VPBRANDLOOP: AIGC ADVERTISING IMAGE GENERATION WITH CROSS-SCENE VANISHING-POINT CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Computational advertising requires scalable creative generation that preserves scene structure and brand identity while supporting reliable pre-deployment quality control. We present VPBrandLoop (Vanishing-Point and Brand-conditioned Closed-Loop Advertising Generation), a framework for multi-scene AIGC advertising image generation that addresses perspective drift, under-utilized visual brand conditions, and the absence of verifiable output-side geometric control. A precomputed dominant vanishing point (VP) and perspective attributes are injected into a latent diffusion model through multi-scale geometric residuals, while product and logo references are encoded as visual brand tokens and fused through decoupled cross-attention. Training uses brand-region-weighted diffusion reconstruction, spatial-gating alignment, and auxiliary brand-discriminative supervision. At inference time, a line-segment detector (LSD) and RANSAC re-estimate the dominant VP of each generated candidate; spatial-consistency quality gating and confidence-weighted geometric feedback then drive candidate resampling. We additionally use a CLIP-based zero-shot protocol to evaluate the multimodal semantic retrievability of brands, product categories, and visually observable attributes. On 1,000 AdsGeo-10K test images evaluated across 12 methods, VPBrandLoop achieves an N-VPE of 0.2558 and an SCPR of 16.9%. Relative to ControlNet, the strongest external baseline on the corresponding metrics, N-VPE decreases by 20.3% and SCPR increases by 4.2 percentage points; both paired differences remain significant after Holm correction. The evidence therefore supports a primary advantage in preserving dominant perspective consistency between the input scene and the generated advertisement.

1 INTRODUCTION

Programmatic delivery, personalized recommendation, and multi-channel creative production require advertising systems to produce large volumes of visual assets quickly and economically. In practice, useful advertising images must satisfy three requirements simultaneously: product and brand elements should obey the dominant perspective of roads, buildings, tabletops, and other scene structures; logos, package colors, and product appearance should preserve a stable visual identity rather than be approximated from text alone; and outputs should expose automatically checkable quality signals so that low-quality candidates can be rejected or resampled before deployment. In short, an advertising generator must not only look plausible, but also place content correctly, preserve identity, and permit verification.

Latent diffusion models (LDMs) provide a strong generative foundation. Stable Diffusion performs diffusion and denoising in latent space to reduce the cost of high-resolution synthesis (Rombach et al., 2022); SDXL improves high-resolution image synthesis (Podell et al., 2024); and Stable Diffusion 3 extends the framework with multimodal diffusion transformers and rectified flow (Esser et al., 2024). Yet general generative capability does not directly guarantee scene geometry, brand fidelity, or verifiable output control.

ControlNet, T2I-Adapter, GLIGEN, and IP-Adapter improve controllability through explicit spatial conditions, lightweight adapters, grounding constraints, and image prompts, respectively (Zhang

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.

et al., 2023; Mou et al., 2024; Li et al., 2023; Ye et al., 2023). However, the presence of an input condition does not imply that the final image continues to satisfy it: img2img reconstruction, brand insertion, and multi-candidate sampling can alter structural lines and local scale relationships. We therefore treat output-side geometric re-verification as a control stage rather than as an afterthought: after generation, the dominant VP is re-estimated and a measurable spatial score determines whether a candidate is accepted, rejected, or resampled.

Advertising images also face a machine-readability requirement. CLIP maps images and text into a shared semantic space (Radford et al., 2021), while BLIP improves vision-language understanding and retrieval (Li et al., 2022). Generative Engine Optimization (GEO) studies discoverability, citation, and ranking in generative engines (Aggarwal et al., 2024). Since actual engine rankings and citations are controlled by external systems, we use GEO only as an application-level motivation and evaluate brand/category/visible-attribute retrievability with a fixed vision-language model; we do not treat GEO as a differentiable training objective for the current generator.

In this paper, *cross-scene* means that one generation and evaluation framework covers indoor home, outdoor street, nature, office, and dining scenes; it does not denote strict leave-one-scene-out domain generalization. Input VP priors are read from precomputed/annotated AdsGeo-10K geometry, whereas output VPs are estimated online with a common LSD+RANSAC verifier. Our contributions are: (1) a unified advertising-generation framework combining scene context, visual brand conditioning, and verifiable output-side geometric control; (2) an output VP re-verification and closed-loop candidate mechanism that turns geometric constraints into detectable, rejectable, and resampleable inference-time signals, together with decoupled brand conditioning and spatial-gating supervision; and (3) a unified re-evaluation of 12 generation/editing methods on AdsGeo-10K using paired tests, module ablations, and scene-wise analysis, including non-significant results and failure modes rather than only favorable metrics.

2 Related Work

2.1 Controllable diffusion and multimodal conditioning

LDMs balance generation quality and efficiency by operating in a low-dimensional latent space (Rombach et al., 2022). SDXL scales model capacity and high-resolution training (Podell et al., 2024), while Stable Diffusion 3 strengthens text understanding through a multimodal diffusion transformer and rectified flow (Esser et al., 2024). ControlNet adds conditional branches for edges, depth, and segmentation (Zhang et al., 2023); T2I-Adapter aligns lightweight external controls with a frozen generator (Mou et al., 2024); GLIGEN introduces open-set grounding and layout constraints (Li et al., 2023); and IP-Adapter fuses text and image prompts through decoupled cross-attention (Ye et al., 2023). Condition expressiveness continues to expand through controllable visual text generation (AnyText (Tuo et al., 2023) and TextDiffuser-2 (Chen et al., 2024a)), transferable visual representations (DINOv2 (Oquab et al., 2023)), monocular depth (Depth Anything V2 (Yang et al., 2024)), efficient high-resolution generation (SANA (Xie et al., 2025)), unified multimodal generation (Janus-Pro (Chen et al., 2025), OmniGen2 (Wu et al., 2025)), and general image editing (Step1X-Edit (Liu et al., 2025)). These works primarily strengthen pre-generation conditioning; we additionally ask whether the generated result can be verified to satisfy the intended geometry.

2.2 Generative advertising and visual brand conditioning

Generative AI is shifting computational advertising from procedural composition toward knowledge-supported, agentic, multimodal creative production. Prior advertising research emphasizes that generative creativity depends not only on automatic content production, but also on domain knowledge, multimodal expression, and downstream effectiveness (Duan et al., 2025). For brand identity, better captions can improve prompt following (Betker et al., 2023), but text remains insufficient for exact package, logo, and product appearance. Visual attention is also spatially structured and predictable (Borji & Itti, 2013; Kummerer et al., 2016). Our implementation does not use saliency maps as a differentiable generator objective; instead, product and logo references are encoded as visual tokens, injected through an independent brand K/V pathway, and constrained by brand-region-weighted reconstruction and spatial-gating supervision.

2.3 VANISHING-POINT GEOMETRY AND OUTPUT VERIFICATION

A vanishing point describes the image-plane convergence of projected parallel 3D lines. Classical detectors combine line segments, voting, clustering, or RANSAC, while learning-based methods include inverse-gnomonic CNNs (Kluger et al., 2017), VPGNet (Lee et al., 2017), geometric-prior approaches (Lin et al., 2022), VPDETR (Chen et al., 2024b), and recurrence-based VP detection (Bharadwaj et al., 2024). In VPBrandLoop, VP information plays two distinct roles: a precomputed input VP provides a dominant-perspective reference, while a weight-free LSD+RANSAC verifier independently re-estimates the VP from each generated candidate. This explicitly separates conditioning from condition satisfaction.

2.4 GEO-ORIENTED VISIBILITY AND SEMANTIC RETRIEVABILITY

CLIP and BLIP support zero-shot vision-language matching and retrieval (Radford et al., 2021; Li et al., 2022), while GEO extends optimization to discoverability and citation in generative engines (Aggarwal et al., 2024). In commercial information environments, algorithm-auditing work has also described context-dependent forms of brand visibility in generative search (Duan & Pu, 2026). We therefore evaluate whether machines can reliably parse brand, product category, and visible attributes. This proxy measures machine interpretability/retrievability, not actual search-engine ranking, citation probability, or exposure.

3 PROBLEM FORMULATION AND TRAINING/EVALUATION MODEL

3.1 INPUT CONDITIONS AND OUTPUT GEOMETRY

Given a scene image, a set of visual brand elements, and a text prompt, our goal is to generate advertising candidates while preserving the scene's dominant composition. AdsGeo-10K supplies precomputed/annotated input VP and perspective attributes. During training, the data structure also stores VP confidence, the number of supporting lines, and the inlier ratio; at inference, these three entries are set to fixed neutral values. Thus, the $-$ Input VP ablation tests geometric-condition injection rather than online input-VP detector accuracy. Let n_{in} denote the supporting-line count and η_{in} the inlier ratio. The 7-D training and inference VP features are

$$\mathbf{g}_{\text{vp}}^{\text{train}} = \Big[\frac{u_{\text{in}}}{W}, \frac{v_{\text{in}}}{H}, c_{\text{in}}, \rho_{\text{in}}, r_{\text{in}}, \frac{n_{\text{in}}}{50}, \eta_{\text{in}}\Big]^{\top}, \quad \mathbf{g}_{\text{vp}}^{0} = \Big[\frac{u_{\text{in}}}{512}, \frac{v_{\text{in}}}{512}, 0.5, \rho_{\text{in}}, r_{\text{in}}, 0.2, 0.5\Big]^{\top}. \tag{1}$$

For candidate k, the output verifier estimates $\mathbf{v}_{\text{out}}^{(k)} = (u_{\text{out}}^{(k)}, v_{\text{out}}^{(k)})$, perspective strength $\rho_{\text{out}}^{(k)}$, regularity $r_{\text{out}}^{(k)}$, and confidence $c_{\text{out}}^{(k)}$:

$$\mathbf{g}_{\text{vp,out}}^{(k)} = [\mathbf{v}_{\text{out}}^{(k)}, \rho_{\text{out}}^{(k)}, r_{\text{out}}^{(k)}, c_{\text{out}}^{(k)}]. \tag{2}$$

The scene is VAE-encoded and noised according to img2img strength. Text, visual brand conditions, and multi-scale geometry from a geometric encoder jointly condition denoising:

$$I_{\text{ad}}^{(k)} = D(z_t; c_t, c_b, C_{\text{vp}}^{(k)}, \text{seed}_k), \qquad C_{\text{vp}}^{(k)} = E_{\text{vp}}(\mathbf{g}_{\text{vp}}^{(k)}). \tag{3}$$

3.2 DIRECT TRAINING OBJECTIVES

Following the actual gradient paths in our implementation, the generator and brand-conditioning pathway are directly optimized by region-weighted diffusion reconstruction, brand spatial-gating alignment, and brand-condition classification:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{diff}}^{\text{rw}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}. \tag{4}$$

For sample i, latent channel c, and spatial position p, let $b_i \in \{0, 1\}$ indicate a valid brand condition, $M_{b,\downarrow}^{i}$ be the brand mask downsampled to latent resolution, and α the extra brand-region weight. Then

$$\mathcal{L}_{\text{diff}}^{\text{rw}} = \frac{\sum_{i,c,p} w_{i,p}(\epsilon_{i,c,p} - \epsilon_{\theta,i,c,p})^2}{\sum_{i,p} w_{i,p}}, \qquad w_{i,p} = 1 + \alpha b_i M_{b,\downarrow}^{i}(p). \tag{5}$$

Spatial-gating alignment is computed only on the set $\mathcal{V}$ of samples with valid brand conditions. If G_b^i is the gate map aggregated from cross-attention layers, then

$$\mathcal{L}_{\text{gate}} = \frac{1}{N_v} \sum_{i \in \mathcal{V}} \frac{\|G_b^i - M_{b,\downarrow}^i\|_F^2}{HW}, \qquad \mathcal{L}_{\text{cls}} = -\frac{1}{N_v} \sum_{i \in \mathcal{V}} \log p_\phi(y_i \mid B_i). \tag{6}$$

The implementation also computes output-geometry auxiliaries, texture diagnostics, and CLIP/recognition diagnostics. Reconstructed images used by these diagnostics are produced on a stop-gradient path, so these terms do not directly back-propagate into the U-Net, geometric encoder, or brand cross-attention; auxiliary proxy networks may still be updated by their own losses.

3.3 Output-space and GEO-oriented evaluation

The independent output verifier uses LSD+RANSAC. We define normalized VP error and line-convergence error as

$$\text{N-VPE} = \frac{\|\mathbf{v}_{\text{out}} - \mathbf{v}_{\text{in}}\|_2}{512}, \qquad \text{LCE} = \frac{1}{512|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} d(\mathbf{v}_{\text{out}}, \ell). \tag{7}$$

Consistent with the generation code, spatial consistency is

$$S_{\text{space}} = 0.45 s_{\text{vp}} + 0.20 s_\rho + 0.15 s_r + 0.20 c_{\text{out}}, \tag{8}$$

with $s_{\text{vp}} = \max(0, 1 - 4\,\text{N-VPE})$, $s_\rho = \max(0, 1 - 2|\rho_{\text{out}} - \rho_{\text{in}}|)$, and $s_r = \max(0, 1 - 2|r_{\text{out}} - r_{\text{in}}|)$. For GEO-oriented evaluation, CLIP ViT-L/14 compares the generated image with N_q text concepts covering brand, category, and visible selling points:

$$R_{\text{sem}} = \frac{1}{N_q} \sum_{j=1}^{N_q} \mathbf{1}\left[\cos(E_I(I_{\text{ad}}), E_T(q_j)) \geq \tau_{\text{clip}}\right]. \tag{9}$$

This is a machine-retrievability proxy, not a real generative-engine ranking score.

4 VPBrandLoop

Figure 1 summarizes the pipeline: inputs; precomputed scene VP and spatial-prior quantification; scene-text-brand-geometry conditioned LDM generation; output VP re-verification and spatial-consistency validation; quality gating and final output. Brand visual conditioning and spatial-gating enhancement are training-side mechanisms, while confidence-weighted closed-loop feedback and candidate resampling operate at inference. GEO-oriented multimodal retrievability is evaluated after generation and is not part of the default feedback control.

4.1 Precomputed VP prior and spatial conditioning

For each input scene, AdsGeo-10K provides a dominant VP and perspective attributes that are directly read as the geometric prior; the current inference pipeline does not run an online input-VP detector. During training, `VanishingPointPrior` is the 7-D vector in Eq. 1. The `GeometricConditionEncoder` maps it through an MLP and zero-initialized 1×1 convolutions to multi-scale features, which are injected into U-Net by residual forward hooks. At inference, confidence, supporting-line count, and inlier ratio are fixed to neutral values, whereas VP coordinates, perspective strength, and structural regularity vary by sample. Accordingly, the $-$ Input VP Injection ablation disables the geometric residual path rather than testing detector accuracy.

4.2 Scene-text-brand-geometry conditioned generation

The scene image is VAE-encoded and noised according to img2img strength. A frozen text encoder supplies semantic conditions; geometry enters via multi-scale residual features; and a separate BrandEncoder produces visual brand conditions. CLIP ViT-L/14 features map each product or logo reference to 16 visual tokens of dimension 768. Product and logo references occupy fixed token slots; a missing branch is replaced by zero tokens, and conditioning dropout improves robustness.

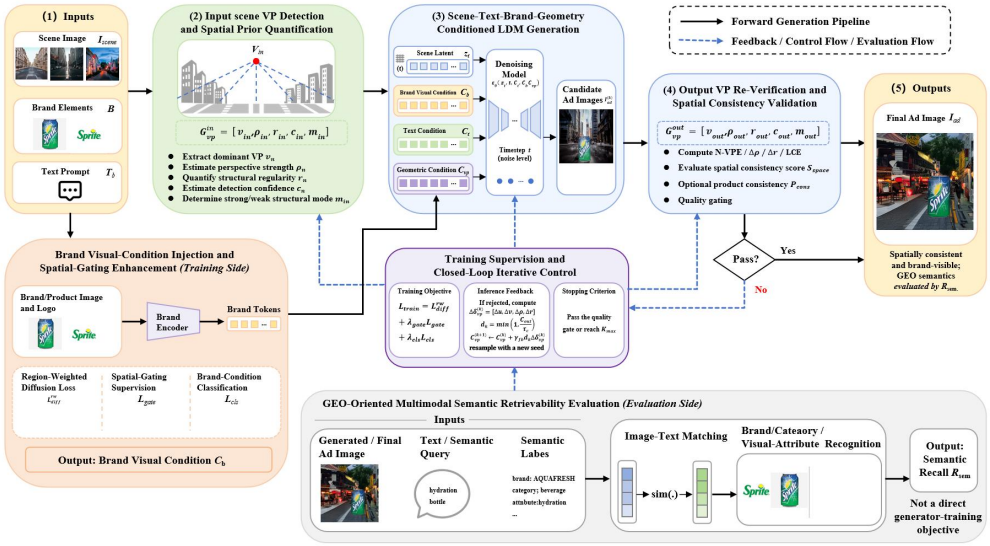

Figure 1: Overall framework of VPBrandLoop. The input VP is an offline precomputed/annotated prior; brand visual conditioning and spatial gating are training-side mechanisms; output geometry re-verification and closed-loop feedback control inference candidates; GEO-oriented semantic retrievability is an evaluation-side module.

Brand tokens are not concatenated to text tokens. Instead, `BrandVPAttnProcessor` adds independent brand K/V projections to U-Net `attn2` layers, creating decoupled text and brand pathways. Brand-condition dropout supports both conditioned and unconditioned inference; text dropout on a subset of branded training samples forces the model to use visual brand tokens for product appearance. Geometry and brand conditions therefore enter by different mechanisms: multi-scale residual features versus cross-attention K/V injection.

4.3 Brand Visual Conditioning and Spatial-Gating Enhancement

Brand annotations are used to construct synthetic training pairs containing foreground products. The default configuration composites a product into the scene according to the foreground placement rule and uses product/logo references as conditions. The brand mask is downsampled to latent resolution for two direct signals: region-weighted diffusion loss up-weights reconstruction error inside the branded region, and spatial-gating supervision aligns the learned brand gate to the target mask. A BrandEncoder classification head adds the auxiliary discriminative loss. We do not claim multi-brand non-overlap, local homography, boundary penalties, saliency heatmaps, or VGG texture terms as effective differentiable layout losses for the current generator.

4.4 Output VP Re-Verification and Spatial Consistency

After each candidate is generated, `VPDetectorProxy` extracts line segments with OpenCV LSD. RANSAC selects the intersection supported by the largest number of lines and robustly refines it using inlier intersections. Output perspective strength is heuristically normalized from the supporting-line count, regularity is the inlier ratio, and confidence combines line support with inlier consistency; no additional learned weights are required.

The current implementation may also infer a product box from the brand spatial gate and compute a product-VP perspective-consistency score. For box $[x_1, y_1, x_2, y_2]$ with $H = W = 512$, define $h_b = y_2 - y_1$, $y_b = y_2$, $x_c = (x_1 + x_2)/2$, and $y_h = \mathrm{clip}(v_{\mathrm{in}}, 0.20H, 0.75H)$. The score components

are

$$s_{\text{size}} = \text{clip}\left(\frac{h_b - 0.25H}{0.20H}, 0, 1\right), \quad s_{\text{bottom}} = \text{clip}\left(\frac{y_b - y_h}{\max(1, 0.95H - y_h)}, 0, 1\right), \tag{10}$$

$$s_{\text{pos}} = \max\left(0, 1 - \frac{\max(0, |x_c - 0.5W| - 0.25W)}{0.25W}\right), \quad s_{\text{prod}} = s_{\text{size}}(0.6s_{\text{bottom}} + 0.4s_{\text{pos}}). \tag{11}$$

If no reliable product box is available, this term is skipped. Candidate k passes the inference gate when

$$Q_k = \mathbf{1}\left[c_{\text{out}}^{(k)} \geq \tau_c^{\text{gen}} \ \wedge \ S_{\text{space}}^{(k)} \geq \tau_s^{\text{gen}} \ \wedge \ \left(a_p^{(k)} = 0 \ \vee \ s_{\text{prod}}^{(k)} \geq \tau_p\right)\right], \tag{12}$$

where $a_p^{(k)}$ indicates whether a product-consistency score is available.

4.5 Closed-loop candidate control and resampling

If a candidate fails the gate, the system changes the random seed and continues sampling under a fixed candidate budget, while retaining the best candidate by output geometric score. The implementation can also update the next VP condition from the residual between output and input geometry. Because the ablation study does not establish a stable independent marginal benefit for residual direction updating, we treat it as an implementation component of closed-loop candidate control rather than as a standalone performance contribution.

At iteration k, the normalized residual is

$$\delta_{\text{vp}}^{(k)} = [\Delta u^{(k)}, \Delta v^{(k)}, \Delta\rho^{(k)}, \Delta r^{(k)}]^\top, \quad \Delta u^{(k)} = \frac{u_{\text{in}} - u_{\text{out}}^{(k)}}{512}, \quad \Delta v^{(k)} = \frac{v_{\text{in}} - v_{\text{out}}^{(k)}}{512}, \tag{13}$$

with $\Delta\rho^{(k)} = \rho_{\text{in}} - \rho_{\text{out}}^{(k)}$ and $\Delta r^{(k)} = r_{\text{in}} - r_{\text{out}}^{(k)}$. To avoid over-correcting from low-confidence VPs,

$$d_k = \min\left(1, \frac{c_{\text{out}}^{(k)}}{\max(\tau_c^{\text{gen}}, 10^{-6})}\right), \qquad \gamma_k = \gamma_{\text{fb}} d_k, \tag{14}$$

and only normalized VP coordinates, perspective strength, and regularity are updated:

$$\mathbf{g}_{\text{vp}}^{(k+1)} = \mathbf{g}_{\text{vp}}^{(k)} + \gamma_k[\Delta u^{(k)}, \Delta v^{(k)}, 0, \Delta\rho^{(k)}, \Delta r^{(k)}, 0, 0]^\top. \tag{15}$$

The next round uses a new seed with a fixed offset from the initial seed. If no candidate passes within the maximum number of rounds, the best generated candidate is returned. The default semantic-guard weight is zero, so candidate selection is primarily geometric/product-consistency based; GEO evaluation does not participate in default feedback. The maximum closed-loop budget is three rounds.

4.6 GEO-oriented multimodal retrievability

The brand-accuracy metric uses CLIP ViT-L/14 for whole-image zero-shot top-1 classification among three target brands. Semantic Recall constructs text concepts from brand name, product category, and visible attributes in `key_selling_points_en`, then applies the thresholded similarity rule in Eq. 9. The evaluation code allows semantic groups to be trimmed so that non-visible concepts such as taste or emotion are not treated as visual targets. Both metrics measure machine parseability under one fixed vision-language model, not independent human brand recognition or real engine exposure.

5 Experiments

5.1 Dataset, implementation, and protocol

AdsGeo-10K contains 10,000 scene images and is split 8:1:1 into train/validation/test. It contains 2,000 images each from indoor home (stable one-/two-point furniture perspective), outdoor street (strong road/building depth), nature (weak explicit line structure), office (stable desk/wall perspective), and dining (occlusion and complex package/table textures). Cross-scene evaluation refers to one model operating across these five distributions, not unseen-domain generalization.

Table 1: Unified re-evaluation on the AdsGeo-10K test set ($n = 1000$ per method). Best values are bold; second-best values are underlined. Brand accuracy has a three-way tie at 100%.

Method	FID↓	VDSR↑	N-VPE↓	LCE↓	SCPR↑	Brand↑	Sem.↑
SD1.5	156.18	44.2	0.4666	0.2512	0.8	58.4	86.3
SDXL	125.17	53.7	0.4231	0.2349	2.2	87.3	90.7
SD3	82.17	61.1	0.4568	0.2208	1.7	98.5	**97.8**
SANA	326.76	22.7	0.3964	0.2174	0.1	38.1	41.0
ControlNet	46.88	64.0	<u>0.3212</u>	0.2332	<u>12.7</u>	97.4	81.5
GLIGEN*	39.70	50.4	0.3413	0.2459	6.5	93.8	87.0
IP-Adapter	48.77	51.3	0.3478	0.2393	7.0	98.7	87.3
T2I-Adapter	194.97	31.0	0.3691	0.2534	1.8	42.0	31.6
Janus-Pro	81.62	62.6	0.4307	<u>0.2088</u>	2.5	89.8	93.0
OmniGen2	46.56	58.7	0.3546	0.2408	7.9	**100.0**	84.2
Step1X-Edit	47.73	**64.9**	0.3298	0.2208	10.0	**100.0**	92.7
VPBrandLoop	**23.96**	63.6	**0.2558**	**0.2069**	**16.9**	**100.0**	81.4

Percent signs are omitted from VDSR, SCPR, Brand, and Sem. columns. N-VPE/LCE significance uses common samples for which both methods obtain a valid VP. GLIGEN* denotes the implementation in our evaluation pipeline; when native GLIGEN weights are unavailable, that pipeline can fall back to SD1.5 img2img, so this row should not be interpreted as an unconditional replacement for the standard GLIGEN implementation.

Training initializes from Stable Diffusion v1.5 at 512×512. We use 4,000 training images, batch size 2, and 30,000 target optimization steps. AdamW learning rates are 5×10^{-5} for U-Net, 10^{-4} for brand-attention layers, and 10^{-5} for new modules, with weight decay 0.01. Brand-region extra weight is $\alpha = 2.0$, spatial-gating weight $\lambda_{\text{gate}} = 0.5$, and brand-classification weight $\lambda_{\text{cls}} = 0.1$. Inference uses DPM-Solver Multistep, 30 diffusion steps, classifier-free guidance 3.0, img2img strength 0.6, and a maximum of three candidate rounds. Evaluation thresholds are calibrated on validation and then frozen: $\tau_c = 0.0008$, $\tau_s = 0.2431$, $\tau_v = 0.0001$ for valid VP detection, and $\tau_{\text{clip}} = 0.20$. We report FID (Inception-v3 pool3 Fréchet distance), VDSR (fraction with $c_{\text{out}} \geq \tau_v$), N-VPE, LCE, SCPR (fraction satisfying both the confidence and spatial-score thresholds), three-brand closed-set CLIP top-1 accuracy, and thresholded CLIP Semantic Recall over brand/category/visible-attribute concepts.

N-VPE and LCE are computed only when the output VP meets the VDSR validity threshold. Image-level statistical tests compare VPBrandLoop with each external method using paired samples: McNemar-type exact tests for binary VDSR, SCPR, and brand accuracy; Wilcoxon/paired bootstrap on common valid samples for N-VPE and LCE and paired analysis for semantic recall. Multi-method comparisons are Holm-corrected within each metric. FID is a set-level distribution metric, so we report it as a common-implementation aggregate rather than treating individual images as independent FID samples.

5.2 Unified comparison across 12 methods

We re-evaluate Stable Diffusion v1.5, SDXL, SD3, SANA, ControlNet, GLIGEN, IP-Adapter, T2I-Adapter, Janus-Pro, OmniGen2, Step1X-Edit, and VPBrandLoop on the same 1,000-image test split, with a common output size, output-VP verifier, and CLIP evaluation protocol. Native inputs differ: SD1.5/SDXL/SD3/SANA/Janus-Pro use text; ControlNet and T2I-Adapter use scene structure plus text; GLIGEN* uses scene, bbox, brand phrase, and text; IP-Adapter and OmniGen2 use scene, product reference, and text; Step1X-Edit uses scene plus a brand-text edit instruction; and VPBrandLoop uses scene, text, product/logo references, and VP. We therefore interpret FID in light of conditioning strength and scene retention.

The clearest gains are in N-VPE and SCPR. VPBrandLoop obtains N-VPE 0.2558, compared with 0.3212 for the strongest external baseline, ControlNet, a 20.3% reduction by the reported means. On the 447 paired images where both methods have a valid VP, the Wilcoxon difference remains highly significant after Holm correction ($p_{\text{adj}} \approx 3.13 \times 10^{-12}$). SCPR is 16.9%, versus 12.7% for ControlNet, an improvement of 4.2 percentage points; the paired exact test over all 1,000 images remains significant after Holm correction ($p_{\text{adj}} \approx 0.0011$). These results indicate that the main geometric benefit is not merely making a VP detectable, but keeping the output dominant perspective closer to the input reference and increasing the probability that a candidate satisfies the combined spatial criterion.

Results for LCE and VDSR are more restrained. The mean LCE of 0.2069 is the lowest among all methods, with Janus-Pro second at 0.2088, but their paired difference is not significant after

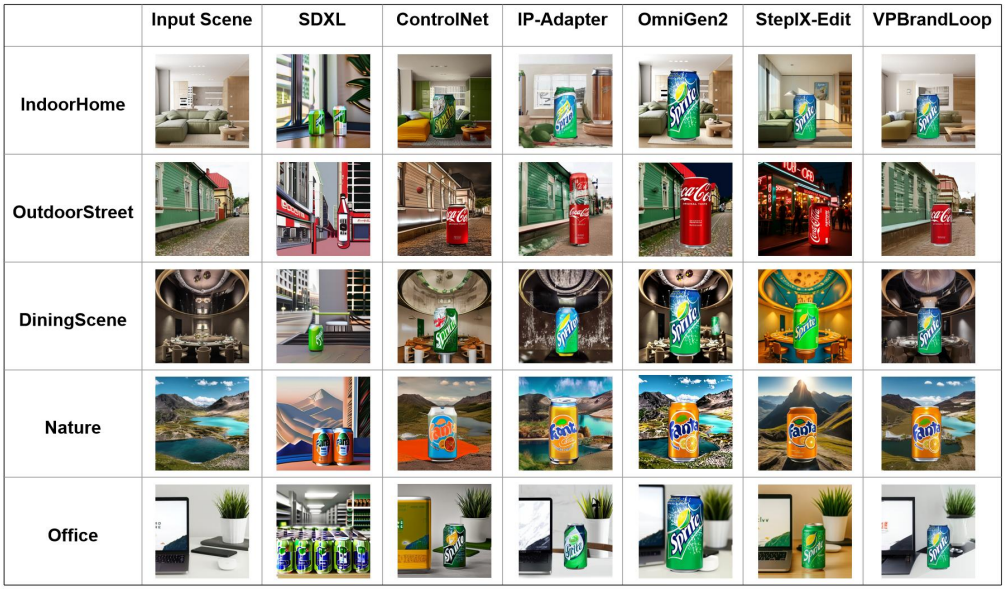

Figure 2: Qualitative comparison across five scene types. Columns show the input scene and outputs from SDXL, ControlNet, IP-Adapter, OmniGen2, Step1X-Edit, and VPBrandLoop.

Table 2: Module-level ablation on 200 validation images.

Variant	N-VPE↓	LCE↓	SCPR↑	Brand↑	Sem.↑
VPBrandLoop	0.1777	0.2155	35.5	27.0	21.5
−Input VP Injection	0.1785	0.2136	37.0	26.5	20.5
−Output Closed Loop	0.2018	0.2272	26.5	15.5	14.5
−Confidence Control	0.1849	0.2184	33.5	26.5	21.5
−Brand Visual Condition	0.1573	0.2108	40.5	7.0	4.5

correction ($p_{\text{adj}} \approx 0.516$). VDSR is highest for Step1X-Edit (64.9%), followed by ControlNet (64.0%) and VPBrandLoop (63.6%); the difference between VPBrandLoop and Step1X-Edit is also non-significant. This contrast suggests that we improve input-output VP position preservation more clearly than internal line-convergence tightness.

FID is lowest for VPBrandLoop (23.96), but native inputs differ substantially across methods: our model directly receives a scene through img2img; ControlNet/T2I-Adapter receive structural scene conditions; IP-Adapter/OmniGen2 additionally receive visual references; and several general models use text only. We therefore do not interpret FID alone as evidence that the generative backbone is universally superior. Brand accuracy reaches 100% for VPBrandLoop, OmniGen2, and Step1X-Edit, but this is a three-brand closed-set CLIP top-1 task with a clear ceiling effect, not open-set logo detection. Semantic Recall does not lead overall: SD3 reaches 97.8%, Janus-Pro 93.0%, Step1X-Edit 92.7%, and VPBrandLoop 81.4%; ControlNet obtains 81.5%, with no significant paired difference from VPBrandLoop ($p_{\text{adj}} \approx 0.940$). Thus, GEO-oriented retrievability is an evaluation dimension rather than a demonstrated advantage of the current method.

5.3 Module-level ablation

We evaluate four module-level ablations on an independent set of 200 validation images: − Input VP Injection, − Output Closed Loop, − Confidence Control, and − Brand Visual Condition. Confidence control jointly affects quality gating and feedback damping and is therefore treated as one control module. Geometric residual feedback and candidate resampling are evaluated as part of the output closed loop rather than as a separate non-orthogonal residual-feedback ablation. GEO/Semantic Recall is an evaluation-side metric, so we do not define a "− GEO loss" variant. Table 2 compares variants only within this validation batch and should not be cross-compared numerically with Table 1.

Disabling output re-verification and the candidate loop increases FID from 174.97 to 203.54, raises N-VPE and LCE, and reduces SCPR from 35.5% to 26.5%; brand accuracy and semantic recall also decrease. This makes output-side verifiable candidate control one of the most stable contributions in the current system. Removing visual brand conditioning reduces brand accuracy from 27.0% to 7.0% and Semantic Recall from 21.5% to 4.5%, while FID rises to 222.50, confirming that the visual brand pathway is actively used by the generator.

In contrast, − Input VP Injection yields slightly higher SCPR than the full model, and − Confidence Control does not show a stable independent advantage. We therefore do not claim that either module must improve every metric in isolation; we treat them as geometric-reference and control components of the closed-loop system. The apparently better N-VPE/SCPR of − Brand Visual Condition may arise because product contours, package text, and brand edges interfere with LSD/RANSAC line estimation. This reveals content dependence in automatic VP evaluation and motivates line-level separation of brand-induced and scene structural lines, together with independent human VP annotation.

5.4 Scene-wise analysis

Scene-wise aggregation over the 1,000 test images gives the following VDSR/N-VPE/LCE/SCPR/Semantic-Recall values (brand accuracy is 100.0% in every group): indoor home, 73.3%/0.2987/0.2136/21.3%/79.1% ($n = 225$); outdoor street, 62.4%/0.1969/0.2116/25.4%/81.5% ($n = 173$); nature, 66.5%/0.2236/0.1722/13.6%/84.5% ($n = 206$); office, 58.4%/0.2904/0.2193/10.7%/87.9% ($n = 214$); and dining, 55.5%/0.2500/0.2226/14.3%/73.1% ($n = 182$). These are not unseen-domain tests, but they reveal substantial scene dependence: outdoor street has the highest SCPR, office the lowest SCPR, and dining the lowest Semantic Recall.

6 Discussion

The strongest evidence for VPBrandLoop is geometric: N-VPE and SCPR significantly improve over ControlNet after Holm correction, whereas VDSR is near but not at the best value and the mean LCE advantage over Janus-Pro is non-significant. The main supported claim is therefore preservation of the dominant output VP relative to the input reference, not simply producing more VP-detectable lines. The absolute SCPR is still only 16.9%, so the closed loop should be viewed as increasing pass probability and enabling rejection/resampling rather than guaranteeing perspective consistency.

Visual brand conditioning strongly improves brand accuracy and Semantic Recall in ablation, but some automatic geometry scores improve when brand conditioning is removed. This may indicate a true brand-versus-background tradeoff or sensitivity of LSD/RANSAC to product edges, package text, and brand contours. GEO is likewise useful as an evaluation perspective rather than a demonstrated performance advantage: Semantic Recall is highest for SD3 and is statistically indistinguishable between VPBrandLoop and ControlNet, consistent with the fact that CLIP/GEO diagnostics are not effective differentiable generator objectives in the current implementation.

The independent benefit of input VP injection is unresolved, possibly because output candidate control compensates for weakened front-end geometry; a future 2×2 Input-VP $\times$ Output-Closed-Loop design should separate these effects. The current input VP is precomputed/annotated, not an online detector. Finally, cross-model interpretation must account for unequal input modalities, and GLIGEN* may fall back to SD1.5 img2img when native weights are unavailable. Future comparisons should match scene input, brand reference, and sampling budget and add open-set brand recognition and independent human geometry evaluation.

7 Conclusion

We presented VPBrandLoop, combining dominant-VP conditioning, decoupled visual brand tokens, spatial brand supervision, and output-side VP verification/resampling. Across 1,000 AdsGeo-10K test images and 12 methods, VPBrandLoop achieves the lowest N-VPE (0.2558) and highest SCPR (16.9%); relative to ControlNet, N-VPE decreases by 20.3% and SCPR rises by 4.2 percentage points, with both paired differences significant after Holm correction. Ablations identify the output closed loop and visual brand conditioning as the most stable contributors.

AI USE STATEMENT

Generative AI tools were used for manuscript language editing, document organization, methodological feedback, reference-formatting assistance, and figure-layout drafting. They were not used to generate experimental measurements. All AI-assisted content, equations, code descriptions, and quantitative claims were reviewed and verified by the authors, who take responsibility for the final manuscript.

REFERENCES

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, et al. Geo: Generative engine optimization. In *Proceedings of the ACM Web Conference*, 2024.

James Betker, Gabriel Goh, Li Jing, et al. Improving image generation with better captions. Technical report, OpenAI, 2023.

Sudeep Bharadwaj, Robert Collins, and Yanxi Liu. Recurrence-based vanishing point detection. *arXiv preprint arXiv:2412.20666*, 2024.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, 2024a.

Tianheng Chen, Xinyi Ying, Jian Yang, Rui Wang, Rong Guo, Bing Xing, and Jian Shi. Vpdetr: End-to-end vanishing point detection transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Chunlin Duan and Yuan Pu. Manipulation and verification: Formation mechanisms, algorithm auditing, and governance paths of brand visibility in generative search engines. *Journal of Nanjing Normal University (Social Science Edition)*, (4):148–157, 2026. In Chinese.

Chunlin Duan, Yuan Pu, and Fang Wei. Model construction and effect study of agents and generative computational advertising creativity. *Journalism and Mass Communication Monthly*, (6):63–73, 2025. In Chinese.

Patrick Esser, Sumith Kulal, Andreas Blattmann, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Florian Kluger, Hanno Ackermann, Michael Ying Yang, and Bodo Rosenhahn. Deep learning for vanishing point detection using an inverse gnomonic projection. In *German Conference on Pattern Recognition*, pp. 17–28. Springer, 2017.

Matthias Kummerer, Thomas S. A. Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

Seokju Lee, Junsik Kim, Jae Shin Yoon, et al. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1947–1955, 2017.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Yuheng Li, Haotian Liu, Qingyang Wu, et al. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

Yiming Lin, Silvia L. Pintea, and Jan van Gemert. Deep vanishing point detection: Geometric priors make dataset variations vanish. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Shiyu Liu, Yuxuan Han, Pengwei Xing, Feng Yin, Rui Wang, Wei Cheng, Jing Liao, Ying Wang, Hongbo Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Chong Mou, Xintao Wang, Liangbin Xie, et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Dustin Podell, Zion English, Kyle Lacey, et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Yuxiang Tuo, Wang Xiang, Jun-Yan He, Yu Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023.

Chengyue Wu, Peng Zheng, Rui Yan, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Enze Xie, Junsong Chen, Jincheng Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, et al. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.