $B = W \otimes \mu_\epsilon \in \mathbb{R}^{NK\times NK}$ *satisfies* $B \succeq 0$*.*

Proof. By the spectral properties of the Kronecker product, if matrix W has eigenvalues $\{\alpha_i\}_{i=1}^N$ and matrix μ_ϵ has eigenvalues $\{\beta_k\}_{k=1}^K$, then the eigenvalues of $W \otimes \mu_\epsilon$ are precisely the set of products:

$$\text{spec}(W \otimes \mu_\epsilon) = \{\alpha_i \cdot \beta_k \mid 1 \leq i \leq N,\ 1 \leq k \leq K\}. \tag{48}$$

From Lemma 1, $\alpha_i \geq 0$ for all $i \in \{1, \ldots, N\}$. From Lemma 2, $\beta_k \geq \epsilon > 0$ for all $k \in \{1, \ldots, K\}$. Consequently:

$$\alpha_i \cdot \beta_k \geq 0, \quad \forall i, k. \tag{49}$$

All eigenvalues of $W \otimes \mu_\epsilon$ are non-negative, confirming that $B = W \otimes \mu_\epsilon \succeq 0$. □

C Proof of Stage-wise Majorization-Minimization Descent

This section formalizes the optimization semantics of the unrolled recurrence. We prove that each unrolled mean-field layer minimizes a globally valid convex surrogate of its stage-specific continuous relaxed free energy, establishing monotonic descent and characterizing the rate of energy dissipation.

C.1 Mathematical Statement

Fix a stage $t \in \{0, \ldots, T_{\text{iter}} - 2\}$ with learned stage parameters $\lambda_t \geq 0$ and $\tau_t > 0$. As established in Appendix A, the stage-specific continuous relaxed free energy over the product probability simplex $\Delta^{N\times K} \triangleq \prod_{i=1}^N \Delta^K$ is:

$$\mathcal{F}_t(\mathbf{Q}) = -\operatorname{Tr}(\mathbf{Q}^\top U) - \frac{\lambda_t}{2} \operatorname{Tr}(\mathbf{Q}^\top W \mathbf{Q} \mu_\epsilon) + \tau_t \sum_{i=1}^N \sum_{k=0}^M Q_{ik} \log Q_{ik}. \tag{50}$$

Let $R = \mathbf{Q}^{(t)} \in \Delta^{N\times K}$ denote the current posterior state. We construct the auxiliary surrogate functional $\mathcal{A}_t : \Delta^{N\times K} \times \Delta^{N\times K} \to \mathbb{R}$ by linearizing the concave quadratic interaction around R:

$$\mathcal{A}_t(\mathbf{Q}, R) \triangleq -\operatorname{Tr}(\mathbf{Q}^\top U) - \lambda_t \operatorname{Tr}(\mathbf{Q}^\top W R \mu_\epsilon) + \frac{\lambda_t}{2} \operatorname{Tr}(R^\top W R \mu_\epsilon) + \tau_t \sum_{i=1}^N \sum_{k=0}^M Q_{ik} \log Q_{ik}. \tag{51}$$

This construction follows the standard majorization-minimization (MM) principle, in which the surrogate upper-bounds the target objective and is tangent at the current iterate (Hunter & Lange, 2004).

Theorem 1 (Stage-wise Majorization-Minimization Descent)**.** *For any current posterior state* $R = \mathbf{Q}^{(t)} \in \Delta^{N\times K}$*:*

(i) ***Global Upper Bound with Tangency:*** *$\mathcal{A}_t(\mathbf{Q}, R)$ globally majorizes the stage-specific relaxed free energy:*

$$\mathcal{A}_t(\mathbf{Q}, R) \geq \mathcal{F}_t(\mathbf{Q}), \quad \forall \mathbf{Q} \in \Delta^{N\times K}, \qquad \text{with tangency} \quad \mathcal{A}_t(R, R) = \mathcal{F}_t(R). \tag{52}$$

(ii) ***Unique Global Minimizer:*** *$\mathcal{A}_t(\mathbf{Q}, R)$ is strictly convex in $\mathbf{Q}$ on the interior of the product simplex, and its unique global minimizer analytically coincides with the unrolled network step:*

$$\mathbf{Q}^{(t+1)} \triangleq \arg\min_{\mathbf{Q}\in\Delta^{N\times K}} \mathcal{A}_t(\mathbf{Q}, R) = \mathrm{Softmax}\left(\frac{U + \lambda_t W R \mu_\epsilon}{\tau_t}\right). \tag{53}$$

(iii) ***Stage-wise Monotone Descent:*** *The updated state satisfies monotonic non-increasing free-energy descent:*

$$\mathcal{F}_t(\mathbf{Q}^{(t+1)}) \leq \mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)}) \leq \mathcal{A}_t(\mathbf{Q}^{(t)}, \mathbf{Q}^{(t)}) = \mathcal{F}_t(\mathbf{Q}^{(t)}). \tag{54}$$

Strict descent ($\mathcal{F}_t(\mathbf{Q}^{(t+1)}) < \mathcal{F}_t(\mathbf{Q}^{(t)})$) is guaranteed whenever $\mathbf{Q}^{(t+1)} \neq \mathbf{Q}^{(t)}$ by virtue of the strict convexity of the negative entropy regularizer.

C.2 Proof of Theorem 1

Proof. **Part (i): Global Upper Bound and Tangency.** Subtracting Eq. 50 from Eq. 51:

$$\begin{aligned}\mathcal{A}_t(\mathbf{Q}, R) - \mathcal{F}_t(\mathbf{Q}) &= \frac{\lambda_t}{2}\operatorname{Tr}(\mathbf{Q}^\top W \mathbf{Q}\mu_\epsilon) - \lambda_t \operatorname{Tr}(\mathbf{Q}^\top W R \mu_\epsilon) + \frac{\lambda_t}{2}\operatorname{Tr}(R^\top W R \mu_\epsilon) \\ &= \frac{\lambda_t}{2}\operatorname{Tr}\Big((\mathbf{Q} - R)^\top W (\mathbf{Q} - R)\mu_\epsilon\Big).\end{aligned} \tag{55}$$

Let $\mathrm{vec}_r(X) \in \mathbb{R}^{NK}$ denote the vector formed by stacking the rows of matrix $X \in \mathbb{R}^{N\times K}$. By the trace-Kronecker vectorization identity:

$$\operatorname{Tr}(X^\top W X \mu_\epsilon) = \mathrm{vec}_r(X)^\top (W \otimes \mu_\epsilon)\, \mathrm{vec}_r(X). \tag{56}$$

Setting $X = \mathbf{Q} - R$, the difference in Eq. 55 becomes:

$$\mathcal{A}_t(\mathbf{Q}, R) - \mathcal{F}_t(\mathbf{Q}) = \frac{\lambda_t}{2}\, \mathrm{vec}_r(\mathbf{Q} - R)^\top (W \otimes \mu_\epsilon)\, \mathrm{vec}_r(\mathbf{Q} - R). \tag{57}$$

From Lemma 3, the compound interaction kernel $W \otimes \mu_\epsilon \succeq 0$ is positive semidefinite. Since $\lambda_t \geq 0$, the quadratic form is non-negative for all $\mathbf{Q}, R \in \Delta^{N\times K}$:

$$\mathcal{A}_t(\mathbf{Q}, R) - \mathcal{F}_t(\mathbf{Q}) \geq 0 \implies \mathcal{A}_t(\mathbf{Q}, R) \geq \mathcal{F}_t(\mathbf{Q}). \tag{58}$$

At the expansion point $\mathbf{Q} = R$, $\mathrm{vec}_r(\mathbf{Q} - R) = \mathbf{0}$, yielding $\mathcal{A}_t(R, R) - \mathcal{F}_t(R) = 0$. Hence, $\mathcal{A}_t(\cdot, R)$ is a globally valid majorizer of $\mathcal{F}_t(\cdot)$ tangent at R.

Part (ii): Strict Convexity and Closed-Form Minimization. Expanding Eq. 51 in coordinate form:

$$\mathcal{A}_t(\mathbf{Q}, R) = -\sum_{i=1}^{N}\sum_{k=0}^{M} [U_{ik} + \lambda_t (W R \mu_\epsilon)_{ik}]\, Q_{ik} + \tau_t \sum_{i=1}^{N}\sum_{k=0}^{M} Q_{ik} \log Q_{ik} + \mathrm{const}(R). \tag{59}$$

The first term is linear in $\mathbf{Q}$. The second term is the negative Shannon entropy scaled by $\tau_t > 0$. The second-order partial derivatives evaluate to:

$$\frac{\partial^2 \mathcal{A}_t}{\partial Q_{ik}^2} = \frac{\tau_t}{Q_{ik}} > 0, \qquad \frac{\partial^2 \mathcal{A}_t}{\partial Q_{ik}\partial Q_{jl}} = 0 \quad ((i,k) \neq (j,l)). \tag{60}$$

The Hessian $\nabla_Q^2 \mathcal{A}_t(\mathbf{Q}, R)$ is strictly positive definite on the interior of the product simplex $\Delta^{N\times K}$. Therefore, $\mathcal{A}_t(\mathbf{Q}, R)$ is strictly convex in $\mathbf{Q}$, guaranteeing a unique global minimizer.

Introducing Lagrange multipliers $\boldsymbol{\beta} \in \mathbb{R}^N$ enforcing $\sum_{k=0}^{M} Q_{ik} = 1$, the Lagrangian is:

$$\mathcal{L}_{\mathcal{A}}(\mathbf{Q}, \boldsymbol{\beta}) = \mathcal{A}_t(\mathbf{Q}, R) + \sum_{i=1}^{N} \beta_i \left(\sum_{k=0}^{M} Q_{ik} - 1\right). \tag{61}$$

Setting the gradient with respect to Q_{ik} to zero:

$$\frac{\partial \mathcal{L}_{\mathcal{A}}}{\partial Q_{ik}} = -\left[U_{ik} + \lambda_t (WR\mu_\epsilon)_{ik}\right] + \tau_t(\log Q_{ik} + 1) + \beta_i = 0. \tag{62}$$

Isolating Q_{ik} and normalizing across the simplex constraints via the steps established in Appendix A:

$$Q^*_{ik} = \frac{\exp\left(\frac{1}{\tau_t}[U_{ik} + \lambda_t (WR\mu_\epsilon)_{ik}]\right)}{\sum_{k'=0}^{M} \exp\left(\frac{1}{\tau_t}[U_{ik'} + \lambda_t (WR\mu_\epsilon)_{ik'}]\right)} \implies \mathbf{Q}^* = \mathrm{Softmax}\left(\frac{U + \lambda_t WR\mu_\epsilon}{\tau_t}\right). \tag{63}$$

Setting $R = \mathbf{Q}^{(t)}$ proves that the layer update $\mathbf{Q}^{(t+1)}$ in Eq. 10 is the unique global minimizer of $\mathcal{A}_t(\mathbf{Q}, \mathbf{Q}^{(t)})$.

Part (iii): Stage-wise Monotone Descent and Dissipation Bound. By the definition of the unique global minimizer:

$$\mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)}) \le \mathcal{A}_t(\mathbf{Q}^{(t)}, \mathbf{Q}^{(t)}). \tag{64}$$

Combining Eq. 64 with the majorization bound in Eq. 52:

$$\mathcal{F}_t(\mathbf{Q}^{(t+1)}) \le \mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)}) \le \mathcal{A}_t(\mathbf{Q}^{(t)}, \mathbf{Q}^{(t)}) = \mathcal{F}_t(\mathbf{Q}^{(t)}). \tag{65}$$

To quantify the exact energy dissipation during each update, we express the surrogate difference in terms of the Bregman divergence induced by the strictly convex negative entropy. For any $Q \in \Delta^{N\times K}$, the surrogate can be rewritten around its minimizer $\mathbf{Q}^{(t+1)}$ as:

$$\mathcal{A}_t(\mathbf{Q}, \mathbf{Q}^{(t)}) - \mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)}) = \tau_t \sum_{i=1}^{N} \mathrm{KL}\left(Q_i \,\middle\|\, Q_i^{(t+1)}\right), \tag{66}$$

where $\mathrm{KL}(p \parallel q) \triangleq \sum_k p_k \log(p_k/q_k)$ is the discrete Kullback-Leibler divergence. Evaluating Eq. 66 at $\mathbf{Q} = \mathbf{Q}^{(t)}$:

$$\mathcal{A}_t(\mathbf{Q}^{(t)}, \mathbf{Q}^{(t)}) - \mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)}) = \tau_t \sum_{i=1}^{N} \mathrm{KL}\left(Q_i^{(t)} \,\middle\|\, Q_i^{(t+1)}\right). \tag{67}$$

Combining Eq. 57 and Eq. 67, the total stage-specific free-energy reduction evaluates to:

$$\begin{aligned}
\mathcal{F}_t(\mathbf{Q}^{(t)}) - \mathcal{F}_t(\mathbf{Q}^{(t+1)}) &= \left[\mathcal{A}_t(\mathbf{Q}^{(t)}, \mathbf{Q}^{(t)}) - \mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)})\right] && (68)\\
&\quad + \left[\mathcal{A}_t(\mathbf{Q}^{(t+1)}, \mathbf{Q}^{(t)}) - \mathcal{F}_t(\mathbf{Q}^{(t+1)})\right] \\
&= \tau_t \sum_{i=1}^{N} \mathrm{KL}\left(Q_i^{(t)} \,\middle\|\, Q_i^{(t+1)}\right) && (69)\\
&\quad + \frac{\lambda_t}{2} \,\mathrm{vec}_r(\Delta\mathbf{Q})^\top (W \otimes \mu_\epsilon)\, \mathrm{vec}_r(\Delta\mathbf{Q}), && (70)
\end{aligned}$$

where $\Delta\mathbf{Q} = \mathbf{Q}^{(t+1)} - \mathbf{Q}^{(t)}$. Both terms on the right-hand side of Eq. 68 are non-negative.

By Pinsker's inequality, $\mathrm{KL}(Q_i^{(t)} \parallel Q_i^{(t+1)}) \ge \frac{1}{2}\|Q_i^{(t)} - Q_i^{(t+1)}\|_1^2$. Therefore, even if the residual $\Delta\mathbf{Q}$ falls into the null space of the compound kernel $W \otimes \mu_\epsilon$, the strict positivity of the entropic divergence guarantees:

$$\mathcal{F}_t(\mathbf{Q}^{(t)}) - \mathcal{F}_t(\mathbf{Q}^{(t+1)}) \ge \frac{\tau_t}{2} \sum_{i=1}^{N} \left\|Q_i^{(t)} - Q_i^{(t+1)}\right\|_1^2 > 0, \qquad \forall \mathbf{Q}^{(t+1)} \ne \mathbf{Q}^{(t)}. \tag{71}$$

Thus, strict descent $\mathcal{F}_t(\mathbf{Q}^{(t+1)}) < \mathcal{F}_t(\mathbf{Q}^{(t)})$ holds whenever the posterior updates, terminating if and only if $\mathbf{Q}^{(t+1)} = \mathbf{Q}^{(t)}$ reaches a stationary fixed point. □

Table 5: Overview of pretraining datasets grouped by task domain. *OpenImages-Caption is a re-annotated version of the OpenImages dataset, created using the annotation pipeline described in (Hu et al., 2026).

Task	Datasets
General Object Detection	Objects365 (Shao et al., 2019), OpenImages (Kuznetsova et al., 2020)
Referring Expression Grounding	Flickr30k Entities (Plummer et al., 2015), Objects365-Caption (Hu et al., 2026), OpenImages-Caption*
Part-Level	PACO (Ramanathan et al., 2023), HumanParts (Li et al., 2018b), PartImageNet (He et al., 2022), PascalParts (Chen et al., 2014)
Long-Tail & Fine-Grained	V3Det (Wang et al., 2023), DeepFashion2 (Ge et al., 2019)

C.3 Stage-wise Optimization vs. Global Variational Continuation

In this subsection, we clarify the structural semantics between untied layer parameterization and global variational descent:

- **Stage-wise Adaptive Annealing (LUMIN Default):** When stage parameters $\{\lambda_t, \tau_t\}_{t=0}^{T_{\text{iter}}-2}$ are untied across depth, the unrolled network does not execute iterations on a static energy function. Instead, it implements a *learned variational continuation method*: early layers operate with higher variational temperatures τ_0 and moderate smoothing λ_0 to bypass spurious local minima, while later stages sharpen boundaries with cooler temperatures. Theorem 1 guarantees that each individual stage executes an exact surrogate minimization step on its dedicated landscape: $\mathcal{F}_t(\mathbf{Q}^{(t+1)}) \leq \mathcal{F}_t(\mathbf{Q}^{(t)})$.
- **Time-Invariant Lyapunov Descent (Tied Parameterization):** If parameters are constrained to be tied across all layers ($\lambda_t = \lambda, \tau_t = \tau$), the objective functional remains static: $\mathcal{F}_t(\cdot) \equiv \mathcal{F}(\cdot)$. In this case, Theorem 1 proves that the feed-forward unrolled architecture executes exact monotonic descent along a unified Lyapunov free-energy landscape:

$$\mathcal{F}(\mathbf{Q}^{(0)}) \geq \mathcal{F}(\mathbf{Q}^{(1)}) \geq \cdots \geq \mathcal{F}(\mathbf{Q}^{(T_{\text{iter}}-1)}). \tag{72}$$

D Detailed Implementation and Experimental Configurations

This section provides comprehensive specifications of the pretraining data composition, architectural components, offline multimodal representation pipelines, spatial graph computation, training objectives, and hardware infrastructure.

D.1 Pretraining Dataset Compilation

To equip LUMIN with open-vocabulary grounding across both object-level and fine-grained compositional regions, the unified pretraining corpus integrates diverse public datasets across four core task domains. Table 5 summarizes the data sources and their functional categorizations.

D.2 Architectural Specification and Parameter Accounting

LUMIN isolates task-specific processing within a compact visual architecture comprising **49.08M trainable parameters**. Table 6 details the module-level parameter breakdown, highlighting the minimal parameter footprint of our unrolled inference dynamics.

Visual Backbone. The image feature extractor is a ConvNeXt-Tiny architecture initialized with self-supervised weights pretrained via DINOv3 on the LVD-1689M dataset. Given an input image $I \in \mathbb{R}^{3\times 640\times 640}$, the backbone outputs multi-scale hierarchical feature maps with channel dimensions $\{96, 192, 384, 768\}$ at downsampling strides $\{4, 8, 16, 32\}$.

Subsystem	Constituent Modules	Trainable Params	Ratio
Visual Backbone	ConvNeXt-Tiny (DINOv3 pretrain)	27.82M	56.68%
Feature Neck	FPN + PAN (Strides 8/16/32, $d = 256$)	12.55M	25.57%
Task Head	DetectwSegment (Box, DFL, Proto, FiLM)	8.71M	17.75%
Unfolding Dynamics	Stage scalars $\{\lambda_t, \tau_t, s, b\}$	**6 scalars** (≈ 0)	$< 0.01\%$
Text Encoder	Frozen Qwen3-VL-Embedding-2B (Offline)	**0.00M**	0.00%
Total Visual Pipeline	**Complete Trainable Architecture**	**49.08M**	**100.0%**

Table 6: Detailed architectural parameter breakdown of the visual pipeline in LUMIN.

Feature Pyramid and Path Aggregation Neck. The neck couples top-down Feature Pyramid Networks (FPN) with bottom-up Path Aggregation Networks (PAN). All pyramid levels are projected to a uniform channel width of $d_v = 256$. While object bounding-box proposals leverage multi-scale representations across strides 8, 16, and 32, our dense text-conditioned variational inference is explicitly anchored on the high-resolution stride-8 feature grid ($H' \times W' = 80 \times 80$, yielding $N = 6,400$ spatial visual nodes).

Task Head and Mask Realization. The unified DetectwSegment head coordinates multiple task branches:

- **Bounding-Box & Distribution Focal Loss (DFL):** Predicts four-directional bounding-box offsets parameterized via continuous distribution focal loss alongside box regression.
- **Prototype Generation Network (ProtoNet):** A convolutional subnet operating on stride-8 features to output $P = 32$ spatial mask prototypes $\mathcal{B} \in \mathbb{R}^{32\times160\times160}$ at stride 4.
- **Lightweight TextFiLM Modulation:** Point-wise linear layers producing affine modulation factors (γ, β) conditioned on linguistic embeddings to adapt visual prototypes dynamically.
- **Final Mask Synthesis:** The background-referenced differential logit field $\ell_m \in \mathbb{R}^N$ acts as a spatial attention prior, which is linearly combined with mask prototypes via dot-product routing to generate full-resolution instance masks at 640×640.

D.3 Language Pipeline and Matryoshka Representation Truncation

Offline Text Embedding Cache. Text representations are generated using the pre-trained Qwen3-VL-Embedding-2B model. To achieve maximal computational efficiency and circumvent repeated text forward passes over identical phrases across training epochs, all **6.33 million unique captions** present in our unified pretraining corpus are embedded offline and stored in a memory-mapped columnar cache. This cache also includes 18,777 global negative category phrases extracted from broader lexical ontologies. During training, the text encoder remains completely frozen, incurring zero backpropagation overhead and zero GPU memory allocation.

Matryoshka Truncation ($2048 \to 768$). Qwen3-VL-Embedding-2B is trained with Matryoshka Representation Learning (MRL). Its native 2048-dimensional feature vectors are sliced to the first $d = 768$ dimensions:

$$t_m = \frac{\tilde{t}_m[:768]}{\|\tilde{t}_m[:768]\|_2}, \qquad \text{where } \tilde{t}_m \in \mathbb{R}^{2048}. \tag{73}$$

This dimensionality reduction provides a 62.5% saving in text storage and matrix multiplication FLOPs while preserving downstream retrieval fidelity.

Dynamic In-Batch Text Sampling. During pretraining, each image is loaded with its ground-truth positive query expressions. To train the dynamic semantic compatibility kernel μ_ϵ and prevent gradient collapse toward trivial solutions, the label space is padded up to a fixed capacity of $K_{\text{train}} = 80$ semantic channels per image by randomly sampling negative phrases from the 18,777 global negative pool.

D.4 STRIDE-8 DENSE SPATIAL GRAPH CONSTRUCTION

For an input image of size 640×640, the stride-8 visual feature map consists of $N = 80 \times 80 = 6,400$ nodes. The spatial graph operations proceed as follows:

1. **Feature Projection and Normalization:** Visual node features $e_i \in \mathbb{R}^{256}$ are linearly projected to a shared embedding space via $P \in \mathbb{R}^{768\times256}$ and normalized onto the unit hypersphere: $\tilde{e}_i = Pe_i/\|Pe_i\|_2$.
2. **Scaled Dot-Product Similarity:** The dense similarity matrix $S \in \mathbb{R}^{N\times N}$ is computed via scaled inner products:
$$S_{ij} = \langle \tilde{e}_i, \tilde{e}_j \rangle. \tag{74}$$
3. **Strict Symmetrization:** To prevent floating-point asymmetry from non-deterministic CUDA matrix multiplication routines, we apply:
$$S_{\text{sym}} = \frac{1}{2}\left(S + S^\top\right). \tag{75}$$
4. **Gaussian Kernel Activation:** The raw affinity matrix is obtained through an exponential kernel:
$$W_{ij}^{\text{raw}} = \exp\left(\frac{S_{\text{sym},ij}}{\sigma^2}\right), \tag{76}$$
where $\sigma > 0$ is a learnable spatial bandwidth parameterized via $\sigma = \text{Softplus}(\theta_\sigma) + 10^{-4}$ (initialized to $\sigma_0 = 1.0$).
5. **Symmetric Degree Normalization:** Denoting $D_{ii} = \sum_{j=1}^{N} W_{ij}^{\text{raw}}$, the doubly normalized visual affinity matrix is formed via:
$$W = D^{-1/2} W^{\text{raw}} D^{-1/2}. \tag{77}$$

The matrix multiplication $WQ\mu_\epsilon$ is executed using FP16 Tensor Cores, incurring a latency of only ≈ 2.4 ms per unfolding step.

D.5 TRAINING LOSS AND OPTIMIZATION PROTOCOLS

Comprehensive Multi-Task Objective. The unified model is trained end-to-end under a composite multi-task loss:
$$\mathcal{L}_{\text{total}} = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{dfl}}\mathcal{L}_{\text{DFL}}, \tag{78}$$
where the loss coefficients are set to $\lambda_{\text{mask}} = 7.5$, $\lambda_{\text{box}} = 7.5$, and $\lambda_{\text{dfl}} = 1.5$.

Mask Loss Formulation. The mask loss $\mathcal{L}_{\text{mask}}$ supervises the final text-conditioned mask prediction produced by prototype routing, rather than the intermediate differential logit ℓ_m directly. Let L_{mask} denote the resulting full-resolution mask logit. We define
$$\mathcal{L}_{\text{mask}} = 0.5 \cdot \mathcal{L}_{\text{BCEWithLogits}}(L_{\text{mask}}, Y) + 0.5 \cdot \mathcal{L}_{\text{Dice}}(\sigma(L_{\text{mask}}), Y), \tag{79}$$
where
$$\mathcal{L}_{\text{BCEWithLogits}}(L_{\text{mask}}, Y) = -\frac{1}{N}\sum_{i=1}^{N}\left[Y_i \log\sigma(L_{\text{mask},i}) + (1 - Y_i)\log(1 - \sigma(L_{\text{mask},i}))\right], \tag{80}$$
and
$$\mathcal{L}_{\text{Dice}}(\hat{Y}, Y) = 1 - \frac{2\sum_i \hat{Y}_i Y_i + \epsilon_{\text{dice}}}{\sum_i \hat{Y}_i^2 + \sum_i Y_i^2 + \epsilon_{\text{dice}}}, \qquad \hat{Y} = \sigma(L_{\text{mask}}), \tag{81}$$
with $\epsilon_{\text{dice}} = 1.0$. The differential logit ℓ_m influences the mask prediction through the prototype-routing module and is therefore optimized indirectly by the final mask supervision.

Optimization Schedule. We summarize the exact hyperparameter configurations in Table 7.

Table 7: Detailed training hyperparameters and execution configuration. * indecates applied on General Object Detection dataset only.

Hyperparameter	**Setting / Value**
Optimizer	Fused AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$)
Base Learning Rate	1.0×10^{-4}
Weight Decay	1.0×10^{-4}
Learning Rate Schedule	1k step linear warmup, step-wise $0.1\times$ decay at epochs 24 and 27
Total Training Epochs	30 epochs
Per-GPU Batch Size	32 images
Effective Batch Size	256 images (across 8 GPUs)
Input Image Dimension	640×640 (Letterbox square padded)
Data Augmentation*	Cached Mosaic, Random Perspective, Color Jitter, Horizontal Flip
Unfolding Stages (T_{iter})	3 stages ($t = 0, 1$ updates + stage 2 energy readout)
Initial Temperature (τ_0, τ_1)	Initialized to $\tau = 1.0$ via $\mathrm{Softplus}(\theta_\tau) + 0.01$
Initial Interaction Step ($\lambda_0, \lambda_1, \lambda_2$)	Initialized to $\lambda = 0.5$ via $\mathrm{Softplus}(\theta_\lambda)$
Regularization Constant (ϵ)	10^{-4} for μ_ϵ positive definiteness
Readout Scale & Bias (s, b)	Initialized to $s = 1.0$ and $b = 0.0$
Gradient Clipping	Maximum ℓ_2-norm of 10.0
Numerical Precision	Automatic Mixed Precision (AMP FP16) with gradient scaling
Hardware Platform	$8\times$ NVIDIA RTX PRO 6000 Blackwell (96 GB VRAM)
Host Configuration	AMD EPYC 9654 96-Core Processor, 1.5 TB RAM
Deep Learning Stack	PyTorch 2.6.0, CUDA 12.8, cuDNN 9.2

D.6 Inference and Post-Processing Protocol

During test-time evaluation, no test-time augmentation (TTA) or morphological post-processing is applied:

1. Input images are resized to 640×640 using letterboxing with grey padding $(114/255)$.
2. The target query t_1 and the additional textual queries t_m ($m \geq 2$) are retrieved from the precomputed text cache, or encoded on-the-fly when a novel query is encountered.
3. The unrolled recurrence executes $T_{\text{iter}} = 3$ forward operations (Eqs. 10–13).
4. The resulting differential logit ℓ_m is used as a spatial prior for text-conditioned prototype routing. The ProtoNet-generated prototypes are modulated by TextFiLM and combined through prototype routing to produce the final mask logit L_{mask}, which is upsampled to the native image resolution $H \times W$ after removing the letterbox padding.
5. Binary mask decisions are obtained by zero-thresholding the final mask logit:

$$\hat{Y}_{ij} = \mathbb{I}\left(L_{\text{mask},ij} > 0\right) \iff \mathbb{I}\left(\sigma(L_{\text{mask},ij}) > 0.5\right). \tag{82}$$

D.7 Streamlined core module workflow

Algorithm 1 presents a simplified pseudocode description of LUMIN; while the specification of certain hyperparameters is omitted, it effectively illustrates the algorithm's workflow.

E More Visualization Results

E.1 Open-Vocabulary Grounding Across Diverse Scenarios

Figures 4 and 5 further demonstrate the open-vocabulary generalization of LUMIN across cluttered indoor environments, complex industrial scenes, and dynamic activities. As illustrated in Figure 4, our unfolded variational framework seamlessly accommodates multi-granularity state spaces without mutual interference: it simultaneously parses scene-level stuff (*sea*, *beach*) and part-level clothing components without boundary leakage (Figure 4a), resolves subtle relative spatial positions and postures (Figure 4b), and accurately grounds compositional prepositional phrases across cluttered surfaces (Figure 4c). Crucially, the text-induced compatibility kernel readily captures fine-grained

Algorithm 1 LUMIN: Language-Conditioned Mean-Field Unfolding

Require: Visual features $\mathbf{E}$, textual queries $\mathbf{T}$, number of stages T
Ensure: Output logits $\mathbf{L}_{\text{out}}$

$\mathbf{U} \leftarrow \text{Unary}(\mathbf{E}, \mathbf{T}), \quad \mu \leftarrow \begin{bmatrix} \text{Softplus}(\theta_{bg}) & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{T}\mathbf{T}^\top \end{bmatrix} + \epsilon\mathbf{I}$
$\mathbf{W} \leftarrow \text{Affinity}(\mathbf{E}; \sigma), \quad \mathbf{Q}^{(0)} \leftarrow \text{Softmax}(\mathbf{U})$
for $t = 0, \ldots, T-2$ **do**
    $\lambda_t \leftarrow \text{Softplus}(\rho_{\lambda_t})$, $\tau_t \leftarrow \text{Softplus}(\rho_{\tau_t}) + \epsilon_\tau$
    $\mathbf{Q}^{(t+1)} \leftarrow \text{Softmax}\Big(\frac{\mathbf{U} + \lambda_t \mathbf{W}\mathbf{Q}^{(t)}\mu}{\tau_t}\Big)$
end for
$\mathbf{L} \leftarrow \mathbf{U} + \lambda_{T-1}\mathbf{W}\mathbf{Q}^{(T-1)}\mu$
$\mathbf{L}_{\text{out}} \leftarrow \exp(\gamma)\big(\mathbf{L}_{1:M} - \mathbf{L}_0\big) + b$
return $\mathbf{L}_{\text{out}}$

physical object states (Figure 4d), isolating the *toilet that is shut* while rejecting the visually identical open bidet nearby.

Furthermore, Figure 5 highlights the precision of our differential energy readout under severe occlusion, scale variation, and rapid motion. In industrial and urban environments, LUMIN easily differentiates co-occurring instances of the same category via fine-grained action semantics (such as *person welding* in Figure 5b) and successfully resolves dozens of highly repetitive, miniature items within dense displays while distinguishing physical humans from two-dimensional poster depictions (Figure 5c). Finally, under acrobatic body deformations and challenging low-light indoor conditions (Figure 5d), the model delineates the sharp airborne contour of the skateboarder without diffusing mass into the ambient background, confirming the structural stability of our continuous variational relaxation.

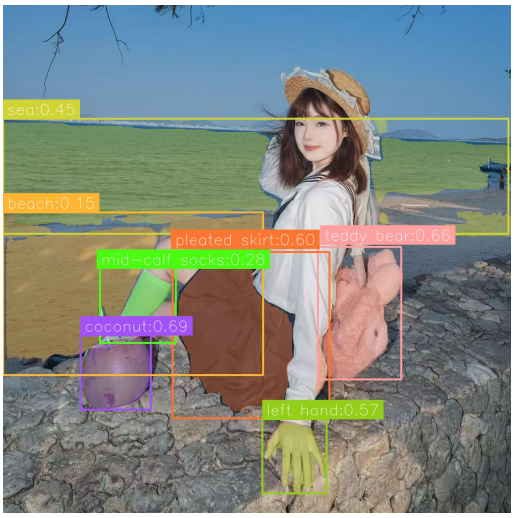
E.2 Internal Unfolding Dynamics and State Evolution

Figures 6b and 6a trace the internal state progression: the unary field U, unary prior $\mathbf{Q}^{(0)}$, probability updates ($\Delta\mathbf{Q}^{(1)}, \Delta\mathbf{Q}^{(2)}$), and final energy F The signed increments $\Delta\mathbf{Q}^{(t)} = \mathbf{Q}^{(t)} - \mathbf{Q}^{(t-1)}$ illustrate directional probability transport on the simplex: the first step ($\Delta\mathbf{Q}^{(1)}$) resolves coarse relational ambiguities by actively shifting probability mass away from distractors to the target, while the second step ($\Delta\mathbf{Q}^{(2)}$) rapidly stabilizes into subtle boundary refinements. This behavior visually validates the fast convergence of our unrolled recurrence within only $T_{\text{iter}} = 3$ stages. Finally, the unnormalized energy field F restores the continuous dynamic range flattened by simplex normalization, providing an uncorrupted, high-contrast substrate for the subsequent differential readout

F Limitations and Future Directions

LUMIN is primarily intended to validate a language-conditioned posterior inference paradigm rather than to exhaustively optimize every component of a state-of-the-art RES system. Several limitations remain. First, the current implementation adopts frozen vision-language representations and a lightweight prototype-based mask realization; therefore, its segmentation quality is still bounded by the underlying visual-language alignment and mask realization capacity. Second, the unfolded inference is deliberately compact, using only three stages and a small number of stage-wise learnable parameters. More expressive unfolding strategies, such as learnable proximal transformations, adaptive preconditioning, or richer stage-wise parameterization, may provide stronger posterior refinement. Third, the current empirical study focuses on RES, while the proposed query-conditioned inference formulation is not intrinsically tied to this task. Extending the same framework to other open-vocabulary dense grounding tasks, such as open-vocabulary detection, is an interesting direction for future work.

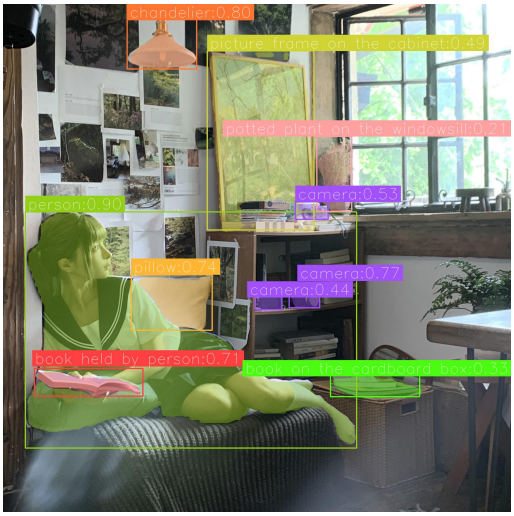
Finally, our current evaluation emphasizes zero-shot transfer and the inference formulation itself, rather than a strictly protocol-matched comparison with fully supervised RES systems. We therefore view the present results as an initial validation of the proposed inference perspective. Future work



(a) coconut. sea. teddy bear. beach. mid-calf socks. left hand. pleated skirt.



(b) person standing on the ladder. the left person.



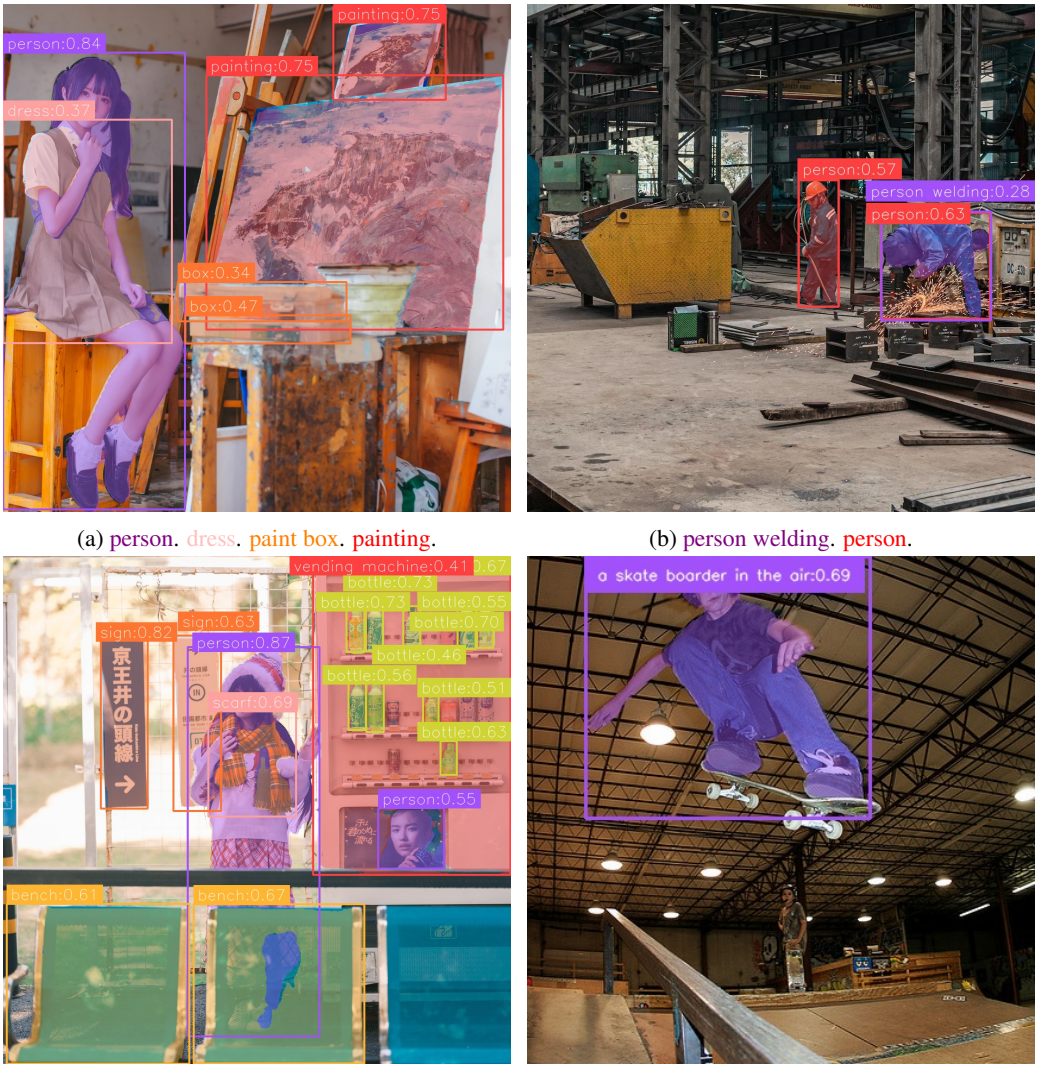
(c) camera. chandelier. picture frame on the cabinet. potter plant on the windowsill. pillow. book on the cardboard box. book beld by person.



(d) toilet that is shut.

Figure 4: **Qualitative results on dense parsing, spatial relations, and physical states.** LUMIN robustly handles: (a) simultaneous parsing of amorphous backgrounds and fine-grained body parts; (b) instance disambiguation via action and relative position; (c) dense compositional grounding with multiple spatial prepositions (*on*, *held by*); and (d) state-dependent visual reasoning (*toilet that is shut*). Query-matched predictions share identical colors.

will investigate stronger vision-language alignment, more expressive unfolding architectures, and broader task instantiations while preserving the optimization-grounded and interpretable structure of the framework.



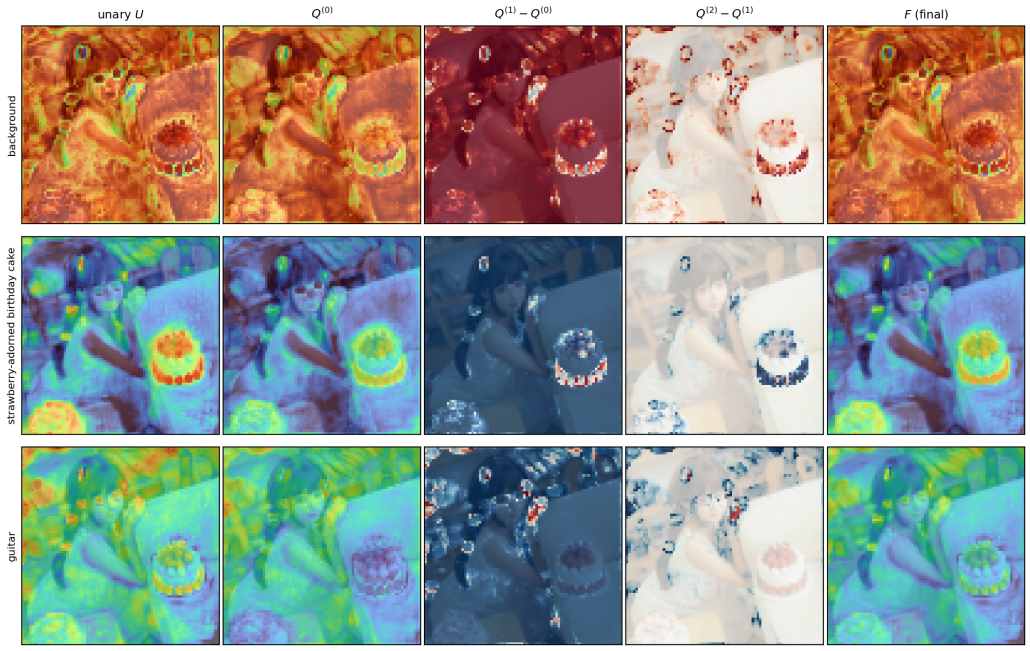
(a) person. dress. paint box. painting.

(b) person welding. person.

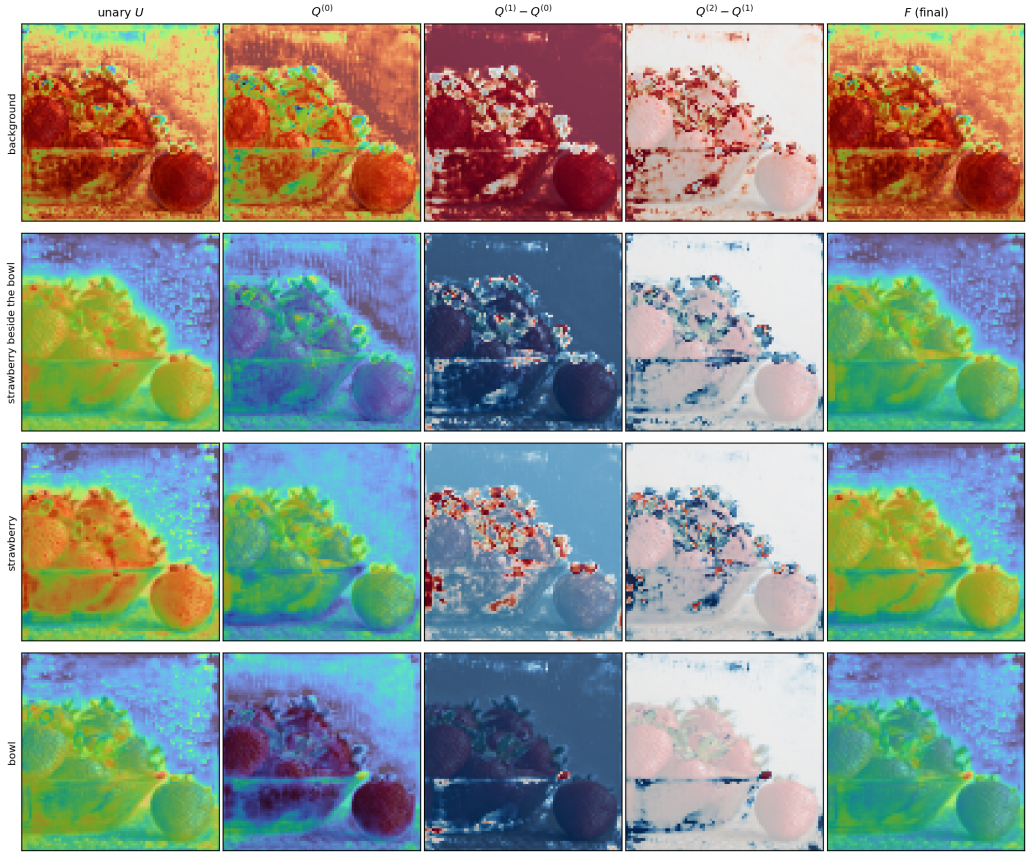
(c) person. bench scarf. sign. bottle. vending machine.

(d) a skate boader in the air.

Figure 5: **Qualitative results under dense clutter, action reasoning, and dynamic poses.** LUMIN demonstrates high fidelity in: (a) artistic scene decomposition; (b) action-driven instance differentiation in industrial clutter (*welding*); (c) extreme multi-instance small-object segmentation (dense bottles) alongside poster-vs-person disambiguation; and (d) non-rigid boundary delineation under airborne athletic poses. Matching colors denote corresponding textual queries.



(a)



(b)

Figure 6: **Visualization of layer-wise mean-field unfolding dynamics.**