CUES AS COMPASSES: CROSS-MODAL CONDITION STEERING FOR UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models perform image understanding and generation within one architecture, yet architectural unification does not guarantee that each task uses its cross-modal condition effectively. Analyzing attention at the decisive stage of each task, we find a common conditioning gap in two task-specific forms: before answering, the model revisits the image but spreads its attention diffusely, whereas during generation, attention to the instruction decays early, leaving later updates weakly constrained. We propose Cross-Modal Condition Steering (CMCS), a training-free framework that contrasts each condition with a targeted reference to adaptively decide where, how strongly, and when to steer according to task demands. For understanding, CMCS redirects the model's existing visual attention toward instruction-relevant regions, and accepts the resulting answer change only to the extent that the reference evidence supports it. For generation, it reinjects, for every semantic unit, the residual between predictions under the original prompt and a counterfactual that alters that unit, shaped spatially at a fixed magnitude and scheduled by the unit's stage-dependent prior. Without updating any parameters, CMCS dynamically adapts to the current query and answer without degrading the original model performance, improving accuracy on visual understanding benchmarks and strengthening prompt adherence and compositional alignment on image generation benchmarks for both backbones, with further improvements in instruction adherence for image editing.

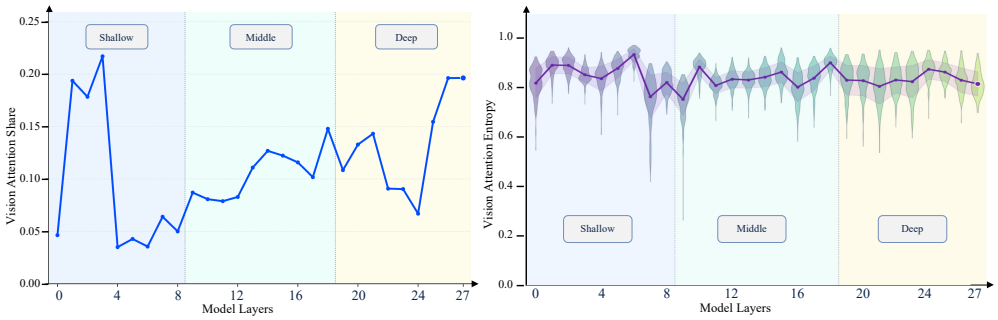
1 INTRODUCTION

Unified multimodal models answer questions about images, synthesize images from text, and edit them within a single architecture (Chen et al., 2025; Xie et al., 2026; Deng et al., 2025). Architectural unification, however, does not guarantee that a model makes effective use of its cross-modal conditions. In understanding, the answer must remain grounded in the image, and in generation, the evolving image must remain constrained by the instruction. Although the direction of conditioning reverses between the two tasks, the requirement is shared: task-relevant evidence from the conditioning modality must continue to govern prediction, even when the model's internal state already favors plausible alternatives.

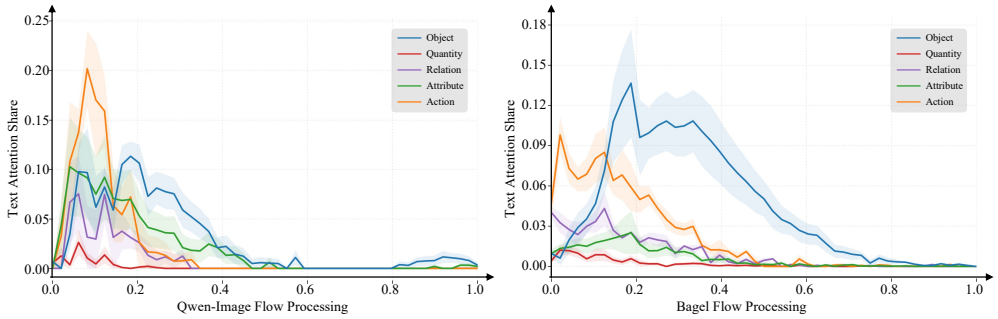
Evidence from task-specific models indicates that representing a condition is distinct from using it: vision-language models can encode visually grounded content that never surfaces in their answers (Li et al., 2025), and text-to-image diffusion models can neglect subjects or misbind attributes despite an explicit prompt (Chefer et al., 2023). For unified models, inference-time methods that rebalance attention to reference images or curate accumulated context improve generation without retraining (Wu et al., 2026; Chen et al., 2026). Condition utilization thus warrants study in its own right, and we ask: how does cross-modal conditioning fall short in understanding and in generation, and can a single steering principle correct both?

We address this question through attention diagnostics at the decisive stage of each task. For understanding, we examine BAGEL as a representative model (Figure 1a): the share of attention assigned to visual tokens rises sharply again in the final layers, just before answer prediction, indicating that the model returns to the image rather than abandoning visual evidence. Once renormalized within the visual tokens, however, this attention retains high entropy across all layers, so the model revisits

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.



(a) Understanding (BAGEL): visual attention share and entropy across layers



(b) Generation: text attention share over generation progress

Figure 1: Attention diagnostics of cross-modal conditioning in understanding (a) and generation (b). Visual revisiting remains diffuse in late layers, while attention to semantic units peaks early and decays. These reveal spatial and temporal conditioning gaps.

the image diffusely. The challenge is thus one of spatial selectivity: a substantial visual attention budget does not by itself ensure a focused inspection of instruction-relevant evidence.

For generation, we track how visual-token queries in Qwen-Image and BAGEL divide attention between the instruction and the evolving visual state along the flow-matching trajectory, grouping instruction tokens into semantic units by type (objects, quantities, relations, attributes, and actions). Attention to every type peaks early and then decays, becoming negligible over the latter part of sampling (Figure 1b), so later updates are computed with little direct attention to the text. This pattern highlights the challenge of temporal persistence: sustaining the influence of textual constraints as the image evolves. The timing of the peaks also varies across semantic types, suggesting that effective conditioning must account for both the persistence and the timing of textual guidance.

Taken together, the diagnostics reveal two manifestations of a common conditioning gap: understanding retains visual attention but distributes it diffusely, whereas generation attends progressively less to its textual constraints. The remedy therefore lies not in raising a modality's overall weight but in regulating where, how strongly, and when its evidence is used: uniform amplification cannot distinguish relevant regions from background, and a fixed enhancement coefficient cannot adapt to the evidence available in each example or generation state. Our key observation is that condition-specific evidence for both localization and strength can be obtained by contrast: when the original condition is compared with a targeted semantic modification of it, changes in attention reveal where the condition acts on visual content, and changes in prediction reveal how strongly to intervene.

Building on this observation, we propose Cross-Modal Condition Steering (CMCS), a training-free framework that derives its steering signals from comparisons between the original condition and task-specific references. For understanding, where spatial selectivity is lacking, neutral references replace question-relevant phrases with generic tokens and thereby reveal instruction-linked visual regions. CMCS redirects the existing visual revisit toward these regions and accepts the resulting answer change only to the extent supported by the reference evidence. For generation, where

temporal persistence is lacking, the difference between the velocity predictions with and without the instruction, which classifier-free guidance already computes, isolates the instruction's semantic residual. CMCS localizes this residual to the positions that attend to each semantic unit and reinjects it along the sampling trajectory, scaling each unit's share by a strength that adapts to the current state and follows the unit's stage-dependent prior. The two pathways thus share one principle, evidence-based condition steering, implemented adaptively through task-appropriate operators rather than an identical modification.

We implement CMCS on BAGEL (Deng et al., 2025), a unified understanding and generation model, and evaluate its transfer to Qwen-Image (Wu et al., 2025a). With all parameters frozen, we evaluate visual understanding, text-to-image generation, and instruction-guided editing, and compare CMCS with spatial controls that retain the intervention but remove or randomize its localization.

Our main contributions are as follows:

1. Through attention diagnostics, we identify a common conditioning gap in unified multimodal inference: diffuse visual revisiting in high-level decision layers during understanding, and rapidly decaying attention to textual constraints during generation.
2. We propose CMCS, a training-free, input-adaptive framework that contrasts each condition with a task-specific reference to determine where and how strongly to steer, and uses stage-dependent semantic priors to determine when to steer during generation.
3. Across visual understanding, text-to-image generation, and instruction-guided editing on two backbones, we isolate the contribution of condition-relevant localization and analyze where the approach helps and where it cannot.

2 Related Work

Multimodal large language models (MLLMs). MLLMs connect a pretrained vision encoder to a language model and learn to follow visual instructions (Li et al., 2023b; Liu et al., 2023a). Open families such as Qwen-VL and InternVL scale model size, resolution, and data (Bai et al., 2025b;a; Chen et al., 2024), and omni models such as GPT-4o are trained end-to-end across text, vision, and audio (OpenAI, 2024). Representing visual evidence does not guarantee its use: tokens tied to visual content can rank highly in intermediate layers yet fail to influence the final decoding (Li et al., 2025), and MLLMs often rely disproportionately on text late in reasoning.

Visual generative models. Diffusion models reverse a gradual noising process (Ho et al., 2020), typically in the latent space of an autoencoder (Rombach et al., 2022), and flow matching and rectified flow learn a velocity field that transports noise to data (Lipman et al., 2022; Liu et al., 2023b). Transformer denoisers have replaced U-Nets (Peebles & Xie, 2023), and multimodal diffusion transformers (MMDiT) with joint text–image attention underlie recent text-to-image models such as Qwen-Image (Esser et al., 2024; Wu et al., 2025a). Autoregressive models instead predict discrete visual tokens in raster order (Sun et al., 2024) or from coarse to fine scales (Tian et al., 2024). Prompt adherence remains a central challenge: models may neglect requested subjects or misbind attributes (Chefer et al., 2023), ELLA improves alignment through a trained timestep-aware connector (Hu et al., 2024), and benchmarks evaluate alignment at the object, compositional, and dense-prompt levels (Ghosh et al., 2023).

Unified multimodal models. Unified models perform understanding and generation in one framework and differ mainly in how generation is coupled to the language model (Deng et al., 2025). Early unified models predict discrete image tokens like text: Chameleon fuses the modalities early (Team, 2024), Janus and Janus-Pro decouple visual encoding for the two tasks (Wu et al., 2025b; Chen et al., 2025), TokenFlow and UniToken share visual tokenizers or encodings across them (Qu et al., 2025; Jiao et al., 2025), and Show-o generates the tokens by discrete diffusion (Xie et al., 2025). Recent models keep next-token prediction for text but generate images in a continuous latent space by diffusion or flow matching. In integrated models, one transformer predicts text and denoises image latents in the same sequence: Transfusion combines next-token prediction with continuous diffusion (Zhou et al., 2025), LMFusion adds image-specific parameters to a frozen language model (Shi et al., 2024), and BAGEL routes an interleaved sequence through understanding and generation experts with shared self-attention and generates images by rectified flow (Deng et al., 2025).

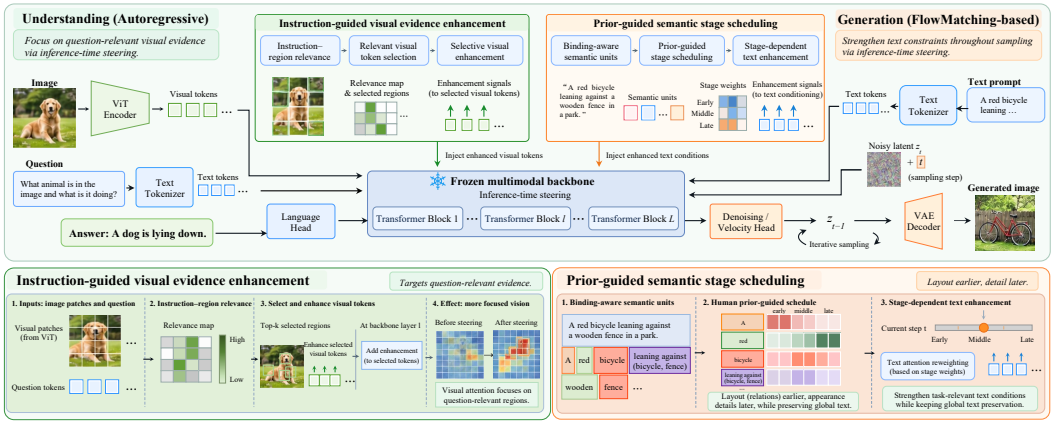


Figure 2: Overview of CMCS: instruction-guided visual evidence enhancement for understanding and stage-dependent semantic residual steering for generation, both applied to a frozen backbone.

In connected models, a multimodal language model encodes the condition for a separate generator: MetaQuery bridges the two with learnable queries (Pan et al., 2025), and InternVL-U and Qwen-Image condition MMDiT generators trained with flow matching on the hidden states of InternVL3.5 and Qwen2.5-VL, respectively (Tian et al., 2026; Wu et al., 2025a). The two tasks thus rely on different prediction mechanisms: an answer is decoded token by token, with the answer query reading the image through attention, whereas an image is generated over many sampling steps, with classifier-free guidance amplifying the instruction's contribution. These models further improve through interleaved pretraining (Deng et al., 2025), reasoning-centric data (Tian et al., 2026), and post-training that couples the two tasks (Su et al., 2026; Hao et al., 2026).

Training-free inference-time intervention. Training-free methods change how frozen models use their conditions at inference time. Visual contrastive decoding mitigates object hallucination in MLLMs by contrasting predictions for an original and a distorted image (Leng et al., 2024). In diffusion and flow models, classifier-free guidance scales the difference between conditional and unconditional predictions to control a condition's influence throughout sampling (Ho & Salimans, 2022), and SEGA adds concept-specific guidance directions with element-wise thresholding and a warm-up period (Brack et al., 2023). In language models, PASTA reweights selected attention heads toward user-specified text spans (Zhang et al., 2024b). Closest to our setting, training-free methods for frozen unified models rebalance attention to reference images in multi-image context generation (Wu et al., 2026) or curate the accumulated context in long-horizon interleaved generation (Chen et al., 2026). CMCS is likewise training-free but, rather than targeting a particular generation setting, acts on both prediction mechanisms of a frozen hybrid model: it redirects the answer query's attention over the image in understanding and reinjects the guidance residual along the sampling trajectory in generation, localized by attention to semantic units and scheduled by stage-dependent priors. We evaluate it on BAGEL and Qwen-Image, one integrated and one connected hybrid model.

3 METHOD

CMCS addresses the spatial and temporal deficiencies described above by separating where to steer from how strongly to steer. It identifies spatial targets through attention to the condition's semantic units and determines intervention strength from prediction differences against a reference. For generation, semantic phase priors further regulate when each condition is emphasized. All spatial maps and intervention strengths are computed from the current input, including the question or prompt and, for generation, the current sampling state. We first introduce the reference-based spatial evidence shared by both tasks (Section 3.1), followed by its use for understanding (Section 3.2) and generation (Section 3.3), as shown in Figure 2.

3.1 Reference-Based Spatial Evidence

A semantic unit u is an instruction span that expresses a single condition, namely an instruction-relevant noun phrase identified by the model's language backbone, and CMCS contrasts the original condition with a neutral or unconditional reference. For understanding, the reference, termed an anchor, replaces a unit's span with an equal-length sequence of *object* tokens, preserving the image, other tokens, and answer position, so the contrast isolates the contribution of the unit's content. For generation, the reference is the instruction-free branch of classifier-free guidance, whose difference from the instruction branch captures the contribution of the instruction. Neither construction requires a reference answer or an evaluation score.

Let $A^F_{h,i}$ and $A^R_{h,i}$ be the attention summaries of unit u for head h and visual position i at a candidate layer under the original condition and under the unit's anchor. In understanding, a summary averages attention from the unit's text queries to visual keys, whereas in generation it sums attention from visual query i to the unit's text keys in the instruction branch, so both index visual positions. With $\mathcal{N}$ denoting normalization over visual positions, $p^F_h = \mathcal{N}(A^F_h)$, and $p^R_h = \mathcal{N}(A^R_h)$, we compute

$$R_{h,i} = [\log(p^F_{h,i} + \epsilon) - \log(p^R_{h,i} + \epsilon)]_+, \qquad g = \mathcal{N}\left(\frac{1}{H}\sum_{h=1}^{H}\mathcal{N}(R_h)\right), \tag{1}$$

where $[x]_+ = \max(x, 0)$, H is the number of heads, and ϵ is a numerical stabilizer. In understanding, the log-ratio cancels attention that the question and its anchor allocate alike, such as generic saliency patterns, so g emphasizes the locations whose attention depends on the unit itself. In generation, the reference contains no unit tokens, so R_h is replaced by p^F_h and g is the unit's head-averaged attention map. Maps are ranked across candidate layers by a reliability score that combines spatial specificity, head agreement, cross-layer agreement, and a spatial-permutation reference (Appendix C).

3.2 Instruction-Guided Visual Evidence Enhancement

CMCS redirects the answer query's visual attention toward instruction-relevant evidence.

Adaptive semantic influence. Not every unit warrants steering, so CMCS first measures each unit's influence on the decision. Let z^F and z^R_u be the answer logits at the first answer position under the original condition and under the anchor of unit u, taken over the answer options for multiple-choice questions and over the vocabulary for open-ended ones. Centering by $C(x) = x - \mathrm{mean}(x)$, which removes shifts shared by all candidates, gives

$$r_u = C(z^F - z^R_u), \qquad B_u = \max r_u - \min r_u, \qquad I_u = \frac{B_u}{\max(B_u + \Delta z, \epsilon)}, \tag{2}$$

where Δz is the original top-two answer-logit margin, so a unit whose replacement barely changes the scores of the candidates receives little weight, while evidence that is substantial relative to the native decision margin receives greater weight.

Enhanced visual review. The review acts on the query that predicts the first answer token at the final decoder layer ($L = 27$ for both backbones), where BAGEL's visual attention rises again (Figure 1a). Its map $g^\star$ is the I_u-weighted average of the unit maps, sharpened by keeping the excess above the spatial mean plus one standard deviation, and its strength is $\beta = \max_u I_u$. If no unit has positive influence, CMCS returns the native answer. With $a_{h,i}$ the attention of head h from this query to visual key $i \in \mathrm{vis}$, $m_h = \sum_{i\in\mathrm{vis}} a_{h,i}$, and $p_{h,i} = a_{h,i}/m_h$, CMCS sets

$$p'_{h,i} = (1-\beta)p_{h,i} + \beta g^\star_i, \qquad a'_{h,i} = m_h p'_{h,i}, \tag{3}$$

leaving nonvisual weights unchanged. This adds $m_h \sum_{i\in\mathrm{vis}}(p'_{h,i} - p_{h,i})V_{h,i}$, with value vectors $V_{h,i}$, to the head output and redirects the diffuse revisit toward relevant regions.

Residual-supported decision. Redirected attention can also shift the answer logits in directions unrelated to the question's units, so CMCS keeps the centered change $\delta = C(z^{\mathrm{review}} - z^F)$ of the reviewed logits only as far as the I_u-weighted mean residual $\bar{r}$ supports it:

$$z' = z^F + \eta\delta, \qquad \eta_0 = \left[\frac{\langle\delta, \bar{r}\rangle}{\max(\|\delta\|_2^2, \epsilon)}\right]_{[0,1]}, \qquad \eta = \eta_0 \min\left(1, \frac{\|\bar{r}\|_2}{\max(\eta_0\|\delta\|_2, \epsilon)}\right), \tag{4}$$

where $[\cdot]_{[0,1]}$ clips to the unit interval. The answer maximizes z'. For open-ended questions, z' selects the first answer token, and the remaining tokens are decoded natively.

3.3 Stage-Dependent Semantic Residual Steering

For generation, CMCS reinforces the instruction during sampling by reinjecting the guidance residual at each unit's positions with a state- and stage-dependent strength.

Same-state residual. At each step t, classifier-free guidance evaluates the velocity model with and without the instruction at the current latent x_t, giving v_t^F and $v_t^\emptyset$, and combines them into the native velocity v_t. CMCS steers v_t with the semantic residual $r_t = v_t^F - v_t^\emptyset$, the direction that guidance amplifies (backbone conventions in Appendix C.6). Because both predictions share one state, r_t is, at the flow-matching optimum, proportional to the difference between the clean latents predicted with and without the instruction (Appendix B.2), isolating the instruction's contribution to the current update.

Stage-dependent injection. As with native guidance, strength factors are computed once per step from the residual's magnitude relative to the native velocity, its alignment with the native velocity, and its consistency with the preceding residual. Together with a fixed envelope $a_{\max}$ and the unit's phase prior, a Gaussian over sampling progress whose center and width depend on the unit's semantic type (Table 6), these factors set the strength $\alpha_{t,u}$ of each unit u of the prompt. CMCS localizes the residual with each unit's map, normalizes each localized share $\widehat{r}_{t,u}$ to the root mean square of v_t, and adds the shares to the native velocity, $v_t' = v_t + \sum_u \alpha_{t,u}\widehat{r}_{t,u}$, which the solver integrates (Appendix C.3). If no unit passes the map-reliability and residual tests, the native prediction is kept.

4 Experiments

Our experiments ask whether CMCS improves condition use across tasks and transfers from BAGEL to Qwen-Image, which capabilities benefit and where steering cannot help (Section 4.2), and whether the gains depend on condition-relevant localization (Section 4.3).

4.1 Experimental Setup

Models, benchmarks, and baselines. We apply CMCS to BAGEL-7B-MoT (Deng et al., 2025) and Qwen-Image (Wu et al., 2025a) (Appendix D.1), evaluating understanding for Qwen-Image with its Qwen2.5-VL component (Bai et al., 2025b). Understanding is evaluated on MME (Fu et al., 2026), SEEDBench-Img (Li et al., 2023a), MMBench (Liu et al., 2024), MathVista (Lu et al., 2024), and MMVP (Tong et al., 2024), which include multiple-choice and open-ended questions. Text-to-image generation is evaluated on GenEval and DPG-Bench (Ghosh et al., 2023; Hu et al., 2024), with controlled comparisons on the 800 compositional prompts of GenEval, and instruction-guided editing on the English and Chinese splits of GEdit-Bench (Liu et al., 2025). Baselines include the understanding-specialized InternVL2.5 (Chen et al., 2024) and the unified models Chameleon, Show-o, Emu3, Janus, Janus-Pro, UniToken, Ovis-U1, MetaQuery-XL, and TokenFlow-XL (Team, 2024; Xie et al., 2025; Wang et al., 2024; Wu et al., 2025b; Chen et al., 2025; Jiao et al., 2025; Wang et al., 2025; Pan et al., 2025; Qu et al., 2025), as well as Transfusion, GPT-4o, and OmniGen for generation (Zhou et al., 2025; OpenAI, 2024; Xiao et al., 2024) and GPT-4o, OmniGen, and Gemini 2.0 for editing.

Metrics and settings. Evaluation follows GenEditEvalKit (Tian et al., 2026) and LMMs-Eval (Zhang et al., 2024a). Every comparison among native inference, CMCS, and its spatial controls fixes the checkpoint, input processing, decoding, and evaluator, and the matched GenEval comparisons also share the initial noise, a 512×512 resolution, and 50 sampling steps. Unit identification and reference construction use neither benchmark labels nor evaluator outputs. Understanding benchmarks report accuracy (%), except MME-P (perception) and MME-S (perception and cognition), which keep their native scales. GenEval (0–1 scale) and DPG-Bench (%) are evaluated on their full prompt sets with their official protocols. On the matched GenEval comparisons, Soft-TIFA scores each semantic atom by the probability that a locally deployed Qwen3.8-27B judge

Table 1: Visual understanding. Bold marks the better result in each native–CMCS pair, and † marks native results from our local evaluation.

Model	MME-P	MME-S	SEEDBench-Img	MMBench	MathVista	MMVP
Baseline models						
InternVL2.5-8B	–	2344	–	84.6	64.4	–
Chameleon	–	–	47.4	35.7	13.6	–
Show-o	1097.2	–	–	–	–	–
Emu3	1244	–	68.2	58.5	–	36.6
Janus	1338	–	63.7	69.4	–	–
UniToken-StageIII	–	–	69.9	71.1	38.5	–
Janus-Pro-1B	1444	–	68.3	75.5	–	–
Janus-Pro-7B	1567.1	–	72.1	79.2	–	–
Ovis-U1	–	–	–	77.8	69.4	–
MetaQuery-XL	1685	–	–	83.5	–	–
TokenFlow-XL	1546	–	–	68.9	–	–
Evaluated backbones						
BAGEL	**1687**	**2388**	78.5	87.8	73.1	69.3
+ CMCS	1681	2379	**79.1**	**89.0**	**73.4**	**71.7**
Qwen-Image†	1635	**2347**	**76.9**	82.0	68.0	62.5
+ CMCS	**1648**	2341	76.5	**83.9**	**68.3**	**62.9**

(Qwen Team, 2026) assigns to acceptable answers, and averages atom scores within each prompt by an arithmetic mean (AM), which permits compensation across atoms, or a geometric mean (GM), which is dominated by the weakest atom and thus emphasizes prompt-level completeness. Editing reports semantic consistency (SC), perceptual quality (PQ), and overall score (O) on a 0–10 scale.

4.2 Main Results

Understanding gains concentrate where the question names localized evidence. The largest understanding gains arise on MMBench (1.2 and 1.9 points for BAGEL and Qwen-Image) and on BAGEL's MMVP (2.4 points, Table 1). Both benchmarks pose questions that name the object, attribute, or relation to inspect, and the answer typically depends on a localized region: MMBench probes abilities such as attribute recognition, spatial relations, and object localization (Liu et al., 2024), and MMVP pairs images that CLIP-style encoders embed similarly, so each answer hinges on a detail such as orientation, state, color, or count (Tong et al., 2024). A diffuse revisit (Figure 1a) dilutes such details among irrelevant regions, whereas concentrating it on the queried region lets them inform the answer. The strengthened capability is thus instruction-conditioned visual grounding, which the gains on both backbones show is not specific to one architecture.

Gains are small where reasoning, not evidence selection, is the bottleneck. MathVista improves by 0.3 points on both backbones, suggesting that CMCS better highlights useful visual evidence, but correctness there largely depends on multi-step reasoning and precise value reading (Lu et al., 2024), which CMCS does not modify, so the gains are consistent but modest.

Little change is expected where perception or recognition is the bottleneck. MME and SEEDBench-Img change little in either direction. MME's yes/no questions largely test low-level perception and recognition, such as existence, count, color, text, and celebrities or landmarks, alongside cognition subtasks such as numerical calculation and code reasoning (Fu et al., 2026). Success depends mainly on what the visual encoder represents and on knowledge acquired in training, which CMCS cannot supply because it neither trains the model nor modifies the encoder. The largest SEEDBench-Img dimensions, such as scene understanding and instance counting, require evidence spread across the image rather than in a named region (Li et al., 2023a). The relative maps then provide weak evidence, and CMCS keeps the native answer when no unit has positive influence.

Generation gains align with sustained textual guidance. On the full GenEval and DPG-Bench benchmarks (Table 2), CMCS raises DPG-Bench scores, whose dense prompts specify many entities, attributes, and relations (Hu et al., 2024), by 0.60 points for BAGEL and 1.16 points for Qwen-Image. GenEval changes by only $+0.02$ for Qwen-Image and -0.01 for BAGEL, as its short, object-centric prompts leave little headroom for backbones that already exceed 0.8.

Table 2: Text-to-image generation. Bold marks the better of each pair.

Model	GenEval	DPG-Bench
Baseline models		
GPT-4o	0.84	85.15
OmniGen	0.68	81.16
Chameleon	0.39	–
Show-o	0.68	–
Emu3	0.54	80.60
Janus	0.61	79.68
TransFusion	0.63	–
UniToken-StageIII	0.63	–
Janus-Pro-1B	0.73	82.63
Janus-Pro-7B	0.80	84.19
Ovis-U1	0.89	83.72
Evaluated backbones		
BAGEL	**0.82**	85.07
+ CMCS	0.81	**85.67**
Qwen-Image	0.87	88.32
+ CMCS	**0.89**	**89.48**

Table 3: Image editing on GEdit-Bench. Bold marks the better of each pair.

	GEdit-Bench					
	English			Chinese		
Model	SC	PQ	O	SC	PQ	O
Baseline models						
Gemini 2.0	6.73	6.61	6.32	5.43	6.78	5.36
GPT-4o	7.85	7.62	7.53	7.67	7.56	7.30
OmniGen	5.96	5.89	5.06	–	–	–
Evaluated backbones						
BAGEL	7.36	6.83	6.52	7.34	**6.85**	6.50
+ CMCS	**7.54**	**7.42**	**6.91**	**7.52**	6.82	**6.57**
Qwen-Image	7.69	**7.62**	**7.63**	**7.68**	7.48	7.50
+ CMCS	**7.71**	7.51	7.61	7.62	**7.54**	**7.58**

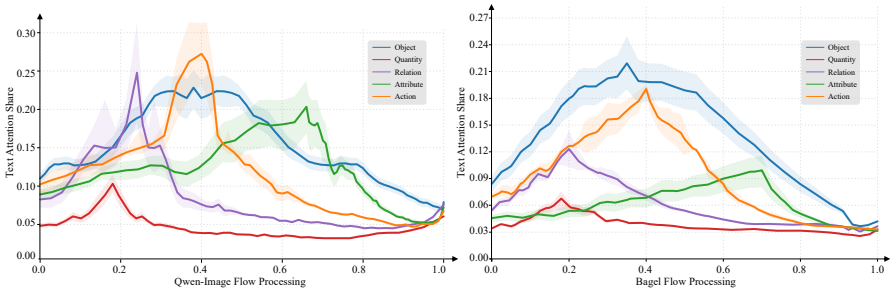


Figure 3: Text attention under CMCS for Qwen-Image (left) and BAGEL (right), measured as in Figure 1b on the same 200 GenEval samples.

Steering sustains attention to the instruction and staggers it by unit type. We repeat the diagnostic of Figure 1b with CMCS applied (Figure 3) and find two differences from native sampling. First, attention to the instruction no longer vanishes: every unit type retains a share of at least 0.025 until the final step, and the object share at 80% of sampling remains about 0.12 for Qwen-Image and 0.08 for BAGEL. Relative gains are largest for weakly attended types: on BAGEL, the peak shares of quantity, relation, and attribute units rise from at most about 0.04 to 0.07–0.12. Second, under the phase priors, the peaks are staggered by type. Later updates are thus computed with sustained attention to the instruction, and each constraint type is emphasized at the stage its prior assigns.

Editing benefits when instruction adherence is the bottleneck. On GEdit-Bench (Table 3), CMCS raises BAGEL's semantic consistency by 0.18 points in both English and Chinese, so requested changes are executed more faithfully regardless of the instruction language. Its overall score rises by 0.39 points in English and 0.07 in Chinese, a difference that stems from perceptual quality, which increases markedly in English but decreases slightly in Chinese. Qwen-Image gains approximately 0.11 points across all metrics. Building on its baseline semantic consistency scores of 7.69 in English and 7.68 in Chinese, it approaches or matches GPT-4o (7.85 and 7.67, respectively).

4.3 Ablation Analysis

Where the revisit is directed matters more than how concentrated it is. We compare CMCS with two review controls (Table 4): uniform enhancement, which removes spatial selectivity, and random routing, which keeps the map values but breaks their correspondence with image locations. CMCS achieves the highest accuracy in every setting. On BAGEL's MMBench, random routing leaves accuracy at the native level (87.86 versus 87.83), so concentrating the revisit without regard to the question does not help, whereas routing toward instruction-relevant regions yields a gain

Table 4: Understanding ablations (accuracy).

	MMBench		MathVista	
Method	BAGEL	Qwen-Image	BAGEL	Qwen-Image
Native	87.83	82.00	73.1	68.0
Uniform	82.59	82.67	72.6	68.1
Random	87.86	82.97	72.8	68.1
CMCS	**88.99**	**83.93**	**73.4**	**68.3**

Table 5: Generation ablations (Soft-TIFA).

Method	BAGEL		Qwen-Image	
	AM	GM	AM	GM
Native	66.20	19.10	81.00	35.10
Uniform	66.21	19.45	81.57	35.91
Random	66.09	18.99	81.26	35.59
CMCS	**66.38**	**19.60**	**81.87**	**36.59**

(88.99). Uniform enhancement lowers accuracy by 5.24 points, suggesting that BAGEL's native revisit, although diffuse, carries spatial structure that flattening erases and that CMCS preserves by mixing the native allocation with the routed map (Equation (3)). On Qwen-Image, both controls slightly exceed native inference, so improvement over native inference alone is insufficient evidence of localization, but CMCS still exceeds the stronger control by 0.96 points on MMBench and 0.2 on MathVista. The comparison thus evaluates spatial selection within the complete review procedure.

Matched GenEval gains concentrate on weakly satisfied constraints. Under matched noise and evaluation (Table 5), CMCS improves AM and GM on both backbones, with larger relative gains for GM (2.6% for BAGEL and 4.2% for Qwen-Image) than for AM (0.3% and 1.1%). Because GM is dominated by the lowest-scoring atom in each prompt, this pattern suggests that the gains come disproportionately from weakly satisfied constraints, consistent with the phase priors, which emphasize quantity and relation units early in sampling (Table 6 and Figure 3), the constraint types on which text-to-image models commonly fail (Ghosh et al., 2023). The strengthened capability is thus compositional constraint satisfaction. The larger gains on Qwen-Image are consistent with Figure 1b, in which its text attention vanishes earlier than BAGEL's, leaving more room for sustained guidance, and the persistent gap between AM and GM (66.38 versus 19.60 for BAGEL) indicates substantial room for improvement in atomic constraint satisfaction.

Spatial shaping must follow the semantic condition. The generation controls in Table 5 retain semantic residual steering but remove spatial shaping (uniform) or randomize its assignment (random), with each unit's injected magnitude fixed at a given state and strength (Equation (22)). Uniform enhancement already improves GM on both backbones (by 0.35 and 0.81 points), so sustaining the semantic residual counteracts the decay of textual guidance even without localization, and spatial shaping adds further gains, most clearly on Qwen-Image (0.68 GM points over uniform enhancement, compared with 0.15 for BAGEL). Random assignment falls below uniform enhancement on both backbones and below native inference on BAGEL. Injecting a unit's share of the residual into unrelated regions is thus worse than spreading it evenly, plausibly because it can realize the constraint on the wrong object or disturb content that should remain fixed. Spatial variation alone is thus insufficient, and the benefit depends on matching the semantic condition.

5 Conclusion and Limitations

Conclusion. We introduced CMCS, a training-free framework that contrasts each cross-modal condition with a reference and localizes the result with attention to semantic units to decide where, how strongly, and when to steer a unified multimodal model, while bounding the injected magnitude in generation. Its gains concentrate on questions that name localized evidence and on prompts with dense or compositional constraints, spatial controls tie them to condition-relevant localization, and attention statistics show that generation steering keeps the instruction attended throughout sampling, with each unit type emphasized at the stage its prior assigns. These results position inference-time condition utilization as a direction complementary to architectural unification and model training.

Limitations. CMCS improves how a model uses evidence it already represents, but cannot add perceptual features or knowledge that the backbone lacks, which bounds its gains on perception-limited benchmarks. Its signals can also mislead: an inaccurate relative map can misdirect the review, and in generation an imprecise unit map or a mistimed phase prior can reinforce unintended regions or act too late. Map-reliability tests and evidence-dependent strength control mitigate these risks but cannot eliminate them, as the model's own representations may themselves be misleading.

AI USE STATEMENT

Generative AI assisted with the following tasks: polishing the text, correcting grammatical errors, and aiding in code optimization. The numerical results reported in the paper are derived from existing experimental records. The authors are responsible for verifying methodological details, references, data sources, and final conclusions prior to submission.

ETHICS STATEMENT

Stronger condition adherence can improve the reliability of useful multimodal applications, but may also make image editing and generation more effective for deceptive or harmful requests. Inference-time steering should not be treated as a safety mechanism or used to bypass a backbone's safeguards. Improved aggregate benchmark performance does not guarantee factual reliability, fairness, or appropriate handling of sensitive content. The present evaluation does not establish these properties.

REPRODUCIBILITY STATEMENT

We define the steering operators (Section 3), give pseudocode for both pathways (Appendix A), and derive each operator from the native computation of the backbone (Appendix B). We further specify the relative maps, decision rule, phase profiles, numerical conventions, and experimental configuration (Appendix C) and consolidate all settings in one table (Table 7). Both backbones are publicly released checkpoints evaluated with frozen weights (Appendix D.1), and we describe the evaluation protocols, including the Soft-TIFA judge used for the matched GenEval comparisons (Appendix D.2).

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Haoyu Chen, Qing Liu, Yuqian Zhou, He Zhang, Zhaowen Wang, Mengwei Ren, Jingjing Ren, Xiang Wang, Zhe Lin, and Lei Zhu. How long can unified multimodal models generate images reliably? taming long-horizon interleaved image generation via context curation. *arXiv preprint arXiv:2603.07540*, 2026.

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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Ke Hao, Yuanzhi Liang, Tingxi Chen, Rui Li, Haibin Huang, Chi Zhang, Yun Gu, and Xuelong Li. Dreaming in flow: Generative grounding feedback for self-evolving unified multimodal models. *arXiv preprint arXiv:2609.08282*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. *arXiv preprint arXiv:2504.04423*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023b.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2545–2555, 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. LMFusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*, 2026.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qi Wei, Jinhui Yin, Xue Yang, Zhihang Zhong, Qi Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, Gen Luo, Hongsheng Li, Yu Qiao, Kai Chen, and Hongjie Zhang. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

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, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025b.

Mingrui Wu, Hang Liu, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Micon-bench: Benchmarking and enhancing multi-image context image generation in unified multimodal models. *arXiv preprint arXiv:2602.19497*, 2026.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024a. arXiv:2407.12772.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *International Conference on Learning Representations*, 2024b.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

A ALGORITHMS

We summarize the understanding and generation procedures used in all experiments (Algorithms 1 and 2). Both compute unit maps with Equation (1) at every candidate layer and keep the layer with the highest reliability in Equation (18) (Appendix C). Model parameters remain frozen throughout, and no step uses benchmark labels or evaluator outputs.

Algorithm 1 CMCS for visual understanding (instruction-guided visual evidence enhancement).

Require: Image $\mathcal{I}$, question x, answer candidates $\mathcal{Y}$ (the options, or the vocabulary for open-ended questions), frozen model $\mathcal{M}$, anchor token `object`, candidate layers $\mathcal{L}$, review layer L

Ensure: Answer $\hat{y} \in \mathcal{Y}$; for open-ended questions, the first answer token

$z^F \leftarrow$ logits of $\mathcal{M}(\mathcal{I}, x)$ over $\mathcal{Y}$ at the first answer position; $\Delta z \leftarrow$ top-two margin of z^F
$\mathcal{U} \leftarrow$ question-relevant noun phrases identified by the model's language backbone
for each unit $u \in \mathcal{U}$ **do**
    $x_u^R \leftarrow x$ with the span of u replaced by the anchor token, repeated to the same length
    $z_u^R \leftarrow$ logits of $\mathcal{M}(\mathcal{I}, x_u^R)$ over $\mathcal{Y}$; $r_u \leftarrow C(z^F - z_u^R)$
    Compute B_u and I_u (Equation (2))
    $g_{\ell,u} \leftarrow$ relative map of u against its anchor at each layer $\ell \in \mathcal{L}$ (Equation (1))
    $\ell_u^\star \leftarrow$ layer with the highest reliability q_ℓ (Equation (18)); $g_u \leftarrow g_{\ell_u^\star,u}$
end for
$W \leftarrow \sum_u I_u$
if $W = 0$ **then**
    return $\arg\max_y z_y^F$ ▷ no unit has positive influence
end if
$g_{\text{agg}} \leftarrow \mathcal{N}\big(\sum_u I_u g_u\big)$; $g^\star \leftarrow \mathcal{N}\big([g_{\text{agg}} - \text{mean}(g_{\text{agg}}) - \text{std}(g_{\text{agg}})]_+\big)$, or g_{agg} if empty
$\beta \leftarrow \max_u I_u$; $\bar{r} \leftarrow W^{-1} \sum_u I_u r_u$
for each head h at layer L, for the answer-prediction query **do**
    $m_h \leftarrow \sum_{i \in \text{vis}} a_{h,i}$; $p_{h,i} \leftarrow a_{h,i}/m_h$; $p'_{h,i} \leftarrow (1-\beta)p_{h,i} + \beta g_i^\star$
    Add $m_h \sum_{i \in \text{vis}} (p'_{h,i} - p_{h,i}) V_{h,i}$ to the head output ▷ Equation (3)
end for
$z^{\text{review}} \leftarrow$ reviewed logits over $\mathcal{Y}$; $\delta \leftarrow C(z^{\text{review}} - z^F)$
$\eta \leftarrow$ Equation (4)
return $\arg\max_y (z^F + \eta\delta)_y$ ▷ open-ended answers continue with native decoding

Inference cost. Let $|\mathcal{U}|$ be the number of semantic units in a question or prompt. Understanding evaluates the prompt $1 + |\mathcal{U}|$ times (the original condition and one anchor per unit), reads its attention summaries from these evaluations, and adds one review evaluation: a cached answer-query evaluation for BAGEL and a full forward pass for Qwen-Image (Appendix C.2). Generation adds no model evaluations: its reference is the instruction-free branch that native classifier-free guidance (CFG) already evaluates at every step, for Qwen-Image and for BAGEL's text guidance. Its overhead consists of reading unit attention from the instruction branch at the candidate layers, using at most 256 sampled visual queries per layer, and of the permutation nulls, which operate on the resulting maps and require no further model evaluations.

Algorithm 2 CMCS for flow-based generation (stage-dependent semantic residual steering). For Qwen-Image, r_t and v_t follow the conditional-proxy convention of Appendix C.6, and native renormalization is applied before the solver step.

Require: Prompt c; semantic units $\mathcal{U}$ with phase profiles (μ_u, σ_u); frozen velocity model v_θ with native classifier-free guidance; descending time grid $\xi_0 > \cdots > \xi_K$; initial noise x_0; candidate layers $\mathcal{L}$; envelope $a_{\max}$; acceptance thresholds, including $p_{\max}$ and $\kappa_{\min}$ (Table 7)

Ensure: Final latent x_K

$r_{\text{prev}} \leftarrow \emptyset$; $g_{\text{acc},u} \leftarrow \emptyset$ for every $u \in \mathcal{U}$
for $t = 0, \ldots, K-1$ **do**
  $\tau_t \leftarrow t/49$ ▷ phase coordinate for 50 nominal steps
  $v_t^F \leftarrow v_\theta(x_t, \xi_t, c)$; $v_t^\emptyset \leftarrow v_\theta(x_t, \xi_t, \emptyset)$ ▷ native guidance branches; record unit attention
  $v_t \leftarrow$ native guided velocity; $r_t \leftarrow v_t^F - v_t^\emptyset$; $v_t' \leftarrow v_t$ ▷ residual shared by all units
  Compute ρ_t, π_t, s_t (Equation (20)), with $s_t = 1$ if $r_{\text{prev}} = \emptyset$
  if r_t passes the projection and residual-cosine tests **then**
    for each unit $u \in \mathcal{U}$ **do**
      $g_{\ell,t,u} \leftarrow$ instruction-branch map of u at each layer $\ell \in \mathcal{L}$ (Equation (1), $R_h = p_h^F$)
      $\ell^\star \leftarrow$ layer with the highest reliability q_ℓ (Equation (18)); $g_{t,u} \leftarrow g_{\ell^\star,t,u}$; $\hat{p} \leftarrow \hat{p}_{\ell^\star}$
      map_ok $\leftarrow \big(\hat{p} \le p_{\max}\big) \wedge \big(\psi_{\ell^\star} > \psi^{\text{null}}_{\ell^\star,95}\big)$
      map_ok $\leftarrow$ map_ok $\wedge \big(g_{\text{acc},u} = \emptyset \vee \cos(g_{t,u}, g_{\text{acc},u}) \ge \kappa_{\min}\big)$
      if map_ok **then**
        $\alpha_{t,u} \leftarrow a_{\max}\, \phi_u(\tau_t)\, \rho_t\, \pi_t\, s_t$ ▷ Equation (21)
        $G_{t,u} \leftarrow g_{t,u} / \operatorname{RMS}(g_{t,u})$
        $\widehat{r}_{t,u} \leftarrow \operatorname{RMS}(v_t)\, (r_t \odot G_{t,u}) / \operatorname{RMS}(r_t \odot G_{t,u})$
        $v_t' \leftarrow v_t' + \alpha_{t,u} \widehat{r}_{t,u}$; $g_{\text{acc},u} \leftarrow g_{t,u}$ ▷ Equation (22)
      end if
    end for
  end if
  $r_{\text{prev}} \leftarrow r_t$ ▷ stability uses consecutive residuals, even after a rejection
  $x_{t+1} \leftarrow \textsc{SolverStep}(x_t, v_t', \xi_t, \xi_{t+1})$ ▷ native scheduler and update convention
end for
return x_K

B Derivations

This section derives both pathways from the native computation of each backbone and marks the stages at which CMCS intervenes. Native stages are unchanged by the method, and stages labeled CMCS are the only points of intervention. The derivations use the definitions of Section 3 and Appendix C, including the stabilizer ϵ.

B.1 Understanding: From Attention to the Answer Decision

Step 1 (native): attention of the answer query. At a decoder layer, the answer-prediction query of head h attends over visual keys vis and text keys txt:

$$a_{h,j} = \frac{\exp(\mathbf{q}_h^\top \mathbf{k}_{h,j}/\sqrt{d_{\text{head}}})}{\sum_{j' \in \text{vis}\cup\text{txt}} \exp(\mathbf{q}_h^\top \mathbf{k}_{h,j'}/\sqrt{d_{\text{head}}})}, \qquad o_h = \sum_{j \in \text{vis}\cup\text{txt}} a_{h,j} V_{h,j}, \tag{5}$$

where $\mathbf{q}_h$ and $\mathbf{k}_{h,j}$ are the query and key vectors, d_{head} is the head dimension, and o_h is the head output before the output projection. The remaining computation maps the head outputs to the answer logits z at the first answer position, over the answer options or, for open-ended questions, the vocabulary.

Step 2 (native): how much versus where. Let $m_h = \sum_{i \in \text{vis}} a_{h,i}$ and, for $m_h > 0$, $p_{h,i} = a_{h,i}/m_h$. Then

$$o_h = \sum_{j \in \text{txt}} a_{h,j} V_{h,j} + m_h \sum_{i \in \text{vis}} p_{h,i} V_{h,i}. \tag{6}$$

The mass m_h determines how strongly the head reads the image, and the distribution p_h determines where it reads. We report these two quantities across layers, the visual share m_h and the normalized entropy of p_h (Figure 1a). Diffuse revisiting is a property of p_h at substantial m_h, so the intervention below acts on p_h.

Step 3 (CMCS): reference evaluations. For each unit u, the model is evaluated again with the span of u replaced by its anchor (Section 3.1). The logits give the centered residual $r_u = C(z^F - z_u^R)$ and the influence I_u in Equation (2), and the attention from the unit's text queries to visual keys gives the relative map in Equation (1) and, after aggregation, the review map $g^\star$. Because $B_u \geq 0$ and the native margin satisfies $\Delta z \geq 0$,

$$0 \leq I_u = \frac{B_u}{\max(B_u + \Delta z, \epsilon)} \leq 1, \tag{7}$$

so the review strength $\beta = \max_u I_u$ lies in $[0, 1]$.

Step 4 (CMCS): visual review. At the review layer L, CMCS keeps the text terms and m_h in Equation (6) and replaces p_h by $p'_h = (1-\beta)p_h + \beta g^\star$ (Equation (3)). Substituting,

$$o'_h = o_h + m_h \sum_{i \in \text{vis}} (p'_{h,i} - p_{h,i}) V_{h,i} = o_h + \beta\, m_h \sum_{i \in \text{vis}} (g^\star_i - p_{h,i}) V_{h,i}, \tag{8}$$

which is the head-output correction stated in Section 3.2. Since $\beta \in [0, 1]$ and p_h and $g^\star$ are distributions, p'_h is a distribution. Hence $\sum_{i \in \text{vis}} a'_{h,i} = m_h \sum_i p'_{h,i} = m_h$, the text weights are unchanged, and the modified weights still sum to one. The correction vanishes when $g^\star = p_h$. Propagating o'_h through the remaining computation gives the reviewed logits z^{review}.

Step 5 (CMCS): residual-supported decision. The softmax over candidates is invariant to adding a constant to every logit, so only centered logits affect the decision. The method therefore compares the centered review change $\delta = C(z^{\text{review}} - z^F)$ with the centered residual $\bar{r}$. The coefficient along δ that best explains $\bar{r}$ in the least-squares sense satisfies

$$\frac{d}{d\kappa} \|\bar{r} - \kappa\delta\|_2^2 = -2\langle \delta, \bar{r} - \kappa\delta \rangle = 0 \quad \Longrightarrow \quad \kappa^\star = \frac{\langle \delta, \bar{r} \rangle}{\|\delta\|_2^2}. \tag{9}$$

The decision rule stabilizes this denominator by ϵ, restricts $\kappa^\star$ to $[0, 1]$, and caps the norm of the accepted change (Equation (4)). For $\kappa^\star \in [0, 1]$, the change before the cap is $\kappa^\star\delta$, the projection of $\bar{r}$ onto δ. Three consequences follow. First, $0 \leq \eta \leq \eta_0 \leq 1$, so the accepted change $z' - z^F = \eta\delta$ neither reverses nor amplifies the review change. Second, if $\langle \delta, \bar{r} \rangle \leq 0$, then $\eta_0 = 0$ and $z' = z^F$: an unsupported review direction cannot change the decision. Third, for $\eta_0 > 0$,

$$\|z' - z^F\|_2 = \eta_0 \|\delta\|_2 \min\left(1, \frac{\|\bar{r}\|_2}{\max(\eta_0\|\delta\|_2, \epsilon)}\right) \leq \|\bar{r}\|_2 \frac{\eta_0\|\delta\|_2}{\max(\eta_0\|\delta\|_2, \epsilon)} \leq \|\bar{r}\|_2, \tag{10}$$

so the accepted change is bounded by the measured semantic residual.

B.2 Generation: From Flow Matching to the Steered Update

Step 1 (native): flow-matching path and optimal velocity. Both generators are flow-matching models (Lipman et al., 2022; Liu et al., 2023b). In the convention of their samplers (Appendix C), a clean latent y and Gaussian noise ε are connected by

$$x_\xi = (1 - \xi)\, y + \xi\varepsilon, \qquad \frac{dx_\xi}{d\xi} = \varepsilon - y, \qquad \xi \in [0, 1], \tag{11}$$

and the velocity model is trained by minimizing $\mathbb{E}\, \|v_\theta(x_\xi, \xi, c) - (\varepsilon - y)\|_2^2$ over latents, noise, time, and conditions c. The minimizer is the conditional mean of the target. Substituting $\varepsilon = (x_\xi - (1-\xi)y)/\xi$ gives $\varepsilon - y = (x_\xi - y)/\xi$, so

$$v^\star(x, \xi, c) = \mathbb{E}[\varepsilon - y \mid x_\xi = x, c] = \frac{x - \hat{y}(x, \xi, c)}{\xi}, \qquad \hat{y}(x, \xi, c) = \mathbb{E}[y \mid x_\xi = x, c]. \tag{12}$$

Step 2 (native): sampling with guidance. Sampling integrates $dx/d\xi = v$ from $\xi_0 = 1$ toward zero on a descending grid $\xi_0 > \xi_1 > \cdots > \xi_K$. With $\Delta t_t = \xi_t - \xi_{t+1} > 0$, the Euler update is

$$x_{t+1} = x_t - \Delta t_t \, v_t, \tag{13}$$

where v_t is the backbone's native prediction at (x_t, ξ_t). Under classifier-free guidance with scale γ, $v_t = v_\theta(x_t, \xi_t, \emptyset) + \gamma\big(v_\theta(x_t, \xi_t, c) - v_\theta(x_t, \xi_t, \emptyset)\big)$. BAGEL additionally applies its text and image scales and renormalization (Appendix C.6). The condition affects the trajectory only through v_t, whose visual queries attend progressively less to the text as sampling proceeds (Figure 1b).

Step 3 (CMCS): semantic residual at the current state. At step t, classifier-free guidance evaluates the velocity model with the instruction c and without it at the same x_t and ξ_t, and CMCS forms $r_t = v_t^F - v_t^\emptyset$ from these two evaluations. For a velocity model at the optimum of Equation (12), the state terms cancel:

$$r_t = -\frac{1}{\xi_t}\Big(\hat{y}(x_t, \xi_t, c) - \hat{y}(x_t, \xi_t, \emptyset)\Big). \tag{14}$$

The residual is thus proportional to the difference between the clean latents predicted with and without the instruction, that is, the instruction's contribution to the current update. It is also the direction that guidance amplifies: without renormalization, $v_t = v_t^\emptyset + \gamma r_t$. When the residual is formed from guidance-scaled predictions, as in the Qwen-Image proxy below, it carries the factor γ. The cancellation requires both predictions at the same state, which native guidance already provides.

Step 4 (CMCS): spatially shaped, stage-dependent injection. The steered prediction $v_t' = v_t + \sum_u \alpha_{t,u} \widehat{r}_{t,u}$ (Equation (22)) replaces v_t in Equation (13), so the steered state is $x_{t+1}' = x_{t+1} - \Delta t_t \sum_u \alpha_{t,u} \widehat{r}_{t,u}$. Substituting $\widehat{r}_{t,u}$ from Equation (22) and Equation (14), with the sum over units for which $\mathrm{RMS}(r_t \odot G_{t,u}) > 0$,

$$\begin{aligned} x_{t+1}' - x_{t+1} &= \Big(\sum_u \lambda_{t,u} \, G_{t,u}\Big) \odot \Big(\hat{y}(x_t, \xi_t, c) - \hat{y}(x_t, \xi_t, \emptyset)\Big), \\ \lambda_{t,u} &= \frac{\Delta t_t \, \alpha_{t,u} \, \mathrm{RMS}(v_t)}{\xi_t \, \mathrm{RMS}(r_t \odot G_{t,u})} \geq 0. \end{aligned} \tag{15}$$

Each step therefore moves the latent further toward the instruction-conditioned clean prediction, with a nonnegative spatial weight $\sum_u \lambda_{t,u} G_{t,u}$ that concentrates the additional guidance at the positions of the units emphasized at that stage. The strengths $\alpha_{t,u} = a_{\max} \phi_u(\tau_t) \rho_t \pi_t s_t$ (Equation (21)) carry the stage dependence through $\phi_u(\tau_t)$ with $\tau_t = t/49$, which is how CMCS sustains each textual constraint after native attention to it has decayed.

Step 5 (CMCS): magnitude of the injection. RMS is positively homogeneous, so $\mathrm{RMS}(\widehat{r}_{t,u}) = \mathrm{RMS}(v_t)$ by Equation (22) for every map $g_{t,u}$, including permuted maps, so each unit's correction has RMS exactly $\alpha_{t,u} \, \mathrm{RMS}(v_t)$. Because each $\widehat{r}_{t,u}$ has the same dimension as v_t and RMS is a scaled Euclidean norm, the triangle inequality gives

$$\mathrm{RMS}(v_t' - v_t) \leq \sum_u \alpha_{t,u} \, \mathrm{RMS}(v_t), \qquad \|x_{t+1}' - x_{t+1}\|_2 \leq \sum_u \alpha_{t,u} \, \|\Delta t_t \, v_t\|_2, \tag{16}$$

with equality when at most one unit is eligible. In Equation (20), $\rho_t \leq 1$ and $\pi_t, s_t \in [0, 1]$. In Equation (23), the bracket lies in $[f, 1]$ with $f \in [0, 0.25]$, $(1 + \zeta_u)/2 \in [1/2, 1]$, and $d_u \in (0, 1]$ combines scope and dependency factors in $(0, 1]$ that attenuate the prior (Appendix C.4). Hence $\phi_u(\tau) \in [0, 1]$ and $\alpha_{t,u} \in [0, a_{\max}]$. Each unit's correction is thus at most $a_{\max}$ times the native step length, the whole step departs from the native one by at most $\sum_u \alpha_{t,u} \leq |\mathcal{U}| \, a_{\max}$ times its length, and the uniform and random controls inject the same per-unit magnitude as CMCS at a common state.

Qwen-Image insertion. For Qwen-Image, the correction is inserted into the conditional prediction before guidance, with $r_t = \gamma(v_{\mathrm{cond},t}^F - v_t^\emptyset)$ and $v_t = \gamma v_{\mathrm{cond},t}^F$, where $v_t^\emptyset = v_\theta(x_t, \xi_t, \emptyset)$ is the unconditional branch of native guidance (Appendix C.6). Adding $\sum_u \alpha_{t,u} \widehat{r}_{t,u}/\gamma$ to the conditional prediction gives the guided velocity

$$v_\theta(x_t, \xi_t, \emptyset) + \gamma\Big(v_{\mathrm{cond},t}^F + \frac{1}{\gamma}\sum_u \alpha_{t,u} \widehat{r}_{t,u} - v_\theta(x_t, \xi_t, \emptyset)\Big) = v_{\gamma,t}^F + \sum_u \alpha_{t,u} \widehat{r}_{t,u}, \tag{17}$$

where $v^F_{\gamma,t}$ is the native guided velocity. Before native renormalization, the guided velocity therefore receives exactly the correction of Step 4, with each $\widehat{r}_{t,u}$ normalized to $\mathrm{RMS}(\gamma v^F_{\mathrm{cond},t})$, so the bound of Equation (16) holds at this injection point.

C Implementation Details

C.1 Relative Maps and Spatial Reliability

In Equation (1), each attention summary is nonnegative and normalized over visual positions before taking a positive log-ratio in understanding, whereas in generation the normalized attention maps of the instruction branch serve directly as head maps. Each head's positive evidence is divided by its spatial sum with a lower bound of ϵ. A zero-evidence head therefore contributes a zero vector to the head average. An empty aggregate is represented by a uniform map, which does not establish condition-specific spatial evidence.

For a candidate layer ℓ, let g_ℓ be the aggregated map, ψ_ℓ the mean pairwise cosine between normalized head maps, and b_ℓ the cosine between g_ℓ and the coordinate-wise median map of the other candidate layers. The spatial null independently permutes visual positions within every head and recomputes the mean pairwise cosine. Let $\psi^{\mathrm{null}}_{\ell,95}$ be the 95th percentile of these null scores. Layer reliability is

$$q_\ell = \left[(1 - \mathcal{H}(g_\ell))[\psi_\ell]_+ [b_\ell]_+ \frac{[\psi_\ell - \psi^{\mathrm{null}}_{\ell,95}]_+}{\max(|\psi_\ell|, \epsilon)} \right]_{[0,1]}, \tag{18}$$

where $\mathcal{H}(g) = -\sum_i g_i \log(g_i)/\log N_{\mathrm{v}}$ is normalized spatial entropy over N_{v} visual positions. The highest-reliability layer is selected, with specificity and head agreement breaking ties. For P permutations with null scores $\psi^{\mathrm{null}}_{\ell,n}$, the empirical tail statistic is $\hat{p}_\ell = (1 + \#\{n : \psi^{\mathrm{null}}_{\ell,n} \geq \psi_\ell\})/(P+1)$.

Understanding selects each unit's layer with its anchor, and the per-unit route g_u is the unit's relative map at that layer. For $W = \sum_u I_u > 0$, the aggregate map and centered semantic residual are

$$g_{\mathrm{agg}} = \mathcal{N}\left(\sum_u I_u g_u \right), \qquad \bar{r} = \frac{1}{W} \sum_u I_u r_u. \tag{19}$$

The final review map is $\mathcal{N}([g_{\mathrm{agg}} - \mathrm{mean}(g_{\mathrm{agg}}) - \mathrm{std}(g_{\mathrm{agg}})]_+)$ when the excess has positive mass, and g_{agg} otherwise. The spatial standard deviation is computed over all visual positions. If $W = 0$, the original answer logits are retained. For understanding, the permutation null enters only through the layer reliability in Equation (18) and is not an acceptance condition for the review.

Generation applies the selected-layer permutation threshold directly as an operational eligibility rule for each unit. It additionally compares each unit's selected map with that unit's most recently accepted map, not necessarily its map at the immediately preceding step. Residual stability, in contrast, compares consecutive residuals, which all units share, even if the preceding intervention was rejected. The first residual has stability factor one, and a unit's first eligible map has no temporal-map requirement. The selected-layer tail statistic is an algorithmic screening criterion, not a multiple-comparison-corrected significance claim.

C.2 Understanding Decision and Execution

The answer is selected with the residual-supported decision in Equation (4). A centered review change with squared norm at most ϵ is assigned zero projection. The decision uses only model outputs, and benchmark labels enter correctness scoring afterward.

BAGEL computes the answer query against the image and text keys in its key–value cache. For Qwen-Image, the understanding evaluation instead uses a full-sequence forward pass with caching disabled and modifies the final query row. Both target the attention computation producing the first answer token, but their cache execution paths differ. For open-ended questions, only the first answer token is steered, and the remaining tokens are decoded natively without further intervention.

C.3 Generation Strength and Injection

We compute the strength factors once per step from three properties of the shared residual r_t (Section 3.3), namely its root-mean-square (RMS) magnitude relative to the native velocity, its alignment with the native velocity, and its consistency with the preceding residual,

$$\rho_t = \min\left(1, \frac{\mathrm{RMS}(r_t)}{\max(\mathrm{RMS}(v_t), \epsilon)}\right), \qquad \pi_t = [\cos(r_t, v_t)]_+, \qquad s_t = [\cos(r_t, r_{t-1})]_+. \tag{20}$$

RMS and cosine similarity are computed over all visual positions and feature dimensions, with $s_t = 1$ at the first step. The phase weight $\phi_u(\tau_t)$ in Appendix C.4 regulates when unit u is emphasized, and the strength of unit u is

$$\alpha_{t,u} = a_{\max}\, \phi_u(\tau_t)\, \rho_t\, \pi_t\, s_t. \tag{21}$$

The envelope $a_{\max}$ is fixed and bounds each unit's intervention, while the remaining factors adapt the strength to the unit and the current state. A unit contributes at step t only if its selected map passes the spatial eligibility tests of Appendix C.1 and the residual passes the minimum projection and residual-cosine tests. Otherwise, its strength is zero, and a step at which no unit is eligible keeps the native prediction.

Let $g_{t,u}$ be unit u's map (Equation (1)) at the layer selected for step t, and let $G_{t,u} = g_{t,u} / \mathrm{RMS}(g_{t,u})$ be its unit-RMS spatial multiplier, applied uniformly across feature channels. Each unit's share of the residual is localized with its multiplier and normalized, and the localized residuals of all units are added,

$$\widehat{r}_{t,u} = \mathrm{RMS}(v_t) \frac{r_t \odot G_{t,u}}{\mathrm{RMS}(r_t \odot G_{t,u})}, \qquad v'_t = v_t + \sum_{u \in \mathcal{U}} \alpha_{t,u} \widehat{r}_{t,u}. \tag{22}$$

A map therefore determines only where its unit's share of the residual acts, and $\alpha_{t,u}$ alone determines how much (Appendix B.2). The native solver then advances the latent state with v'_t, keeping its time grid and update convention.

C.4 Semantic Phase Profiles

The generation phase factor is

$$\phi_u(\tau) = \left[f + (1-f)\exp\left(-\frac{1}{2}\left(\frac{\tau - \mu_u}{\max(\sigma_u T, 0.05)}\right)^2\right)\right] \frac{1+\zeta_u}{2}\, d_u. \tag{23}$$

The center μ_u and width σ_u specify a broad temporal profile, T is the common temperature, f is the floor, and ζ_u is the unit-confidence metadata, defaulting to one when absent. The factor d_u combines scope and dependency metadata. Scope contributes one when explicitly specified and 0.9 otherwise. Dependency weight defaults to one, is capped at 0.75 for an explicitly invalid dependency, and is restricted to $[0.25, 1]$. These factors attenuate the prior but cannot override a failed residual or spatial eligibility test.

Table 6: Default external phase profiles. Explicit valid phase metadata overrides the corresponding type default. These values are fixed priors rather than quantities learned from benchmark outcomes.

Semantic type	Center μ_u	Width σ_u
Quantity	0.18	0.24
Relation	0.24	0.28
Action	0.40	0.32
Object	0.38	0.46
Attribute	0.66	0.30

The implementation restricts centers to $[0, 1]$, widths to $[0.05, 1]$, confidence to $[0, 1]$, temperature to $[0.25, 4]$, and the phase floor to $[0, 0.25]$.

C.5 Numerical and Backbone Conventions

Numerical divisions and logarithms use $\epsilon = 10^{-8}$. Residual RMS is computed in float32 with a lower bound on mean-square energy. The implementation first aligns the residual RMS to the native velocity, applies a unit-RMS spatial multiplier, and then restores the native RMS after spatial multiplication. After cancellation of intermediate positive scales, this operation equals the update in Equation (22), although the implementation retains epsilon-protected divisions. A spatial permutation therefore preserves each unit's injected RMS at a fixed state and dose, up to numerical precision, but not necessarily its direction.

Nonfinite values, incompatible map dimensions, or vanishing residuals suppress correction. These checks are distinct from the residual envelope and acceptance thresholds, which constrain the intervention itself.

Both generation backbones use the same requested candidate-layer indices, discarding indices outside the loaded model. For BAGEL, TaylorSeer caching is disabled, and the original and reference evaluations use independent copies of the input and cache state. Its update is $x_{t+1} = x_t - v'_t \Delta t_t$ with positive Δt_t from the descending time grid. Qwen-Image retains its native scheduler convention. Neither backbone's insertion changes the sign of the residual independently of the solver's native velocity convention.

C.6 Experimental Configuration

Understanding intervention. Understanding combines spatial review with the residual-supported decision in Equation (4). The strict reference construction repeats the neutral token `object` over the original span, retaining token count, masks, image inputs, visual positions, and answer position. The high-level review layer is fixed at zero-based $L = 27$. Map selection uses 32 spatial permutations. The review is activated by residual influence rather than by an additional spatial gate.

Native generation settings. Qwen-Image uses CFG 4.0. BAGEL text-to-image generation uses text/image CFG 4.0/1.0, timestep shift 3.0, guidance interval $[0, 1]$, and global CFG renormalization with minimum 0. These are backbone inference settings, shared across the original and reference evaluations, not additional CMCS coefficients. With 50 nominal steps, Qwen-Image performs 50 updates and BAGEL performs 49 updates over a 50-point time grid. The implemented phase coordinate is $t/49$ in both cases, so BAGEL's final evaluated phase is $48/49$. BAGEL editing uses text/image CFG 4.0/1.5, serial text–image conditioning, 50 time points, seed 42, and 512–1024 pixel bounds. Generation uses bfloat16 where required by the inference configuration.

Backbone-specific insertion points. BAGEL forms the residual from the text-conditioned and text-free predictions of its native classifier-free guidance and adds the summed correction before the native latent update. Qwen-Image inserts the correction through a conditional proxy. Writing its native CFG scale as γ and its unconditional prediction as $v_t^{\emptyset}$, the proxy uses $r_t = \gamma(v_{\mathrm{cond},t}^F - v_t^{\emptyset})$ and $v_t = \gamma v_{\mathrm{cond},t}^F$ in Equations (20)–(22), then adds $\sum_u \alpha_{t,u} \widehat{r}_{t,u}/\gamma$ to the original conditional prediction. Native CFG combination and renormalization follow. Thus, the RMS budget applies at the proxy injection point, not necessarily to the final solver velocity.

Spatial controls. The controls in Tables 4 and 5 change only the spatial map. For understanding, they hold the review layer and β fixed: the random control permutes the positions of $g^\star$, and the uniform control applies the review without spatial selectivity. For generation, the uniform control replaces each $G_{t,u}$ by a constant multiplier while retaining the spatial eligibility tests and the adaptive doses, and the random control permutes map positions before residual shaping and RMS renormalization. Both generation controls therefore inject the same per-unit magnitude as CMCS at a common state (Appendix B.2).

Fixed implementation settings. The adaptive factors in Equation (21) are computed online. The implementation fixes the per-unit residual envelope at 0.35, phase temperature at 1.0, phase floor at 0.15, minimum positive projection at 0.05, and minimum residual cosine at 0.10. Spatial eligibility requires permutation $p \leq 0.05$, positive excess over the null's 95th percentile, and cosine at least 0.20 with the most recently accepted map, when one exists. The generation protocol uses

128 permutations and up to 256 uniformly sampled visual queries per audited layer, with linear interpolation over visual-token indices. Detailed phase profiles and numerical conventions appear in Appendix C.4.

These fixed settings are distinct from the data-dependent strengths and from native sampler settings. The residual envelope, acceptance thresholds, and phase priors affect intervention behavior but are neither learned nor sample-specific. We consolidate all settings in one table (Table 7).

Table 7: Consolidated CMCS settings. All values are fixed before evaluation and shared across benchmarks, and none is tuned on evaluation labels or judge outputs. Layer indices are zero-based.

Stage	Setting	Value
Understanding		
Reference	Neutral anchor replacing each unit span at equal length	`object`
Review	Review layer L (final layer of both 28-layer language models)	27
Routing	Sharpening threshold on the aggregate map	mean + 1 std
Nulls	Permutations for map selection	32
Generation		
Dose	Per-unit residual envelope $a_{\max}$	0.35
Phase	Temperature T; floor f; type profiles	1.0; 0.15
Acceptance	Minimum positive projection; minimum residual cosine	0.05; 0.10
Spatial	Bound $p_{\max}$ on the tail statistic $\hat{p}$; minimum cosine $\kappa_{\min}$ with the last accepted map	0.05; 0.20
Nulls	Permutations; sampled visual queries per audited layer	128; ≤ 256
Sampling	Nominal steps N; phase coordinate τ_t	50; $t/49$
Shared		
Numerics	Stabilizer ϵ	10^{-8}

Semantic units and reference cost. Understanding identifies question-relevant noun phrases with the model's language backbone and evaluates the full condition and one neutral anchor per unit before visual review. Generation steers every semantic unit of the prompt, each with its own attention map and type-specific phase profile (Table 6). Its reference is the instruction-free branch of native classifier-free guidance, so it adds no model evaluations.

D Models and Evaluation Protocols

D.1 Backbone Architectures

We list the components and parameter scales of the two backbones (Table 8) and the language-backbone sizes of understanding baselines with reported configurations (Table 9). CMCS acts only on attention and velocity computations, modifying no weights and adding no parameters. For understanding, the review is applied at layer 27, the final layer of the 28-layer language models of BAGEL and Qwen2.5-VL, which in BAGEL coincides with the late rise of visual attention in Figure 1a. For generation, maps are read from the candidate layers that exist in each generator: layers 0, 10, and 20 of BAGEL's 28-layer transformer, and all seven candidate layers of Qwen-Image's 60-layer multimodal diffusion transformer (MMDiT). The localized residuals enter BAGEL's guided velocity before the latent update and Qwen-Image's conditional prediction before native CFG.

D.2 Evaluation Protocols

Soft-TIFA scoring on GenEval. The controlled comparisons use the 800 compositional prompts of GenEval, each with 3–10 constraints involving objects, attributes, counts, and relations. For generated image I_j, atom question q_{ji}, and acceptable answer set A_{ji}, define $s_{ji} = \sum_{a \in A_{ji}} p_{\mathrm{VQA}}(a \mid I_j, q_{ji})$. Over the $J = 800$ prompts, where prompt j contains m_j atoms, equal prompt weights yield

$$\mathrm{AM} = \frac{100}{J} \sum_{j=1}^{J} \frac{1}{m_j} \sum_{i=1}^{m_j} s_{ji}, \qquad \mathrm{GM} = \frac{100}{J} \sum_{j=1}^{J} \left(\prod_{i=1}^{m_j} s_{ji} \right)^{1/m_j}.$$

Table 8: Architectures and parameter scales of the evaluated backbones, following the released checkpoints. MMDiT denotes a multimodal diffusion transformer and VAE a variational autoencoder. CMCS adds no parameters to either model.

Backbone	Component	Configuration	Parameters
BAGEL-7B-MoT	Visual encoder	SigLIP2-so400m/14; 27 layers, width 1152, native resolution up to 980 pixels	≈0.4B
	Transformer	Mixture-of-Transformers initialized from Qwen2.5-7B-Instruct; 28 layers, width 3584, 28 query and 4 key–value heads	7B active, 14B total
	Autoencoder	FLUX VAE with 16 latent channels	–
Qwen-Image	Condition model	Qwen2.5-VL-7B; vision transformer with 32 layers and width 1280; language model with 28 layers, width 3584, 28 query and 4 key–value heads	≈0.7B vision, ≈7.6B language
	Generator	MMDiT; 60 layers, 24 heads of dimension 128, text-condition width 3584	20B
	Autoencoder	Qwen-Image VAE with 16 latent channels	–

Table 9: Language backbones of understanding baselines, for comparison of scale with Table 8. Sizes are nominal language-backbone sizes.

Model	Language backbone	Size
InternVL2.5-8B (Chen et al., 2024)	InternLM2.5-7B-Chat	7B
Janus-Pro-7B (Chen et al., 2025)	DeepSeek-LLM-7B-base	7B
MetaQuery-XL (Pan et al., 2025)	Qwen2.5-VL-7B	7B
TokenFlow-XL (Qu et al., 2025)	Vicuna-13B	13B

The VQA judge p_{VQA} is a locally deployed Qwen3.8-27B (Qwen Team, 2026), an open-source multimodal model, which is used only for Soft-TIFA. Soft-TIFA serves only for evaluation, and CMCS keeps model parameters frozen.

GenEval and DPG-Bench. Both benchmarks are evaluated on their full prompt sets with their official scoring protocols (Ghosh et al., 2023; Hu et al., 2024): GenEval verifies object-level constraints with its detection-based pipeline, and DPG-Bench checks the propositions of dense prompts by question answering, propagating failures to dependent propositions. Published baseline scores in Table 2 are taken from the respective reports and provide context rather than a matched-protocol ranking.

Editing scores. GEdit-Bench scores are aggregated per example and reported separately for English and Chinese, so the overall score O is not reconstructed from the aggregate SC and PQ.

E Qualitative Examples

We show text-to-image and editing examples from BAGEL with and without CMCS (Figure 4). The text-to-image prompts combine counts, relations, and attributes, the constraint types on which text-to-image models commonly fail (Ghosh et al., 2023). With CMCS, the images contain the requested numbers of bagels and sheep and include the two chairs that native sampling omits, whereas the four requested monkeys are missing with and without CMCS, an instance of the poorly satisfied atoms that remain in the matched GenEval comparison (Section 4.3). In editing, CMCS executes an addition that native BAGEL ignores and changes the person's gaze while preserving the phone that native BAGEL removes, in line with its higher semantic consistency on GEdit-Bench (Table 3).

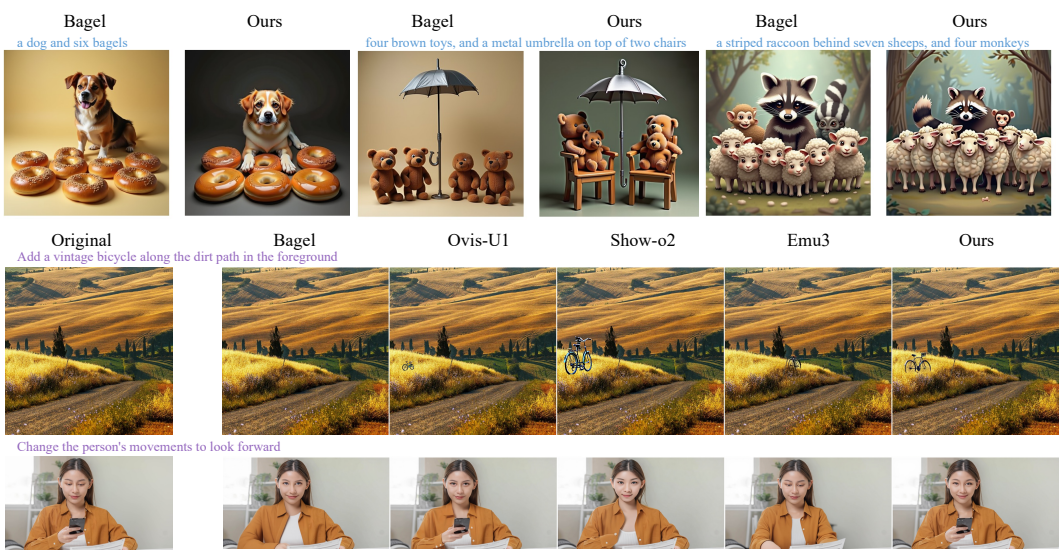


Figure 4: Qualitative examples with BAGEL, where Ours denotes BAGEL with CMCS. Top: text-to-image generation. With CMCS, the image contains six bagels instead of eight, the two chairs missing from the native image appear beneath the umbrella with the four toys, and seven sheep instead of six stand in front of a raccoon with a striped tail. Neither image contains four monkeys. Bottom: instruction-guided editing compared with native BAGEL, Ovis-U1, Show-o2, and Emu3 (Wang et al., 2025; Xie et al., 2026; Wang et al., 2024). With CMCS, BAGEL adds a vintage bicycle beside the dirt path, which native BAGEL omits, and turns the person's gaze forward while keeping the phone in hand, which native BAGEL removes.