CTMR: Cohort-Aware Transformer Multi-Objective Ranker for Personalized Debiased and Diversity-Aware Product Search

Presenter: Liping Zhang
Staff Lead ML Scientist, Adobe Inc.
ICDM ISIR-eCom Workshop 11/12/2025

Authors: Liping Zhang, Subhajit Sanyal, Tracy King

Motivation

- CTR–CVR tradeoff in creative marketplaces
- Fairness across cohorts & long-tail queries
- Need for scalable, responsible re-ranking

CTMR Key Contributions

- Field-Aware Positional Transformer Encoders (FAPTE)
- Task-specific Late-Interaction MaxSim features
- Cohort-Conditioned HyperNetworks (uncertainty-aware)
- IPS debiasing + exposure-diversity regularization
- Unified CTR/CVR multi-task objective

CTMR Architecture (High Level)

- Semantic alignment: Field-aware positional encoders and Late-Interaction features capture token-level semantics with efficiency.
- Cohort personalization: HyperNetworks conditioned on cohort embeddings dynamically generate task mixing and expert routing.
- Responsible optimization: Inverse-propensity weighting debiases training, and a diversity regularizer prevents long-tail neglect.
- *Complete Training Objective: The final training objective combines all loss components with adaptive weighting.*

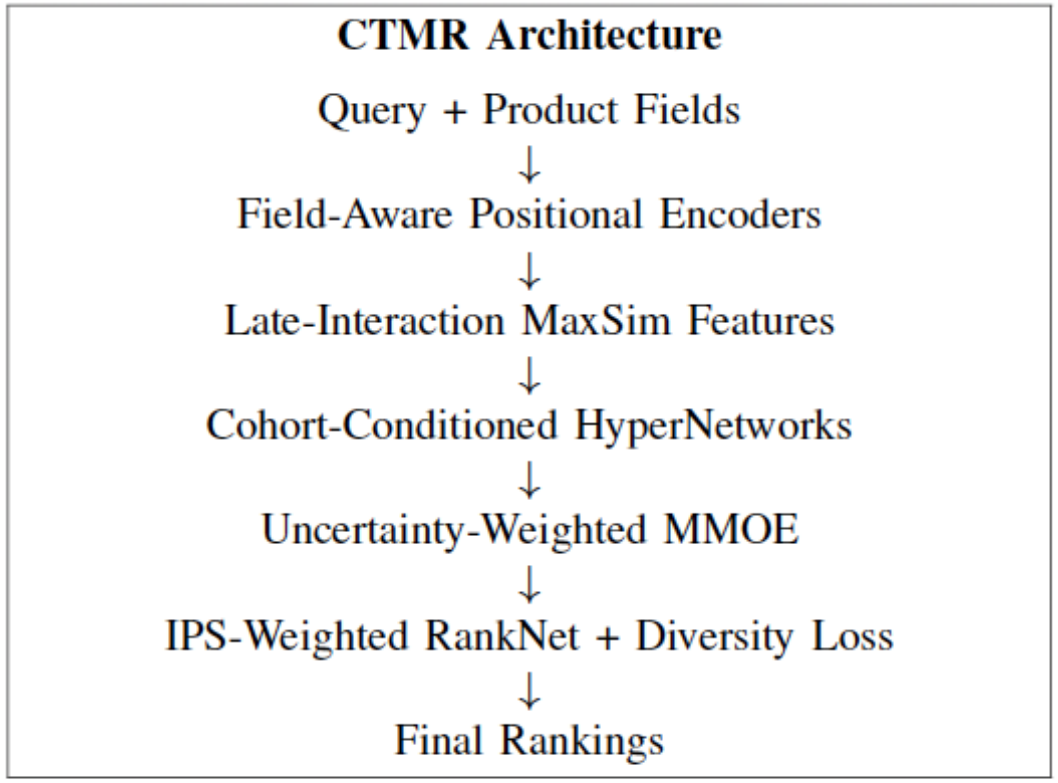


Fig. 1. Overview of CTMR architecture showing the integrated pipeline from field-aware encoding to bias-corrected ranking optimization.

$$\mathcal{L}_{total} = \mathcal{L}_{MTL} + \lambda_{rank}\mathcal{L}_{pair}^{IPS} + \lambda_{div}\mathcal{L}_{div} + \lambda_{reg}\mathcal{L}_{reg} \tag{19}$$

where $\mathcal{L}_{reg}$ represents L2 regularization terms, and the weighting coefficients are learned through hyperparameter optimization.

Field-Aware Positional Transformer Encoders (FAPTE)

- Multi-field tokenization
- Hierarchical positional encodings
- Field-constrained attention

For each token $t_{i,f}$, we construct a composite embedding that integrates three complementary representations:

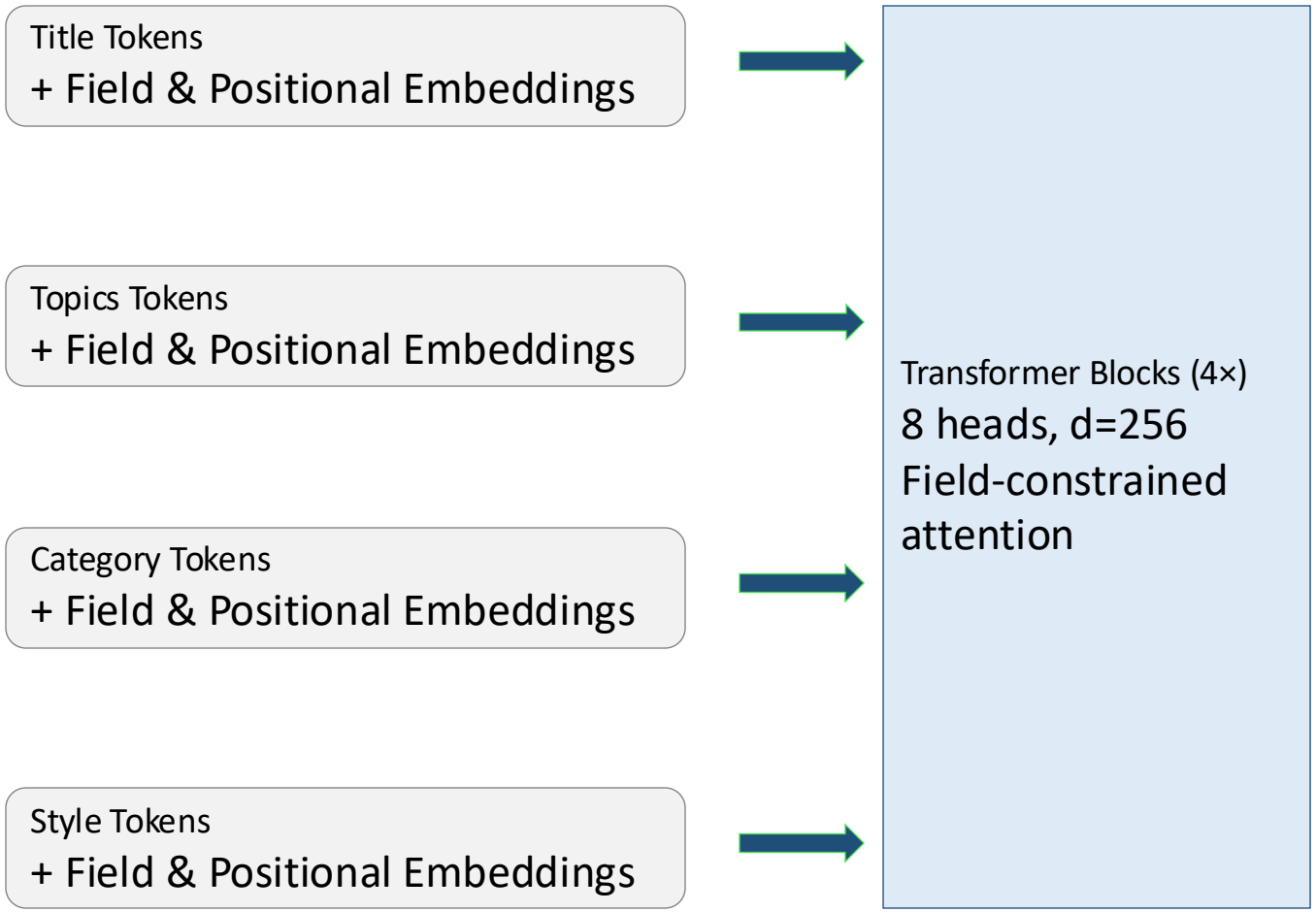
$$\mathbf{e}_{i,f} = \mathbf{W}_{\text{token}} \cdot \mathbf{v}_{t_i} + \mathbf{W}_{\text{field}} \cdot \mathbf{f}_f + \mathbf{W}_{\text{pos}} \cdot \mathbf{p}_i \quad (1)$$

where $\mathbf{v}_{t_i} \in \mathbb{R}^d$ represents the pre-trained token embedding, $\mathbf{f}_f \in \mathbb{R}^d$ denotes the learnable field-type embedding for field f, and $\mathbf{p}_i \in \mathbb{R}^d$ captures positional information using sinusoidal encodings modified for field-aware contexts.

$$\mathbf{p}_i = \alpha \cdot \text{PE}_{global}(i) + \beta \cdot \text{PE}_{field}(i_{local}) + \gamma \cdot \text{PE}_{cross}(f) \quad (2)$$

where $\text{PE}_{global}(i)$ provides absolute positional information across the entire sequence, $\text{PE}_{field}(i_{local})$ encodes the relative position within the current field, and $\text{PE}_{cross}(f)$ captures inter-field relationships. The weighting parameters α, β, and γ are learned during training to optimize field-aware attention patterns.

Field-Aware Positional Transformer Encoders (FAPTE)



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + \mathbf{M}_{field}}{\sqrt{d_k}}\right)V \quad (3)$$

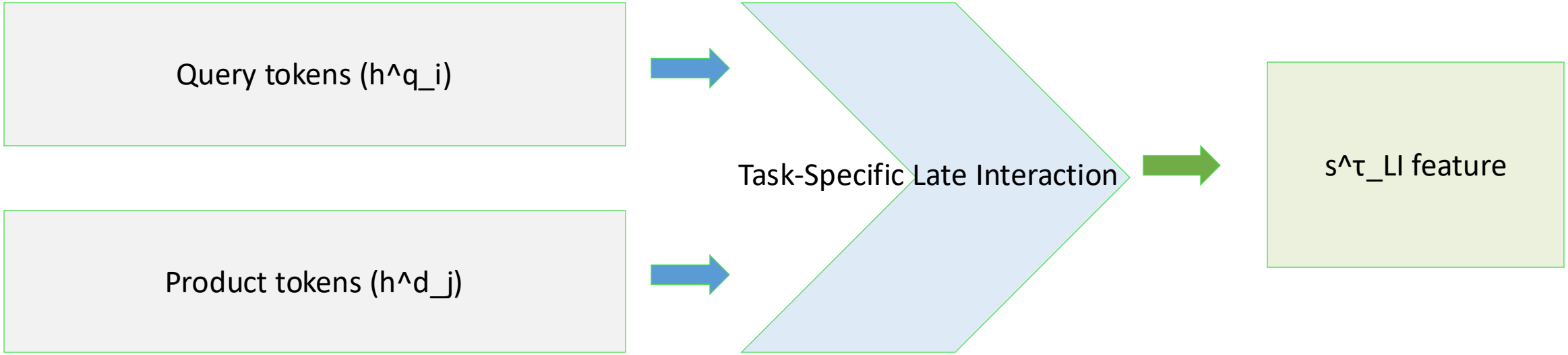
where $\mathbf{M}_{field}$ is a learnable field-relationship matrix that modulates attention weights based on field compatibility patterns observed in Adobe creative content product data.

Late Interaction MaxSim (Task-Specific)

- Token-level semantic matching
- Task-specific projections
- Multi-granularity features

$$s_{LI}^{\tau} = \frac{1}{|q|} \sum_{i=1}^{|q|} \max_{j=1}^{|d|} \langle \mathbf{W}^{\tau} \mathbf{h}_i^q, \mathbf{W}^{\tau} \mathbf{h}_j^d \rangle \quad (4)$$

where $\mathbf{W}^{\tau} \in \mathbb{R}^{d_{model} \times d_{model}}$ represents task-specific projection matrices learned jointly with the multi-task objectives, and $\mathbf{h}_i^q, \mathbf{h}_j^d$ are contextualized token representations from our FAPTE encoders.



Cohort-Conditioned HyperNetworks

- Cohort embedding (locale, temporal, segment, market)
- Dynamic routing & task mixing
- HyperNet generates gates & task weights with uncertainty

$$\mathbf{c} = \text{Concat}[\mathbf{c}_{locale}, \mathbf{c}_{temporal}, \mathbf{c}_{segment}, \mathbf{c}_{market}] \quad (6)$$

where each component captures specific aspects of the search context:

- $\mathbf{c}_{locale}$: Geographic and linguistic context
- $\mathbf{c}_{temporal}$: Seasonal and time-of-day patterns
- $\mathbf{c}_{segment}$: User behavioral segmentation
- $\mathbf{c}_{market}$: Market-specific characteristics

- Cohort-Conditioned HyperNetworks (CCH)
  - dynamically generate expert gating parameters and task balancing coefficients based on contextual signals while incorporating uncertainty quantification for robust decision-ma

$$\theta^{\mu}_{gate}, \theta^{\sigma}_{gate} = \text{HyperNet}_{gate}(\mathbf{c}) \quad (7)$$

$$\theta^{\mu}_{balance}, \theta^{\sigma}_{balance} = \text{HyperNet}_{balance}(\mathbf{c}) \quad (8)$$

where θ^{μ} and θ^{σ} represent the mean and variance of the generated parameters, enabling uncertainty-aware adaptation.

$$\mathbf{g}_e = \text{softmax}(\theta^{\mu}_{gate} + \epsilon \odot \theta^{\sigma}_{gate}) \quad (9)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ introduces controlled stochasticity, and $\mathbf{g}_e$ represents the gating weights for expert e.

Debiasing & Diversity Optimization

- IPS-weighted pairwise RankNet
- Exposure Gini regularization
- Improved long-tail fairness

$$\hat{p}(r) = \frac{1}{r + k} \tag{11}$$

where r represents the ranking position, $k = 2.0$ controls the decay rate, and $\hat{p}(r)$ estimates the probability of examination at position r. This model is calibrated using historical click-through data across different query types and user segments.

$$\mathcal{L}_{pair}^{IPS} = \sum_{(i,j)} w_{ij} \cdot \ell(s_i - s_j, y_{ij}) \tag{12}$$

where the IPS weight is computed as:

$$w_{ij} = \frac{1}{\hat{p}_i(1 - \hat{p}_j)} \tag{13}$$

and $\ell(\cdot)$ represents the pairwise ranking loss function. We employ a robust focal loss variant to handle hard examples:

$$\ell(s_i - s_j, y_{ij}) = -(1 - \sigma(s_i - s_j))^{\gamma} \log \sigma(s_i - s_j) \tag{14}$$

where $\gamma = 2.0$ focuses learning on difficult ranking pairs.

$$\mathcal{L}_{div} = \lambda_{div} \cdot \text{Gini}(\text{Exposure}(B)) \tag{15}$$

where $\text{Exposure}(B)$ represents the exposure values for documents in batch B, computed as:

$$\text{Exposure}(d) = \frac{1}{\log(rank(d) + 2)} \tag{16}$$

The Gini coefficient is computed using the standard formula:

$$\text{Gini}(\mathbf{x}) = \frac{2\sum_{i=1}^{n} i \cdot x_{(i)}}{n\sum_{i=1}^{n} x_{(i)}} - \frac{n+1}{n} \tag{17}$$

where $x_{(i)}$ represents the i-th smallest value in the sorted exposure vector.

Unified Multi-Objective Optimization

- Uncertainty-weighted CTR/CVR losses
- RankNet + diversity + regularization
- Efficient gradient scheduling

$$\mathcal{L}_{MTL} = \sum_{\tau \in \{CTR, CVR\}} \frac{1}{2\sigma_\tau^2} \mathcal{L}_\tau + \log \sigma_\tau \tag{18}$$

where σ_τ represents the learned uncertainty parameter for task τ, and $\mathcal{L}_\tau$ denotes the task-specific loss (binary cross-entropy for both CTR and CVR prediction).

$$\begin{aligned} \mathcal{L}_{total} = \mathcal{L}_{MTL} + \lambda_{rank} \mathcal{L}_{pair}^{IPS} \\ + \lambda_{div} \mathcal{L}_{div} + \lambda_{reg} \mathcal{L}_{reg} \end{aligned} \tag{19}$$

where $\mathcal{L}_{reg}$ represents L2 regularization terms, and the weighting coefficients are learned through hyperparameter optimization.

$$\mathrm{lr}_\tau(t) = \mathrm{lr}_{base} \cdot \frac{\sqrt{1 + \gamma \cdot \|\nabla \mathcal{L}_\tau\|_2}}{1 + \delta \cdot t} \tag{20}$$

where γ and δ control the adaptation rate and decay schedule respectively.

Evaluation

- Adobe Creative Marketplace dataset (4.2M samples)
- Multi-modal features: text, image, KG, behavior
- Strong baselines: MMoE, UWM3R
- Dataset & Features
  - 4.2M query–product interactions; chronological split 70/15/15
  - Text (query/title/topics/style), CLIP embeddings, KG
  - Context (locale/time), behavior, session signals

TABLE I
COMPREHENSIVE FEATURE CATEGORIES IN CTMR SYSTEM. EXTENSIVE FEATURE ENGINEERING IS APPLIED ACROSS MULTIPLE MODALITIES FOR ENHANCED REPRESENTATION LEARNING.

User	Item	Context	Cross
Language	Style	Timestamp	Language × Country
Country	Title	Locale	Language × Region
Region	Mood	Page	Language × Style
User Segment	Creative Intents	Session ID	Region × Country
Behavior History	Topics	Page Context	Region × Category
...	...	...	...

Main Result

- CTMR achieves best AUC, NDCG, MAP, MRR
- Significant gains on long-tail queries

TABLE II
CTMR MODEL COMPARISON ON ADOBE CREATIVE CONTENT DESIGN PLATFORM DATASET. *Tuning:* GRIDS, SEEDS, AND EARLY STOPPING

Model	AUC	LogLoss	Params
MMoE	0.8163	0.1408	75,466
UWM3R	0.9231	0.0827	279,990
CTMR	0.9958	0.016	2,567,588

TABLE III
ABLATION STUDY RESULTS. EACH ROW ADDS ONE COMPONENT TO THE BASE ARCHITECTURE.

Configuration	NDCG	MAP	MMR
Base (MMoE only)	0.661	0.542	0.612
+ Field-Aware Transformers	0.683	0.551	0.623
+ Late-Interaction MaxSim	0.701	0.564	0.638
+ Cohort HyperNetworks	0.718	0.578	0.651
+ IPS Weighting	0.721	0.591	0.669
+ Diversity Regularization	0.729	0.606	0.681

Production Deployment

- Adobe CPF re-ranking microservices
- TorchScript inference
- Caching
- monitoring & A/B testing

Conclusion

- Responsible ranking: relevance + fairness + personalization
- Practical for production-scale search
- Pathway to LLM-enhanced retrieval & conversational search
- Generality beyond creative content, more domains

Paper ID: **S10203**