

New Research Proposes Truly ‘Personalized’ Advertising

Originally published at <https://www.unite.ai/new-research-proposes-truly-personalized-advertising/>



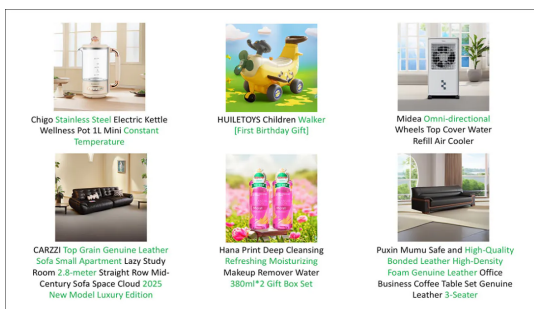
In a redefinition of ‘self-promotion’, a new method mines a user’s own clicks to create bespoke web ads based on their own particular history.

Though ad agencies are [keen to debunk](#) the idea that advertising funnels exist which can serve you ads based on what you just said in the comfort of your home, the extent of ‘personalization’ demonstrated by adverts in websites and social media apps has nonetheless [garnered headlines](#) in recent years.

The ideal scenario for the advertiser has always been that the ad served be an ‘exact fit’ for the viewer. Within the limits of [public pushback](#) about online tracking, and whatever [preventive measures](#) the user might have installed against such monitoring, generative AI (setting aside fears around [LLM advertising](#) in a [post-search world](#)) is quite capable of producing ad images and copy quickly enough for real-time deployment.

However, the main thrust of research and the bulk of implementations in this line to date have been based around [aggregate](#) usage statistics, so that any ad generated for a viewer would be based on the viewer’s guessed [cohort group](#) rather than their own unique history.

Now, a new research collaboration between China and the US presents a system for generating advertising images and text for individual users by learning from their own past clicks when logged into a site, moving beyond the cohort-based assumptions that have governed most personalized advertising research to date:



Example generations depicting individually bespoke ads. Of course, without the user’s history as context, the full impact can only be imagined. [Source](#)

Unusually, the new approach eschews [diffusion-based](#) models in favor of an *autoregressive* architecture – the primary difference being that [diffusion](#) models gradually refine an image from visual noise, whereas autoregressive models generate content [one piece at a time](#), predicting each new element from everything that came before.

To support the new generative model, the authors developed what they claim is the first large-scale image/text dataset for personalized advertising, as well as a novel metric designed to evaluate this very specific task. In tests, they found that their approach outperformed both general baselines and the existing methods and frameworks that currently address this challenge.

Walled Garden

It's worth noting the proposed scope of the work, which does not offer advertisers a way to subvert new measures against third-party tracking, but instead gives a sufficiently large retailer the power to populate a *logged-in customer* with ads that directly relate to that specific person.

This is not necessarily confined to clients who are currently browsing the retailer's own site: depending on the extent to which the user has granted the retailer the power to track them across other sites, they could be presented with targeted ads in any number of other websites that participate in ad auctions that the retailer themselves uses.

This kind of advertising reach tends to be limited to high-volume, high-scale outlets such as Amazon, in the west (and we note that an analogously-sized Chinese retailer has participated in the new work – see below), though any similarly-sized concern (such as a popular social media platform) could in theory generate a similar generative framework.

The [new paper](#) is titled *Design Your Ad: Personalized Advertising Image and Text Generation with Unified Autoregressive Models*, and comes from 18 authors across Sun Yat-Sen University in Guangzhou, Northeastern University, and China's largest retailer, JD.com 📈 JD -0.97% (the latter of which has that precious access to shoppers' histories and habits). The code has been made available [via GitHub](#), and the relevant [checkpoints made available too](#).

Data and Method

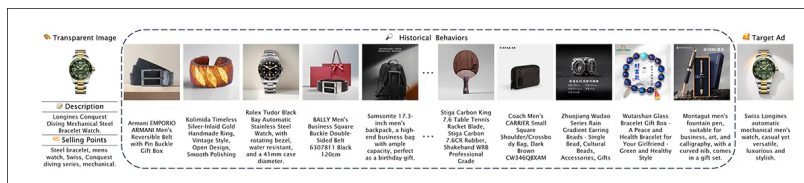
The dataset constructed for the project is titled *Personalized Advertising image-text* (PAD1M), and is powered by data provided by project contributor JD.com. The authors state:

'Each product typically provides more than ten candidate images and texts, ensuring that the diverse preferences can be fully detected. To enable reliable preference modeling, we collect complete user click histories over both images and texts, filtering out users with insufficient activity to reduce noise.'

'This yields a dataset of 1,145,371 users, with 18,923,555 clicked product images and texts, averaging more than sixteen multimodal historical behaviors per user.'

For each user, one previously-clicked image-text pair was selected as the target example, after which the product itself was isolated from the image using [Grounded SAM](#).

Seller-supplied descriptions and selling points were then attached to the record, creating a dataset in which each target advertisement is accompanied by a transparent product image; structured product information; and a history of earlier image and text interactions, intended to capture the user's prior interests and preferences:



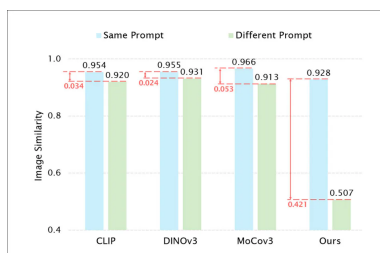
A user profile from the PAD1M dataset, showing a target advertisement alongside the product information used to generate it, and the historical image and text into user's preferences.

The resulting dataset offers a scale of over a million users, and nearly 19 million clicked-image and text records, with the authors stating that the collection is substantially larger than previous personalization datasets.

Additionally, the data, unusually for this strand of research, combines both images and text, allowing user preferences to be modeled across multiple modalities, rather than within a single domain.

PAD1M also features individual-level preference tracking; unlike the run of prior advertising datasets, which were built around click-through rates aggregated across large groups, PAD1M links interactions to specific users from the JD.com data.

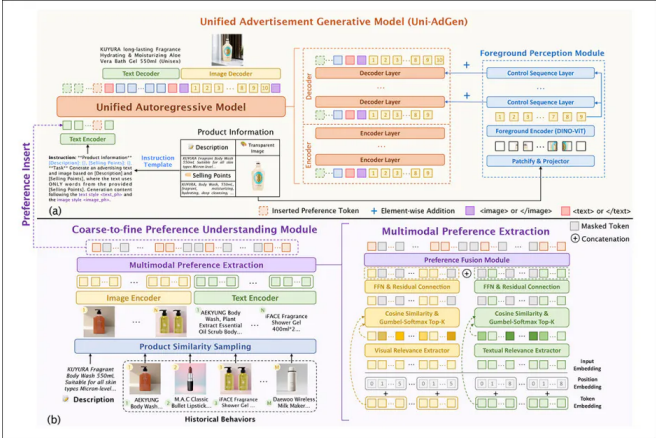
For metrics, besides the standard choices of [BLEU](#) and [ROUGE](#), the researchers developed their own custom measurement titled *Product Background Similarity* (PBS). Based on the prior [MoCo-v3](#) initiative, PBS was trained on 681,123 image pairs showing the same product against different backgrounds, allowing the metric to focus on contextual variation rather than the product itself:



Product Background Similarity (PBS) assigns markedly different similarity scores to advertisements that contain the same product, but place it in different visual contexts. Conversely, competing metrics produce much smaller separations.

During training, each image was paired with itself as a positive example, while an image of the same product placed in a *different setting* served as a negative example, a training strategy intended to increase sensitivity to background context. Evaluation results, the paper contends, indicate bigger similarity differences between matching and non-matching backgrounds than those produced by [CLIP](#), [DINO v3](#), or the aforementioned MoCov3.

As shown in the upper-left section of the image below*, the researchers' *Unified Advertisement Generative (Uni-AdGen)* model uses an autoregressive vision-language architecture to generate both advertising text and images. The process is guided by a structured instruction that includes the task definition, and a product description, along with key selling points:



Method overview.

Special delimiting tokens define the portion of the sequence reserved for advertising copy. After the text has been generated, a dedicated image token triggers image generation, while a closing image token marks its completion, with generated tokens subsequently sent to separate text and image decoders.

For images, [LlamaGen's](#) VQ-GAN decoder is used to convert discrete image tokens back into pixels.

In this way, the unified architecture generates text and images within a single [next-token prediction](#) framework, rather than relying on separate pipelines – the method adopted for earlier advertising systems with a similar ambit.

During training, the model learns both modalities together, with text tokens predicted based on the input sequence and previously-generated text. Image tokens are then predicted using the input sequence, the generated text, and previously-generated image tokens.

To keep generated advertisements tied to the promoted product, Uni-AdGen uses a *foreground-perception module* based on [DINO v2](#), to inject information from transparent product images into the autoregressive model.

[Instruction-tuning](#) (training the model to follow product-specific generation instructions derived from descriptions and selling points) was also used to improve adherence to seller-provided descriptions and selling points, with [GPT-4o](#) used to filter unsuitable training examples.

Personalization relied on a [coarse-to-fine](#) preference-understanding module. Historical interactions were first filtered through a Product Similarity Sampling (PSS) pipeline to favor products resembling the target item. The remaining records were then processed by a *Multimodal Preference Extraction* stage designed to identify the visual and textual elements most likely to reflect user interests – with those preferences inserted into the prompt, to guide generation.

Tests

The authors state that their testing approach is derived from [DeepSeek's Janus-Pro 7B](#).

The model was trained at a [batch size](#) of four, under the [AdamW](#) optimizer at a [learning rate](#) of 5e-5. The base model was fine-tuned via [LoRA](#), with the foreground perception and multimodal preference extraction *fully fine-tuned* (i.e., unlike with LoRA, the base model weights were permanently altered).

All tests were run on a NVIDIA B200 GPU with 192GB of VRAM. For image generation, [PickScore](#), [ImageReward](#), and [ASE](#) were used to measure visual quality, while m-BLEU and m-ROUGE[†] were used to evaluate advertising text. Human evaluators additionally assessed image realism and layout quality, along with textual accuracy and fluency, with all metrics computed across 500 products.

For image generation, the baselines comprised [Qwen2.5-VL](#) and GPT-4o for creating background prompts from product images, followed by [ReliableAd](#), [PosterMaker](#), and [Flux-Fill](#) for generating the final advertisements. Text-generation comparisons were conducted against Qwen2.5, [Qwen3](#), and [DeepSeek-R1](#).

Initial baseline quantitative results for ad generation are shown below:

Method	Image Generation				Text Generation		
	ASE ↑	IR↑	PS↑	Human Eval†	m-BLEU↑	m-ROUGE↑	Human Eval†
<i>Image Generation Models</i>							
GPT-4o [1] + ReliableAd [18]	5.26	-1.491	20.888	93.80	—	—	—
Qwen2.5vl [2] + ReliableAd [18]	5.29	-1.516	20.890	95.20	—	—	—
GPT-4o [1] + PosterMaker [21]	5.42	-1.332	20.970	90.00	—	—	—
Qwen2.5vl [2] + PosterMaker [21]	5.34	-1.399	20.977	90.80	—	—	—
GPT-4o [1] + Flux-Fill [30]	5.21	-1.281	20.926	88.00	—	—	—
Qwen2.5vl [2] + Flux-Fill [30]	5.14	-1.311	20.918	86.80	—	—	—
Ours-image	5.42	-1.351	21.081	92.40	—	—	—
<i>Text Generation Models</i>							
Qwen2.5 [73]	—	—	—	—	0.530	0.651	97.00
Qwen3 [48]	—	—	—	—	0.562	0.652	99.60
DeepSeek-R1 [22]	—	—	—	—	0.533	0.653	97.80
Ours-text	—	—	—	—	0.559	0.650	97.80
<i>Image and Text Joint Generation Models</i>							
Ours	5.25	-1.244	21.002	92.60	0.551	0.654	98.20

Performance on the general advertising-generation benchmark. Uni-AdGen matched or exceeded the strongest image-generation baselines on aesthetic quality a image-and-text model achieved the highest m-ROUGE score among all text-generation approaches. Human evaluation results remained competitive across both

Of these results, the authors state:

'[Our] method achieves the best performance in ImageReward and ranks second in both PickScore and human evaluation, demonstrating its superior performance in aesthetic and high available rate. While ReliableAd leads in human evaluation, it lags significantly in aesthetic metrics. Conversely, PosterMaker and Flux-Fill generate visually appealing images but suffer from noticeable usability limitations.'

'Thanks to effective control approaches, our method successfully achieves an optimal balance between visual content and practical utility.'

Personalized ad-generation was evaluated on 500 users with recorded interaction histories, using the aforementioned PBS to measure image similarity, and BLEU and ROUGE to compare generated text against products the users had actually clicked.

Because the general advertising baselines used in the previous experiment could not incorporate user histories, the comparisons were shifted to systems designed for personalization. For image generation, [Flux-Kontext](#) and [Pigeon](#) were selected as baselines. Flux-Kontext was supplied with a grid of historical user images alongside the target product image, allowing prior preferences to influence generation.

Since Pigeon does not natively support controlled product placement, the foreground-perception module developed for Uni-AdGen was integrated to preserve product consistency. For text generation, Qwen3 and DeepSeek-R1 were used, with historical product descriptions inserted directly into their instruction templates to provide user-specific context:

Method	PBS ↑	BLEU↑	ROUGE↑
<i>Image Generation Models</i>			
Pigeon [72]	0.624	–	–
Flux-Kontext [31]	0.514	–	–
<i>Text Generation Models</i>			
Qwen3 [73]	–	0.345	0.580
DeepSeek-R1 [22]	–	0.373	0.622
<i>Image and Text Joint Generation Models</i>			
Baseline	0.617	0.225	0.525
w/ history	0.606	0.427	0.650
w/ PSS	0.622	0.430	0.652
Ours	0.634	0.435	0.662

Personalized ad generation results. Uni-AdGen outperformed Flux-Kontext, Pigeon, Qwen3, and DeepSeek-R1 across all reported personalization metrics, while the ablation study indicated that historical user data, Product Similarity Sampling (PSS), and multimodal preference extraction each contributed measurable gains.

Here the authors comment:

'The visualized results [included in image below] show that Flux-Kontext fails to understand user preferences and remains susceptible to sample-level noise, resulting in significant deviation from ground truth, such as the irrelevant items in the motorcycle image.'

Advertising Image Generation				Advertising Text Generation			
Flux-Kontext	Pigeon	Ours	Ground Truth	Qwen3: Kouzi Cellar rich cellar aroma Baijiu 50% ABV 500ml*1 Bottle Mid-Autumn Festival Gift Set	Qwen3: Xijiu Cellar Reserve 1988 Soy Sauce-Aroma Baijiu 719ml 53% ABV Shipped in Original and Full Case Gift Set of 6 Bottles	Qwen3: Xinbei Portable Electric Heating Cup, Insulated Constant Temperature Cup, Milk warming magic tool 8275-1	Qwen3: Xinbei Portable Electric Heating Cup, Insulated Constant Temperature Cup, Milk warming magic tool 8275-1
				DeepSeek-R1: Kouzi Cellar 1998 Cellar Reserve time-honored heritage fragrance rich cellar aroma Baijiu 50% ABV 500ml * 1 Bottle	DeepSeek-R1: Xijiu Cellar Reserve 1988 Soy Sauce-Aroma Baijiu 719ml 53% ABV Full Case of 6 Bottles Shipped in Original Case Gift Set	DeepSeek-R1: Xinbei Wireless Portable Constant Temperature Cup 8275-1 Milk warming magic tool	DeepSeek-R1: Xinbei Wireless Portable Constant Temperature Cup 8275-1 Milk warming magic tool
				Ours: Kouzi Cellar 1998 Cellar Reserve Blended-aroma Baijiu 50% ABV 500ml*1 bottle Single-bottle pack, gift box set	Ours: Xijiu Cellar Reserve 1988 719ml 53% ABV Soy Sauce-Aroma Baijiu 2021-2022 Vintage Gift Set Full Case Large-Capacity 719ml Bottles 6 Bottles per Case Shipped in Original Case	Ours: Xinbei Wireless Portable Constant Temperature Cup Constant temperature insulation Baby and Children's Travel Bottle for Heating Milk, Milk warming magic tool 8275-1	Ours: Xinbei Wireless Portable Constant Temperature Cup Constant temperature insulation Baby and Children's Travel Bottle for Heating Milk, Milk warming magic tool 8275-1
				Ground Truth: Kouzi Cellar 1998 Cellar Reserve Blended-aroma Baijiu 50% ABV 500ml*1 bottle Single-bottle pack	Ground Truth: Xijiu Cellar Reserve 1988 719ml 53% ABV Soy Sauce-Aroma Baijiu Large Capacity Gift Set 2021-2022 Vintage 53% ABV 719ml 6 Bottles Full Case (Shipped in Original Case)	Ground Truth: Xinbei Wireless Portable Constant Temperature Cup Baby Water Bottle Insulated Children's Travel Bottle for Heating Milk 8275-1	Ground Truth: Xinbei Wireless Portable Constant Temperature Cup Baby Water Bottle Insulated Children's Travel Bottle for Heating Milk 8275-1

Examples of personalized-ad generation. Compared with Flux-Kontext, Pigeon, Qwen3, and DeepSeek-R1, Uni-AdGen produced images that more closely match advertisements users actually clicked, while generating text that captured a larger proportion of the product attributes and selling points present in the ground-truth highlighted in green.

The qualitative examples, the authors contend, indicate that Flux-Kontext and Pigeon often produced outputs diverging from the visual characteristics of advertisements that users had previously clicked; meanwhile, text generated by Qwen3 and DeepSeek-R1 omitted some selling points present in the ground-truth examples.

Conclusion

The utility of this project depends entirely on user opt-in, and extending the reach of this 'predictive' system beyond the scope of the domain controlling the user history – in this case, JD.com – requires an even more relaxed set of explicit user permissions, in most territories.

However, the system is predicated on the kind of hyperscale network effect at work in such a scenario, and on the (perhaps slightly hopeful) idea that users will find this kind of truly personalized and even prescient recommender system useful rather than intrusive, at least within the context of a retail behemoth's walled garden.

* This image builds on the worrying new trend of 'collated figures' in research papers, wherein illustrations that would once have been 3-4 different figures are collated into one (for the purpose of obeying submission guidelines on the maximum length of the main paper) and used solely as reference material, often without adequate explanation in the accompanying caption.

† 'm'-prefix indicates comparison with multiple candidate texts.

First published Tuesday, June 2nd 2026. Amended 18:21 EET to correct final 'wall' to 'walled' in last para.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More