PROGRESSIVE LEARNING FROM CONTENT TO RELEVANCE TO PERSONALIZATION FOR CTR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal click-through rate (CTR) prediction is central to search systems, where ranking depends on item content, query relevance, and user preference. Existing approaches incorporate content through dense embeddings or coarse discrete IDs, yet effectively integrating fine-grained heterogeneous content with ID-based predictors remains challenging. We propose FineSID, a progressive learning framework that unifies content tokenization, query–item semantic pretraining, and personalized post-training via a shared fine-grained discrete representation. Each stage optimizes an explicit objective while preserving token-level access to content. First, a pretrained multi-scale visual tokenizer and text tokenizers convert heterogeneous content into fine-grained Semantic IDs (SIDs), explicitly aligning them with the ID features. Second, full-parameter semantic pretraining under query–item relevance supervision establishes a robust matching prior, jointly optimizing the interactions of multimodal SIDs and conventional fields within a shared predictor. Third, this initialized predictor is directly fine-tuned with user click supervision to capture implicit personalized preferences. Extensive offline and online evaluations demonstrate substantial gains in both standard CTR prediction and core business metrics. Furthermore, FineSID improves within-product creative discrimination, demonstrating its semantic generalization beyond conventional ID memorization.

1 INTRODUCTION

In billion-scale industrial search platforms like Taobao, accurate multimodal click-through rate (CTR) prediction is critical for both user experience and revenue. The system must simultaneously model three progressive aspects: what an item contains (*Content*), how it matches the query (*Relevance*), and whether it appeals to the user (*Personalization*). While industrial CTR predictors heavily rely on sparse identity (ID) features (Gui et al., 2023; Zhu et al., 2025; Jiang et al., 2026), pure ID-based models struggle with cold-start items and fail to directly leverage the multimodal semantic knowledge available from modern Vision-Language Models (VLMs).

Integrating multimodal content into ID-based models remains a persistent challenge. One-stage joint training and two-stage continuous embedding fusion suffer from computation overhead and face distribution disparities between dense continuous features and the sparse ID space. Recently, Semantic IDs (SIDs) (Yu et al., 2022; Singh et al., 2024) have been proposed to bridge this gap by quantizing global content embeddings into discrete codes. However, standard quantization schemes (e.g., Residual-Quantized Variational AutoEncoders or RQ-VAE) compress information aggressively, discarding spatial structures and textual details necessary for fine-grained semantic matching.

To address this, we propose FineSID, a progressive learning framework that exposes fine-grained multimodal content to a shared ID-based CTR predictor. FineSID structures learning across an explicit *Item* → *Item–Query* → *Item–Query–User* progression:

- **Stage I: Fine-grained Content Tokenization (*Item*).** Instead of relying on global compression, we preserve structural and semantic details by maintaining independent, field-reserved representations for different modalities. We employ a pretrained multi-scale visual tokenizer (Tian et al., 2024) alongside source-specific text tokenizers. These extract

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.

multi-scale SIDs that map visual and textual details into the categorical ID space, ensuring the predictor retains token-level access to multimodal content.

- **Stage II: Semantic Pretraining (*Item–Query*).** The shared predictor is trained under query–item relevance supervision. Visual SIDs, textual SIDs, and conventional fields interact in a common prediction space. By retaining fine-grained tokens rather than a single global embedding, the interaction network is guided to learn explicit cross-modal semantic alignment rather than simply memorizing global representations.
- **Stage III: Personalized Post-training (*Item–Query–User*).** The initialized shared predictor is then fine-tuned end-to-end with user click supervision. This stage adapts the pretrained semantic interactions to incorporate personalized user behaviors and context.

We evaluate FineSID on the industrial Taobao platform and the public MMCTR dataset. Offline results on Taobao show that FineSID yields consistent AUC and GAUC improvements over the evaluated baselines, with notable gains on cold-start items. Furthermore, we evaluate within-product creative discrimination using the NSCTR metric (Yang et al., 2024b). By comparing creatives sharing the exact same product ID, the results indicate that FineSID captures fine-grained content differences independent of conventional ID memorization. A completed online A/B experiment also reports positive changes in both user clicks and Gross Merchandise Value (GMV).

In summary, our key contributions are:

1. We propose FineSID, a three-stage progressive learning framework that integrates multimodal content into ID-based CTR prediction via a discrete interface.
2. We utilize multi-scale, fine-grained Semantic IDs to represent multimodal content. This approach aims to minimize information loss through reconstruction-oriented visual tokenization and source-specific text tokenization, helping the predictor capture spatial and textual details for explicit semantic alignment.
3. We report offline comparisons, ablations, and online A/B outcomes. The results show consistent improvements in standard product-level CTR prediction, cold-start scenarios, and within-product creative discrimination.

2 Problem Formulation

Let an exposure be defined as $x = (u, q, i, c)$, where u denotes user and context features, q a search query, i an item ID, and c its fine-grained multimodal content (e.g., displayed image and textual fields). Notably, different content (creatives) may belong to the exact same item.

Progressive Prediction Tasks. Our ultimate objective is to predict the user click probability:

$$p_{\theta,\psi_{\text{ctr}}}(x) = \Pr(y_{\text{ctr}} = 1 \mid u, q, i, c), \qquad y_{\text{ctr}} \in \{0, 1\}. \tag{1}$$

θ denotes shared feature-interaction parameters, and ψ_{ctr} is the task-specific CTR output head.

Instead of training from scratch, FineSID progressively learns θ. We first pretrain the shared parameters θ under query–item relevance supervision y_{rel} :

$$\pi_{\theta,\psi_{\text{rel}}}(x) = \Pr(y_{\text{rel}} \mid u, q, i, c), \tag{2}$$

where ψ_{rel} is a dedicated relevance head. After establishing this semantic prior, the input interface and the shared network θ are entirely retained, and the model is fine-tuned on the personalized CTR task via click supervision.

Evaluation Scenarios. We evaluate the performance of the predictor under two distinct scenarios:

- **Product-level CTR prediction:** Candidate exposures span distinct item IDs, testing overall ranking capacity.
- **Within-product creative discrimination:** The item ID i is fixed, and candidates vary only in their displayed content c.

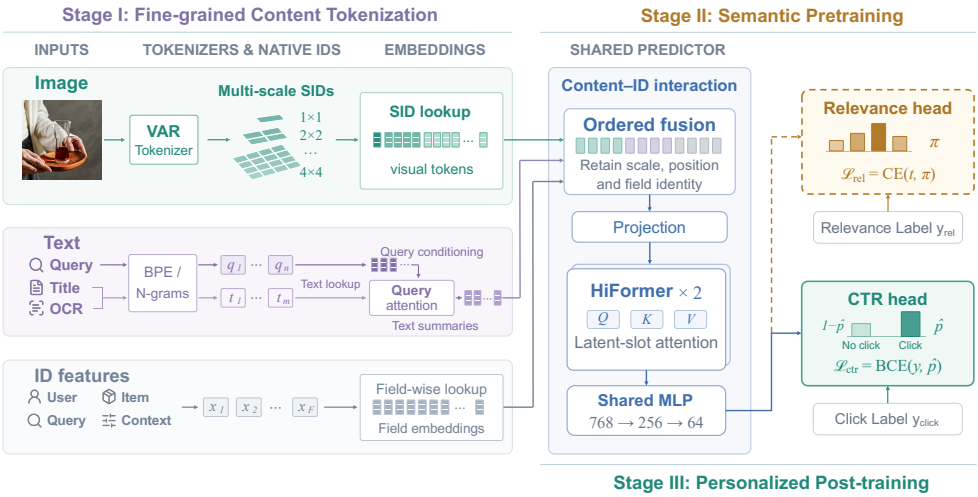

Figure 1: The overall architecture of the FineSID framework. The progressive learning process is decoupled into three stages. **Stage I: Fine-grained Content Tokenization** extracts fine-grained discrete representations, where images are converted into multi-scale visual tokens via a VAR tokenizer, and text fields are tokenized into discrete text IDs and mapped to embeddings through lookup tables. **Stage II: Semantic Pretraining** trains the entire shared predictor (including ordered fusion, latent-slot attention via HiFormer, and the shared MLP) using query–item relevance labels to establish a strong semantic matching prior. **Stage III: Personalized Post-training** fine-tunes the relevance-initialized shared predictor under user click supervision for final CTR prediction. For simplicity, y_{rel} denotes the teacher-provided soft relevance target $t(x)$, $y_{\text{click}} = y$, and π and $\hat{p}$ denote the model's relevance distribution and click probability, respectively.

3 FineSID

As illustrated in Figure 1, FineSID is a progressive learning framework designed to integrate multimodal content into ID-based CTR prediction. The overall pipeline is decoupled into three progressive stages: (1) fine-grained content tokenization (Stage I), (2) full-parameter semantic pretraining (Stage II), and (3) personalized CTR post-training (Stage III). At the core of Stage II and Stage III lies a shared feature interaction architecture (θ) that captures complex cross-modal relations.

3.1 Stage I: Fine-Grained Content Tokenization

In ID-based CTR predictors, continuous multimodal embeddings suffer from distributional misalignment and are often relegated to late fusion or auxiliary similarities. To seamlessly align multimodal content with the sparse ID space, FineSID discretizes it into Semantic IDs (SIDs), sharing the exact same lookup interface as categorical IDs for native feature interactions. Furthermore, our tokenization aims to limit compression-induced information loss by retaining the multi-scale spatial and structural details necessary for fine-grained ranking.

Multi-scale Visual Tokenization. Specifically, we employ the pretrained multi-scale visual tokenizer of VAR (Tian et al., 2024) to encode images into discrete visual SIDs. Its reconstruction-based training encourages high-fidelity preservation of visual content, supplying a spatially aware discrete interface before our downstream relevance and click objectives are applied. For an observed image I, we retain four native code maps across progressively finer scales:

$$S_\ell^v \in \{0, \dots, V-1\}^{h_\ell \times w_\ell}, \qquad (h_\ell, w_\ell) \in \{(1,1), (2,2), (3,3), (4,4)\}, \tag{3}$$

where V is the visual codebook size. This yields 30 spatially indexed SIDs $(1+4+9+16)$ that capture coarse-to-fine visual semantics. Because these discrete codes are extracted offline, deploying

them provides deep semantic evidence without incurring the heavy latency of running an autoregressive image generator during click prediction.

During both Stage II and Stage III, the shared predictor maps each of these fixed visual SIDs to a task-trainable embedding:

$$e^v_{\ell,p} = m_{\ell,p} E_v[S^v_{\ell,p}], \qquad z_v = [\mathrm{vec}(e^v_1); \mathrm{vec}(e^v_2); \mathrm{vec}(e^v_3); \mathrm{vec}(e^v_4)], \tag{4}$$

where p indexes the spatial position within scale ℓ, $m_{\ell,p}$ acts as a validity mask for missing codes, and E_v is the randomly initialized, task-specific lookup table in our shared predictor (distinct from the tokenizer's original reconstruction codebook). The $\mathrm{vec}(\cdot)$ operation flattens the embeddings in a fixed spatial order, preserving both the scale and spatial position slots before they enter the feature interaction layers. Code zero is a valid ID and does not itself indicate padding.

Textual Tokenization. To complement the visual SIDs, we process textual content (e.g., search query, item title, and OCR texts) using Byte Pair Encoding (BPE) and n-gram tokenizers. Instead of compressing heterogeneous text into a single summary vector, we maintain independent, *field-reserved* token sequences. Each sequence preserves its source identity and term position, exposing textual details through discrete lookups E_t, just as visual SIDs expose spatial image content. By maintaining separate vocabularies and fine-grained discrete interfaces for both modalities, Stage I ensures the shared predictor has comprehensive, token-level access to the multimodal content.

3.2 Shared Feature Interaction Architecture (θ)

To effectively map the fine-grained multimodal SIDs and conventional ID features into a common semantic space, we deploy a shared feature interaction network (parameterized by θ). This architecture is utilized in both Stage II and Stage III.

Ordered Fusion. For textual content, we apply cross-attention with search-query embeddings as queries (Q) and each text field's token embeddings as keys and values (K, V), producing query-conditioned embeddings a_t. These attention outputs are passed to the shared predictor through ordered fusion. We concatenate the conventional base-field embeddings v_f, the multi-scale visual representation z_v, and the query-conditioned text embeddings in a fixed order:

$$z = [v_1; \ldots; v_F; z_v; \{a_t\}_t]. \tag{5}$$

This ordering preserves the scale and spatial slots of the visual codes and the field identities of the query-conditioned text representations before subsequent feature interactions.

Interaction Backbone. The high-dimensional concatenated vector z is first projected into a unified 768-dimensional global state. Since standard self-attention over thousands of raw tokens and ID fields is computationally prohibitive, we deploy a highly efficient latent-slot attention backbone inspired by HiFormer (Gui et al., 2023). This network maps the global state into compact latent slots to capture high-order cross-feature interactions without the quadratic bottleneck.

Finally, a Shared MLP configured with dimensions $768 \to 256 \to 64$ processes the refined global state to output the ultimate shared representation h. This compact representation h is then fed into task-specific heads (ψ_{rel} or ψ_{ctr}) in the subsequent progressive learning stages.

3.3 Stage II: Full-Parameter Semantic Pretraining

Memorization vs. Generalization. User satisfaction in search relies on two distinct dimensions: explicit *query–item semantic relevance* and implicit *personal preference* (e.g., visual appeal or price sensitivity). Conventional ID-based CTR models heavily rely on memorizing historical co-occurrences to capture the latter, but struggle with the former due to ID sparsity. Stage II injects a generalized semantic prior into the model. Crucially, without fine-grained content representation, models tend to devolve into merely memorizing global content embeddings. By preserving token-level details via FineSID, we guide the shared predictor to learn explicit cross-modal semantic alignment rather than resorting to rote memorization.

Relevance Objective. Using the complete exposure input x, we train the shared feature-processing network to judge whether the fine-grained candidate content satisfies the query intent. For each eligible exposure x, a Qwen3.5 teacher model, fine-tuned on million-scale expert annotations, provides a soft relevance distribution $t(x) \in \Delta^3$ over four matching levels. We optimize both the shared predictor parameters θ and a dedicated relevance head ψ_{rel} using a soft-label cross-entropy loss:

$$\mathcal{L}_{\text{rel}} = -\mathbb{E}_{x\sim\mathcal{D}_{\text{rel}}} \sum_{k=1}^{4} t_k(x) \log \pi_{\theta,\psi_{\text{rel}},k}(x). \tag{6}$$

Importantly, relevance gradients update the *entire* shared trainable parameter group θ—including SID embeddings, categorical field lookups, interaction blocks, and normalization parameters. The associated normalization statistics are updated during training. Thus, the complete CTR backbone acquires a robust semantic prior.

3.4 Stage III: Personalized CTR Post-training

Having established a generalizable semantic prior, the model must now learn the implicit, hard-to-articulate user preferences from behavioral logs. The primary advantage of FineSID is its seamless compatibility with existing industrial pipelines. Because multimodal content has already been aligned into the discrete ID space (Stage I) and the feature interactions have been semantically initialized (Stage II), Stage III requires no complex modality-fusion heuristics or contrastive losses.

Click Objective. We initialize the CTR model directly with the pretrained shared parameters θ and its normalization states. We discard the relevance head, attach a randomly initialized CTR output layer ψ_{ctr}, and jointly fine-tune the entire network with personalized user behavior:

$$\mathcal{L}_{\text{ctr}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{ctr}}} \left[y \log p_{\theta,\psi_{\text{ctr}}}(x) + (1-y) \log(1 - p_{\theta,\psi_{\text{ctr}}}(x)) \right], \tag{7}$$

where $y \in \{0, 1\}$ is the observed user click. Unlike standard embedding-only transfer methods that discard the pretrained dense network, FineSID transfers and fine-tunes all shared predictor parameters (full-parameter post-training). This simple, standard binary cross-entropy optimization proves sufficient to adapt the robust semantic interactions to highly personalized CTR predictions.

4 Experiments

We conduct extensive experiments to evaluate the effectiveness of FineSID. We examine whether the proposed three-stage progressive learning framework can simultaneously improve standard product-level CTR prediction and fine-grained within-product creative discrimination.

4.1 Datasets

Industrial Taobao Search. We use a chronological split of raw search impression logs: August 25–31, 2026 for CTR training and September 1, 2026 for offline evaluation. Each row is a candidate-item exposure with a binary click label and user, normalized-query, and item identifiers. We retain exposure rows rather than deduplicating repeated user–query–item interactions. Table 1 summarizes the raw CTR populations; its training counts aggregate all seven days before distinct entities are counted. We count distinct nonempty user IDs, normalized query IDs, and item IDs over each complete split rather than summing daily distinct counts. These statistics characterize the raw CTR data, before any method-specific feature or teacher-label availability conditions; they do not include the separate relevance-supervision corpus.

Table 1: Approximate size of the raw Taobao CTR splits in 2026. Exposure totals are rounded from row counts; User, Query, and Item are estimated distinct nonempty identifiers over the entire split.

Split	Dates	Exposures	Users	Queries	Items
Train	Aug 25–31	51.2B	251M	200M	418M
Eval	Sep 1	453M	30.2M	15.4M	1.40M

4.2 Evaluation Metrics

For product-level prediction, we report Area Under the Curve (AUC) and Group AUC (GAUC). GAUC is calculated by aggregating AUC computed within individual query groups, weighted by impression counts (Zhou et al., 2018): $\text{GAUC} = \sum_g w_g \text{AUC}_g / \sum_g w_g$. Furthermore, to evaluate the performance under data sparsity, we report metrics on a *cold-start subset*, which contains items with fewer than 1,000 page views (PV) on the evaluation day (i.e., $\mathcal{I}_{\text{cold}} = \{i : \text{PV}(i) < 1000\}$). For within-product creative ranking, we utilize the Normalized Simulated CTR (NSCTR) (Yang et al., 2024b), detailed in Section 4.7. For the industrial Taobao dataset, results are expressed as improvements relative to the baseline (percentage points (pt) for AUC/GAUC, and relative percentage (%) for NSCTR). Absolute AUC is reported for the public MMCTR dataset.

4.3 Comparison Methods

We compare FineSID with an ID-based baseline and several state-of-the-art approaches that integrate content semantics or relevance supervision into CTR prediction:

- **HiFormer** (Gui et al., 2023): the foundational baseline using conventional ID features.
- **LEARN-style embedding** (Jia et al., 2025): directly augments the baseline with Qwen3.5 continuous content embeddings.
- **Visual ID** (Yu et al., 2022): quantizes a visual embedding into discrete IDs.
- **RQ-VAE SID** (Singh et al., 2024): uses residual-quantization IDs to compress content.
- **COINS** (Zhao et al., 2026): aligns content and collaborative information through semantic IDs for cold-start items.
- **PRECTR** (Chen et al., 2025): couples personalized search relevance and CTR prediction.
- **Joint relevance + CTR** (Chen et al., 2026): jointly optimizes relevance and user engagement in a multitask ranking model.

4.4 Taobao Offline Evaluation

Table 2: Taobao offline improvements over the HiFormer plus ID features baseline. AUC and GAUC columns report absolute changes in percentage points (pt); NSCTR reports relative improvement (%). 'Cold' denotes items with evaluation-day PV < 1000.

Method	AUC (pt)	GAUC (pt)	Cold AUC (pt)	Cold GAUC (pt)	NSCTR (%)
HiFormer + ID features (Gui et al., 2023)	–	–	–	–	–
LEARN-style embedding (Jia et al., 2025)	+0.05	+0.04	+0.14	+0.09	+0.18
Visual ID (Yu et al., 2022)	+0.14	+0.14	+0.16	+0.21	+0.24
RQ-VAE SID (Singh et al., 2024)	+0.12	+0.15	+0.20	+0.25	+0.22
COINS (Zhao et al., 2026)	+0.16	+0.13	+0.18	+0.28	+0.27
PRECTR (Chen et al., 2025)	+0.10	+0.15	+0.11	+0.16	–
Joint relevance + CTR (Chen et al., 2026)	+0.04	+0.03	-0.04	-0.07	–
FineSID (Ours)	**+0.25**	**+0.26**	**+0.38**	**+0.34**	**+0.75**

As shown in Table 2, FineSID achieves the largest gains across all reported metrics. Its overall AUC and GAUC improvements over the baseline are +0.25 and +0.26 points, respectively. Notably, these gains are amplified on the cold-start subset (+0.38 and +0.34 points), which is consistent with our hypothesis: when behavioral ID evidence is sparse, the predictor successfully leverages the robust semantic prior established by fine-grained content tokenization and relevance pretraining. Furthermore, the 0.75% relative improvement in NSCTR demonstrates the framework's effectiveness in within-product creative discrimination, outperforming the evaluated global-quantization baselines.

4.5 Ablation Study

To understand the contribution of each design component, we conduct ablations in Table 3. Removing the Stage II semantic pretraining reduces both overall and cold-start metrics, while driving the

Table 3: Ablation study on Taobao. Each row reports its gain over the HiFormer baseline.

Configuration	AUC (pt)	GAUC (pt)	Cold AUC (pt)	Cold GAUC (pt)	NSCTR (%)
FineSID (Full)	**+0.25**	**+0.26**	**+0.38**	**+0.34**	**+0.75**
w/o Relevance pretraining (Stage II)	+0.20	+0.21	+0.31	+0.30	+0.51
Pooling FineSID	+0.12	+0.14	+0.22	+0.18	+0.44
w/o Fine-grained SIDs (Stage I)	+0.03	+0.02	-0.03	+0.04	+0.02

NSCTR improvement down from 0.75% to 0.51%. This highlights that end-to-end CTR training alone is prone to missing explicit query–item relevance signals. Furthermore, the *Pooling FineSID* variant, which replaces the query-conditioned cross-attention and multi-scale slots with flat average pooling, yields smaller AUC gains over HiFormer: 0.12 percentage points versus 0.25 for the full model. This supports spatially indexed visual features and query-conditioned text embeddings over flat average pooling. Crucially, entirely removing the fine-grained SIDs (reverting to pure ID features) substantially diminishes the improvements, resulting in a 0.03-percentage-point decrease in cold-start AUC relative to HiFormer and minimal NSCTR improvement (+0.02%). This confirms that fine-grained content tokenization is a critical component for multimodal CTR improvements.

4.6 Public Benchmark: MMCTR

Table 4: Performance on the public MMCTR dataset. Changes (Δ) are absolute AUC differences from the HiFormer baseline, expressed in percentage points.

Method	AUC	ΔAUC (pt)
HiFormer + ID features	0.9767	Reference
LEARN-style embedding	0.9823	+0.56
COINS	0.9825	+0.58
FineSID (w/o relevance pretraining)	**0.9844**	**+0.77**

To validate the generalizability of our framework, we evaluate on the public MMCTR benchmark (Zhu et al., 2021). Because the MMCTR dataset lacks the explicit query-item relevance labels required for Stage II, we specifically evaluate the FineSID variant *without* relevance pretraining. As shown in Table 4, even relying solely on Stage I (fine-grained content tokenization) and Stage III (CTR post-training), FineSID achieves the largest AUC gain (+0.77 pt) among all reported methods, further validating the utility of representing fine-grained SIDs as native categorical features.

4.7 Within-Product Creative Evaluation

Standard product-level CTR evaluation often confounds a model's true semantic understanding with its capacity to memorize historical ID co-occurrences. To isolate the benefit of fine-grained multimodal representations, we evaluate *within-product creative selection*. In this scenario, the item ID is identical across candidates; therefore, conventional ID-based memorization becomes uninformative. The predictor must therefore rely on resolving the fine-grained differences in the multimodal content tokens (e.g., varied images or titles for the exact same product) to accurately rank the creatives.

To quantify this, we utilize the Normalized Simulated CTR (NSCTR) introduced in Peri-CR (Yang et al., 2024b). For each logged exposure, the model scores candidate creatives of the same item and selects the highest-scoring one. A replay match occurs when that selection equals the creative actually displayed. Let N_i be the original exposure count of item i, and let M_i and C_i be the matched exposure and click counts. For items where $M_i > 0$, the NSCTR is defined as:

$$\text{NSCTR} = \frac{\sum_i N_i (C_i/M_i)}{\sum_i N_i}. \tag{8}$$

This metric weights the matched CTR of each item by its original exposure count, effectively evaluating the quality of the selected creative. As reported in Table 2, FineSID achieves a +0.75% relative improvement in NSCTR, more than doubling the improvement of COINS (+0.27%), demonstrating

the framework's ability to generalize its semantic understanding of visual and textual details beyond simple ID memorization.

4.8 Online A/B Testing

Table 5: Seven-day online A/B testing results on Taobao's main search traffic.

Outcome	Relative change vs. Control
Item Page View (IPV / Clicks)	+0.21%
Gross Merchandise Value (GMV)	+0.85%

To verify the practical business value of the proposed framework, we deployed FineSID in the Taobao main search engine and conducted online A/B testing against the production baseline for seven days. As summarized in Table 5, FineSID yielded a +0.21% improvement in Item Page Views (Clicks) and a +0.85% increase in Gross Merchandise Value (GMV). Both improvements passed statistical significance testing. In a mature, billion-scale industrial platform like Taobao, such improvements represent meaningful business impact, demonstrating that the semantic prior learned through our progressive framework translates directly to user engagement and revenue gains.

5 Related Work

Multimodal Content Representations. Prior work incorporates visual content into CTR prediction and recommendation through visual encoders and task-aware representation learning (Ge et al., 2018; Liu et al., 2020; Xv et al., 2022; Yang et al., 2024a). Fine-grained vision–language modeling provides complementary approaches: FashionBERT jointly models image patches and text (Gao et al., 2020), FILIP uses token-level late interaction (Yao et al., 2022), and MM-Rec models relations between image regions and text for news recommendation (Wu et al., 2022). These studies establish the value of content representations and fine-grained interactions across different tasks. Industrial CTR deployment requires balancing content granularity with the computational cost of feature interaction. For FineSID, the relevant design question is how to retain separately accessible content details when constructing inputs to a highly efficient shared CTR predictor.

Semantic IDs as a Content Interface. Visual ID and residual-quantization semantic IDs encode content representations as discrete codes for downstream prediction (Yu et al., 2022; Singh et al., 2024). COURIER also examines cluster IDs as a downstream interface for pretrained visual representations (Yang et al., 2024a). More recent approaches explore content–collaborative alignment for cold-start prediction and prototype-composed semantic identifiers (Zhao et al., 2026; Liu et al., 2026). These approaches motivate the use of discrete content interfaces, while the organization and granularity of the codes remain important design choices. To retain explicit spatial structure, FineSID uses spatially indexed, multi-scale codes from VAR's reconstruction-trained tokenizer (Tian et al., 2024), together with field-specific textual tokens. Its emphasis is on retaining structured content units whose task-trainable embeddings can interact with conventional features during relevance and CTR learning.

Relevance Supervision and CTR Transfer. Prior work combines relevance and behavioral supervision through representation pretraining, coupled prediction modules, and joint objectives. SEMINAR learns reusable multimodal representations with multiple pretraining objectives, including query–item relevance (Shen et al., 2024). PRECTR pretrains a relevance module before joint training with a CTR base model (Chen et al., 2025); PRECTR-V2 further develops a lightweight text encoder using relevance supervision and LLM distillation (Cao et al., 2026). Joint relevance–engagement ranking instead optimizes the two tasks together and combines their outputs for ranking (Chen et al., 2026). FineSID emphasizes the scope of relevance pretraining and transfer: relevance supervision trains the shared SID embeddings, content-interaction layers, and prediction backbone, which are then transferred and fine-tuned for personalized CTR prediction. This connects a fine-grained multimodal input interface with relevance-to-CTR transfer of the shared predictor.

6 Discussion and Limitations

While FineSID demonstrates significant improvements in multimodal CTR prediction and creative discrimination, it presents several limitations that suggest valuable directions for future research:

Towards End-to-End Semantic Tokenization. Currently, the multi-scale visual tokenizer (Stage I) is pretrained for image reconstruction and kept frozen during relevance pretraining (Stage II). While this discrete interface effectively bridges modalities, it is not strictly optimized for search semantics. Future work could explore jointly fine-tuning the visual tokenizer alongside the relevance pretraining task, enabling an end-to-end semantic tokenization process that could yield a higher upper bound for cross-modal matching.

Broader Generalization and LLM Integration. We have validated the effectiveness of the pretrained semantic prior primarily through personalized CTR post-training (Stage III). However, whether this progressive pretraining paradigm can seamlessly generalize to other downstream search tasks (e.g., early-stage retrieval, conversion rate prediction) remains to be fully explored. Integrating the fine-grained SID representation with Large Language Models (LLMs) to build a more comprehensive, universal foundational model for search systems is an exciting open direction.

Computational Overhead in Ranking vs. Retrieval. Replacing a single global image embedding with 30 multi-scale tokens inevitably increases the width of the concatenated feature vector and the computational cost of the initial bottom-layer projection. However, because the heavy Stage I token extraction is executed entirely offline, online inference only incurs the cost of extra discrete ID lookups and this widened initial projection, before the representation is mapped into fixed-capacity latent slots. In practice, this additional computational overhead is negligible; our online deployment confirmed no significant increase in serving latency during the late-stage ranking phase. Nonetheless, extending this fine-grained tokenization framework to extreme low-latency early-stage retrieval systems (which process millions of candidates) remains an area for future optimization.

7 Conclusion

In this paper, we introduced FineSID, a three-stage progressive learning framework that integrates multimodal content into ID-based CTR prediction. By discretizing multimodal content into fine-grained Semantic IDs, the framework seamlessly unifies content tokenization, semantic relevance pretraining, and personalized post-training within a shared ID-based architecture. This decoupled design enables the predictor to establish a robust relevance-informed semantic prior before adapting to implicit, personalized user behaviors.

Offline evaluations on the industrial Taobao dataset and the public MMCTR benchmark, alongside completed online A/B testing, demonstrate the effectiveness of our approach. Specifically, FineSID not only improves standard product-level CTR prediction but also exhibits strong capability in within-product creative discrimination. These results validate that retaining fine-grained multi-scale structural details through reconstruction-oriented visual tokenization is crucial for unlocking the full potential of multimodal semantics in industrial search systems.

Reproducibility and AI Assistance

To ensure reproducibility, detailed formulations of our evaluation metrics and comprehensive implementation details (including network configurations and hyperparameter settings) are provided in the Appendix.

In accordance with the ICLR 2027 policy on AI assistance, we disclose that Large Language Models (LLMs) were utilized during the preparation of this manuscript. Specifically, AI tools were employed to: (1) aid in the retrieval and discovery of related literature; (2) assist with research ideation and execution; (3) draft initial sections of the paper; and (4) polish the final writing, organize references, and format equations. All AI-generated content was thoroughly reviewed and critically revised by the human authors. The experimental designs, data analyses, and ultimate scientific claims remain entirely the responsibility of the authors.

REFERENCES

Shuzhi Cao, Rong Chen, Ailong He, Shuguang Han, and Jufeng Chen. PRECTR-V2:Unified Relevance-CTR Framework with Cross-User Preference Mining, Exposure Bias Correction, and LLM-Distilled Encoder Optimization. arXiv:2602.20676, 2026. URL https://arxiv.org/abs/2602.20676.

Luming Chen, Jiaqi Xi, Raghav Saboo, Martin Wang, and Sudeep Das. Joint Optimization of Relevance and Engagement in Multi-Task Ranking for E-Commerce with Efficient LLM Supervision. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4513–4517, 2026. doi: 10.1145/3805712.3808481. URL https://doi.org/10.1145/3805712.3808481.

Rong Chen, Shuzhi Cao, Ailong He, Shuguang Han, and Jufeng Chen. PRECTR: A Synergistic Framework for Integrating Personalized Search Relevance Matching and CTR Prediction. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 152–161, 2025. doi: 10.1145/3701716.3715256. URL https://doi.org/10.1145/3701716.3715256.

Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2251–2260, 2020. doi: 10.1145/3397271.3401430. URL https://doi.org/10.1145/3397271.3401430.

Tiezheng Ge, Liqin Zhao, Guorui Zhou, Keyu Chen, Shuying Liu, Huimin Yi, Zelin Hu, Bochao Liu, Peng Sun, Haoyu Liu, Pengtao Yi, Sui Huang, Zhiqiang Zhang, Xiaoqiang Zhu, Yu Zhang, and Kun Gai. Image Matters: Visually Modeling User Behaviors Using Advanced Model Server. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 2087–2095, 2018. doi: 10.1145/3269206.3272007. URL https://doi.org/10.1145/3269206.3272007.

Huan Gui, Ruoxi Wang, Ke Yin, Long Jin, Maciej Kula, Taibai Xu, Lichan Hong, and Ed H. Chi. Hiformer: Heterogeneous Feature Interactions Learning with Transformers for Recommender Systems. arXiv:2311.05884, 2023. URL https://arxiv.org/abs/2311.05884.

Jian Jia, Yipei Wang, Yan Li, Honggang Chen, Xuehan Bai, Zhaocheng Liu, Jian Liang, Quan Chen, Han Li, Peng Jiang, and Kun Gai. LEARN: Knowledge Adaptation from Large Language Model to Recommendation for Practical Industrial Application. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11861–11869, 2025. doi: 10.1609/aaai.v39i11.33291. URL https://ojs.aaai.org/index.php/AAAI/article/view/33291.

Yuchen Jiang, Jie Zhu, Xintian Han, Hui Lu, Kunmin Bai, Mingyu Yang, Shikang Wu, Ruihao Zhang, Wenlin Zhao, Shipeng Bai, Sijin Zhou, Huizhi Yang, Tianyi Liu, Wenda Liu, Ziyan Gong, Haoran Ding, Zheng Chai, Deping Xie, Zhe Chen, Yuchao Zheng, and Peng Xu. TokenMixer-Large: Scaling Up Large Ranking Models in Industrial Recommenders. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 7489–7500, 2026. doi: 10.1145/3770855.3818494. URL https://doi.org/10.1145/3770855.3818494.

Hu Liu, Jing Lu, Hao Yang, Xiwei Zhao, Sulong Xu, Hao Peng, Zehua Zhang, Wenjie Niu, Xiaokun Zhu, Yongjun Bao, and Weipeng Yan. Category-Specific CNN for Visual-aware CTR Prediction at JD.com. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2686–2696, 2020. doi: 10.1145/3394486.3403319. URL https://doi.org/10.1145/3394486.3403319.

Huanyu Liu, Baining Chen, Hui Liu, Zengyang Li, and Ziyi Huang. PaletteID: Prototype-Composed Semantic Identifiers for Multimodal CTR Prediction. arXiv:2607.29000, 2026. URL https://arxiv.org/abs/2607.29000.

Kaiming Shen, Xichen Ding, Zixiang Zheng, Yuqi Gong, Qianqian Li, Zhongyi Liu, and Guannan Zhang. SEMINAR: Search Enhanced Multi-modal Interest Network and Approximate Retrieval for Lifelong Sequential Recommendation. arXiv:2407.10714, 2024. URL https://arxiv.org/abs/2407.10714.

Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, Ed Chi, and Xinyang Yi. Better Generalization with Semantic IDs: A Case Study in Ranking for Recommendations. In *18th ACM Conference on Recommender Systems*, pp. 1039–1044, 2024. doi: 10.1145/3640457.3688190. URL https://doi.org/10.1145/3640457.3688190.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 84839–84865, 2024. doi: 10.52202/079017-2694. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9a24e284b187f662681440ba15c416fb-Abstract-Conference.html.

Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2560–2564, 2022. doi: 10.1145/3477495.3531896. URL https://doi.org/10.1145/3477495.3531896.

Guipeng Xv, Si Chen, Chen Lin, Wanxian Guan, Xingyuan Bu, Xubin Li, Hongbo Deng, Jian Xu, and Bo Zheng. Visual Encoding and Debiasing for CTR Prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4615–4619, 2022. doi: 10.1145/3511808.3557721. URL https://doi.org/10.1145/3511808.3557721.

Jia-Qi Yang, Chenglei Dai, Dan OU, Dongshuai Li, Ju Huang, De-Chuan Zhan, Xiaoyi Zeng, and Yang Yang. COURIER: Contrastive User Intention Reconstruction for Large-Scale Visual Recommendation. arXiv:2306.05001, 2024a. URL https://arxiv.org/abs/2306.05001.

Zhiguang Yang, Liufang Sang, Haoran Wang, Wenlong Chen, Lu Wang, Jie He, Changping Peng, Zhangang Lin, Chun Gan, and Jingping Shao. Parallel Ranking of Ads and Creatives in Real-Time Advertising Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9278–9286, 2024b. doi: 10.1609/aaai.v38i8.28780. URL https://ojs.aaai.org/index.php/AAAI/article/view/28780.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cpDhcsEDC2.

Tan Yu, Zhipeng Jin, Jie Liu, Yi Yang, Hongliang Fei, and Ping Li. Boost CTR Prediction for New Advertisements via Modeling Visual Content. In *2022 IEEE International Conference on Big Data (Big Data)*, 2022. doi: 10.1109/BigData55660.2022.10020786. URL https://doi.org/10.1109/BigData55660.2022.10020786.

Qihang Zhao, Zhongbo Sun, Xiaoyang Zheng, Xian Guo, Siyuan Wang, Zihan Liang, Mingcan Peng, Ben Chen, and Chenyi Lei. COINS: Semantic Ids Enhanced Cold Item Representation for Click-through Rate Prediction in E-commerce Search. In *Proceedings of the ACM Web Conference 2026*, pp. 8529–8532, 2026. doi: 10.1145/3774904.3792902. URL https://doi.org/10.1145/3774904.3792902.

Guorui Zhou, Xiaoqiang Zhu, Chengru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep Interest Network for Click-Through Rate Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1059–1068, 2018. doi: 10.1145/3219819.3219823. URL https://doi.org/10.1145/3219819.3219823.

Jie Zhu, Zhifang Fan, Xiaoxie Zhu, Yuchen Jiang, Hangyu Wang, Xintian Han, Haoran Ding, Xinmin Wang, Wenlin Zhao, Zhen Gong, Huizhi Yang, Zheng Chai, Zhe Chen, Yuchao Zheng, Qiwei Chen, Feng Zhang, Xun Zhou, Peng Xu, Xiao Yang, Di Wu, and Zuotao Liu. RankMixer: Scaling Up Ranking Models in Industrial Recommenders. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6309–6316, 2025. doi: 10.1145/3746252.3761507. URL https://doi.org/10.1145/3746252.3761507.

Jieming Zhu, Jinyang Liu, Shuai Yang, Qi Zhang, and Xiuqiang He. Open Benchmarking for Click-Through Rate Prediction. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 2759–2769. ACM, 2021. doi: 10.1145/3459637.3482486. URL https://doi.org/10.1145/3459637.3482486.

A Measurement Details

For a specific evaluation group g (e.g., a query group) containing n_g^+ positive samples (clicks) and n_g^- negative samples (non-clicks), let W_g denote the number of correctly ordered positive–negative

pairs plus half of the tied pairs. The exact AUC for this group is calculated as $\mathrm{AUC}_g = W_g/(n_g^+ n_g^-)$. For the set of valid groups $\mathcal{G}_+ = \{g : n_g^+ n_g^- > 0\}$, the Group AUC (GAUC) is defined as:

$$\mathrm{GAUC} = \frac{\sum_{g \in \mathcal{G}_+} n_g \mathrm{AUC}_g}{\sum_{g \in \mathcal{G}_+} n_g}, \qquad n_g = n_g^+ + n_g^-. \tag{9}$$

The denominator effectively weights the contribution of each valid group by its total impression count n_g. Note that while Item GAUC compares exposures across different users and contexts for a given item, it is fundamentally distinct from NSCTR (defined in Section 4.7), which specifically evaluates the fine-grained ranking of different creatives within the exact same item ID.

B Implementation and Reproducibility Details

Tokenizer and Feature Configurations. The framework explicitly distinguishes between the upstream image tokenizer and the downstream trainable SID lookup tables. For Stage I visual tokenization, we utilize a pretrained VAR checkpoint. We retain code maps corresponding to resolutions of 1×1, 2×2, 3×3, and 4×4, yielding exactly 30 spatially indexed discrete codes $(1 + 4 + 9 + 16)$ per image. Note that this multi-scale retention is an explicit subset of the source tokenizer's full scale schedule. Missing visual codes are handled via a dedicated validity mask $m_{\ell,p}$. Concurrently, text sources (e.g., query, title) are encoded using distinct BPE and n-gram vocabularies to preserve source-specific sequence identities.

Network and Full-Parameter Transfer (Stage II to Stage III). The shared CTR predictor incorporates 65 dynamic categorical embedding tables alongside the multimodal SIDs. During the Stage III personalized post-training, the model undergoes a strict full-parameter transfer protocol. Specifically, all 197 dense network states and normalization entries learned during the Stage II relevance pretraining are exactly retained and loaded. The sparse Adagrad accumulators are reset to avoid carrying relevance-stage adaptive learning-rate statistics into CTR training, and a fresh training step is initialized for the CTR task.

Training Hyperparameters and Hardware. The shared predictor's dense parameters and sparse embeddings are optimized using Adagrad with a fixed learning rate of 0.01. Training is distributed across 64 NVIDIA H20 GPUs, with a per-GPU batch size of $4{,}096$ and a global batch size of $262{,}144$. No learning-rate warmup or decay schedule is used. For the multi-scale VAR tokenizer, the discrete visual codebook size is fixed at $V = 4{,}096$.