CARVE: CASCADED ADAPTIVE ROUTING OF VISUAL-CONDITION EMBEDDINGS FOR MULTI-REFERENCE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-reference controllable generation represents each reference with many visual-condition tokens. As the number of references grows, these tokens create substantial attention cost in existing transformer architectures. Yet many tokens contribute little to the current generation while still being processed at every layer and denoising step. Existing compression methods use fixed, hand-crafted rules to select or merge tokens. Such rules cannot always accurately measure each reference token's contribution to the current generation and may mistakenly remove important reference-specific details. We propose **CARVE**, the first learnable and input-adaptive framework for visual-condition-token compression. Rather than predefining token importance or pruning locations, CARVE learns to preserve the tokens that contribute to the current generation at every denoising step and to remove tokens after their reference information has become redundant. The resulting compression adapts to each input and denoising computation while preserving the information required for faithful multi-reference generation. Across diffusion-transformer and unified multimodal generation backbones, CARVE achieves the strongest generation quality among compared compression methods while substantially reducing token processing and inference latency.

1 INTRODUCTION

Modern image generators increasingly rely on multiple reference images to preserve subjects, compose objects, and transfer style or scene attributes. A common in-context formulation encodes each reference as visual tokens and concatenates them with the noised target tokens (Tan et al., 2025a; Black Forest Labs et al., 2025; Wu et al., 2025a; Deng et al., 2025). Although effective, this design becomes expensive in transformer architectures, where the growing number of visual-condition tokens leads to increasing attention cost at every block and denoising step. At the same time, the computation does not use reference content equally. Background patches, duplicated regions, and information already represented in the target sequence can make only a limited contribution at a given computation (Figure 1), yet remain costly to process. Visual-condition-token compression should remove these redundant tokens to reduce computation while retaining the reference details required for faithful multi-reference generation.

The key to visual-condition-token compression is deciding which tokens to remove. Existing approaches reduce computation in two ways. One line of work directly shortens the visual-condition token sequence: ToMe, ToMeSD, and ToMA (Bolya et al., 2023; Bolya & Hoffman, 2023; Lu et al., 2025) merge tokens by feature similarity; ToPi (Lin et al., 2026) ranks tokens with an attention–value heuristic; and OminiControl2 (Tan et al., 2025b) applies a uniform spatial downsampling ratio to every reference. The other prunes attention connections rather than visual-condition tokens: PKA (Zhou et al., 2026) uses position alignment and keyword scoping to sparsify attention. Both use fixed, hand-crafted heuristic criteria. These criteria are simple, efficient, and can be effective in their intended settings, but they are fixed before inference rather than derived from the generator's present representations. Therefore, these criteria cannot always accurately measure the actual contribution of each visual-condition token to the generator's current denoising computation. They may remove visual-condition tokens carrying important identity or texture details while retaining redun-

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.

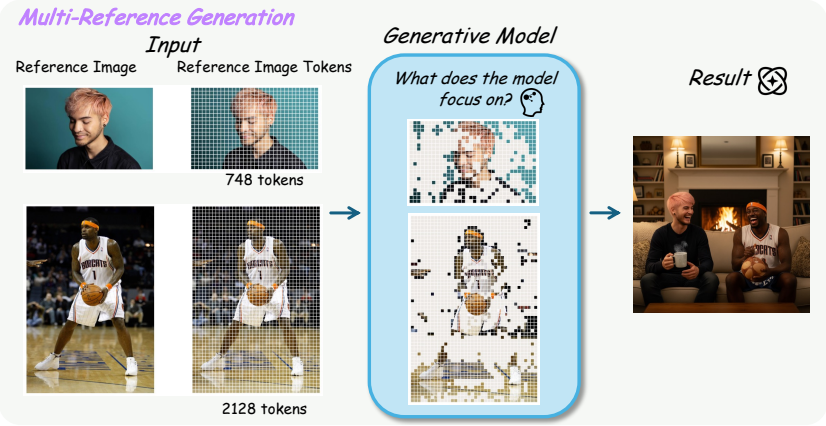


Figure 1: Motivation. Multi-reference generation encodes each reference as many visual-condition tokens, although only a subset is emphasized for a given output.

dant tokens. With multiple references, merging or uniform reduction can further blend or remove the distinct identity, color, and appearance cues from each image.

To address this problem, we introduce **CARVE** (**C**ascaded **A**daptive **R**outing of **V**isual-condition **E**mbeddings). Its central component is a lightweight, learnable *visual-condition Router*. Rather than rank tokens with a fixed proxy, the Router reads the current visual-condition states and a summary of the noised target, then scores the visual-condition tokens for the current input and denoising state. At each Router layer, it keeps a budgeted subset and physically deletes the rest; the surviving tokens form the input to the next Router, progressively shortening the sequence throughout the remaining transformer blocks. Across generative models, visual-condition redundancy can emerge at different layers. We therefore introduce Dependency-Drop Selection (DDS), a model-aware placement criterion that measures condition-token redundancy in the frozen model and identifies suitable layer transitions for Router insertion. We train only the Routers through full-context teacher–student distillation, using the frozen full-token model as the teacher on existing training samples. Random-budget training enables the same Router policy to operate over a continuous range of compression ratios without token-level labels or additional data.

On DreamOmni2Bench, CARVE even surpasses the full-token model, doubling inference efficiency while improving performance by approximately 10%. This result further indicates substantial redundancy in multi-reference visual context. On OmniContext with Qwen-Image, CARVE compresses roughly 50% of the condition tokens with no more than a 3% performance loss. Finally, ablations confirm that the gains arise from the proposed dynamic routing and DDS-based layer selection.

Our contributions are:

- We introduce CARVE, a learnable visual-condition-token compression framework built around an input-adaptive Router that physically deletes selected tokens in a cascade.
- We train only the Routers with full-context distillation and random budgets, allowing one frozen-backbone model to support a continuous range of compression ratios without additional labels or data.
- We introduce DDS as a model-aware measure of condition-token redundancy for selecting Router layers, and show that CARVE improves generation quality and efficiency across diffusion-transformer and unified multimodal backbones.

2 Related Work

2.1 Multi-Reference Visual Conditioning

Early adapter-based methods inject image features through cross-attention, such as IP-Adapter (Ye et al., 2023) and BLIP-Diffusion (Li et al., 2023). More recent generative models place reference

visual tokens and the noised target in one sequence for joint self-attention, including FLUX.1 Kontext (Black Forest Labs et al., 2025), Qwen-Image (Wu et al., 2025a), Step1X-Edit (Liu et al., 2025), OmniGen2 (Wu et al., 2026), In-Context LoRA (Huang et al., 2024), OminiControl (Tan et al., 2025a), and Less-to-More (Wu et al., 2025b). Unified multimodal models such as BAGEL (Deng et al., 2025) similarly process reference ViT/VAE tokens together with target tokens. This in-context design provides strong multi-reference control, but every visual-condition token is processed throughout generation. Its cost therefore increases with the number of references, motivating visual-condition-token compression in multi-reference settings.

2.2 Visual-Condition-Token Compression

Existing visual-condition-token compression mainly uses fixed, hand-crafted heuristics. ToMe and ToMeSD (Bolya et al., 2023; Bolya & Hoffman, 2023) merge tokens by feature similarity; ToMA (Lu et al., 2025) uses merge–unmerge; ToPi (Lin et al., 2026) ranks context tokens with an attention–value heuristic; and OminiControl2 (Tan et al., 2025b) uniformly downsamples each reference before encoding. PKA (Zhou et al., 2026) instead sparsifies multi-condition attention through position alignment and keyword scoping, without shortening the token sequence. Because these predesigned rules do not observe the generator's current condition and target states, they cannot always accurately measure a token's contribution to the current denoising computation. Adaptive condition-token selection with actual sequence reduction therefore remains unresolved.

2.3 Learned and Cascaded Token Selection

Learned token selection in discriminative vision, such as DynamicViT (Rao et al., 2021) and EViT (Liang et al., 2022), predicts token scores from intermediate states and progressively prunes image tokens across layers. Related token-reduction studies also consider vision-language and multimodal models (Chen et al., 2024; Cai et al., 2025; Shang, 2025; Ye et al., 2025; Yang et al., 2025; Alvar et al., 2025; Gao et al., 2026; Li et al., 2026a;b; Xu et al., 2026; Sun et al., 2026; Xiong et al., 2026; Zhu et al., 2026). Dynamic sparse-attention and cache-selection methods (Gao et al., 2025; Yuan et al., 2025; Jiang et al., 2024; Zhang et al., 2023) show that attention computation can be reduced by data-dependent sparsity. These methods target visual recognition, language models, or attention sparsity rather than multi-reference visual conditioning in generative models. They do not establish how a frozen generator should select visual-condition tokens, where cascaded deletion should occur, or how the selection can be trained without token-level labels in generative settings. CARVE addresses these questions through adaptive routing and generative distillation.

3 Method

3.1 Overview and Notation

Figure 2 shows CARVE: lightweight, input-adaptive *visual-condition token Routers* prune a frozen generator, while DDS selects layers where condition information has become redundant. A forward pass processes

$$\mathbf{X} = [\,\mathbf{X}^{\text{tgt}} \parallel \mathbf{X}^{\text{cond}}\,], \tag{1}$$

where $\mathbf{X}^{\text{tgt}}$ are noised target tokens and $\mathbf{X}^{\text{cond}}$ are L_c visual condition tokens. For Qwen-Image, these are concatenated reference VAE latents after the target; for BAGEL, they are each reference's ViT and clean-VAE tokens. We insert N Routers at DDS-selected layers $\{\ell_1 < \cdots < \ell_N\}$ (Sec. 3.5); target tokens are never deleted.

CARVE makes selection part of the frozen generator's computation: a reference token may matter for one input or denoising state but not another. Each Router reads the current hidden states, selects the surviving condition tokens, and passes only them to later blocks; the target sequence remains intact throughout the diffusion sampling process.

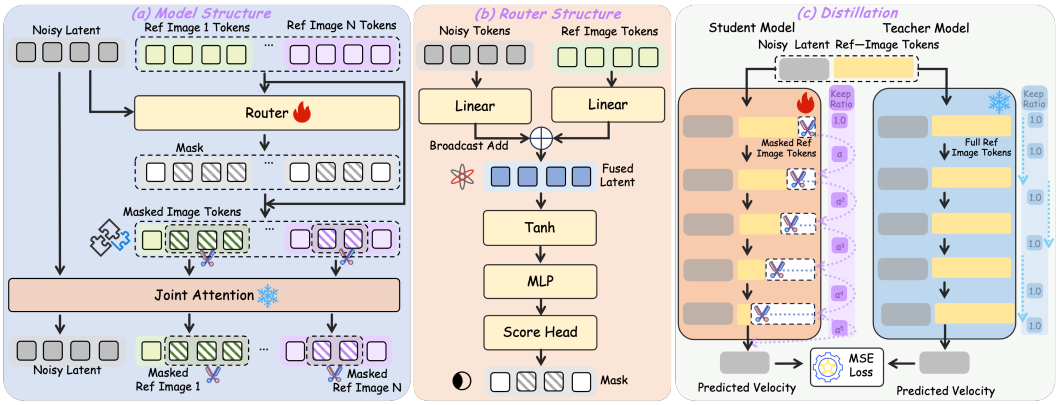


Figure 2: Overview of CARVE. (a) Cascaded Router masking in a frozen backbone. (b) The lightweight Router scores visual-condition tokens using the current target and condition states. (c) Full-context teacher–student distillation trains the Routers under an exponential keep-ratio schedule.

3.2 Adaptive Learnable Visual-Condition Token Router

At each Router layer, a lightweight scorer (Figure 2b) assigns a keep-score to every *alive* condition token by fusing a global summary of the target tokens with each condition token:

$$\mathbf{q} = \frac{1}{L_t} \sum_{i=1}^{L_t} \mathbf{W}_q \, \mathbf{h}_i^{\text{tgt}}, \tag{2}$$

$$s_j = \sigma\big(\mathbf{w}^\top \phi(\mathbf{W}_m \tanh(\mathbf{q} + \mathbf{W}_c \mathbf{h}_j^{\text{cond}}))\big), \tag{3}$$

where $\mathbf{h}_i^{\text{tgt}}$ and $\mathbf{h}_j^{\text{cond}}$ are the target and condition hidden states, $\mathbf{W}_q$, $\mathbf{W}_c$, and $\mathbf{W}_m$ are learned projections, ϕ is a pointwise nonlinearity, and σ produces one keep-score $s_j \in (0, 1)$ per condition token.

The target summary represents the current generation state, while each condition state retains token content and position. Their fusion scores tokens against the target. Routers are the only learned additions; at selected layers, they choose tokens according to the current input and denoising state rather than applying a static mask.

3.3 Cascaded Physical Deletion and Position-Encoding Preservation

At inference, Router i (at layer ℓ_i) scores the currently alive condition tokens, keeps the top-k_i, and *physically removes* the rest from the token sequence, RoPE tables, and attention mask. The keep count is anchored to the original condition count L_c^{orig} under an exponential schedule:

$$k_i = \max\big(1,\ \min(L_c^{\text{alive}},\ \text{round}(L_c^{\text{orig}}\, r_i))\big), \tag{4}$$

$$r_i = \alpha^{\,i}, \tag{5}$$

with per-layer factor $\alpha \in (0, 1]$, giving cumulative retention α^N. Deletion is irreversible: later layers process the shortened condition sequence, while target, text, and understanding tokens remain unchanged.

The cascade lets Routers remove tokens at different stages. A token can be kept early and deleted later once its reference information has been incorporated into the target states. Removing the token physically, rather than only setting its attention weight to zero, makes all later layers process a shorter sequence, reducing attention computation.

Position-encoding preservation. A kept condition token retains its original RoPE position. Training implements deletion with an attention-mask bias over the full rotary embedding, preserving the same visible-token set and geometry as inference-time physical deletion.

3.4 TRAINING: SELF-DISTILLATION WITH RANDOM BUDGETS

CARVE freezes the diffusion backbone, text encoder, and VAE, and trains only the Routers. Each step uses a full-token teacher and Router-compressed student, optimized by flow matching and velocity distillation:

$$\mathcal{L} = \lambda_{\text{fm}}\, w(t) \left\| \mathbf{v}^{\text{stu}} - \mathbf{v}^{\text{tgt}} \right\|^2 + \lambda_{\text{dist}} \left\| \mathbf{v}^{\text{stu}} - \text{sg}[\mathbf{v}^{\text{tea}}] \right\|^2, \tag{6}$$

where $\mathbf{v}^{\text{tgt}} = \boldsymbol{\epsilon} - \mathbf{x}_0$, $w(t)$ is a timestep weight, and $\text{sg}[\cdot]$ denotes stop-gradient. Hard top-k is trained with a straight-through soft mask, while accumulated masks enforce the cascade across Router layers. At each step, we sample $\alpha \sim \mathcal{U}[0.3, 1.0]$ and set $r_i = \alpha^i$, allowing one trained model to support multiple inference budgets. Each base model uses a checkpoint fine-tuned on MICo-150K. We train only the Routers on a 10K-sample subset of the same dataset, with the backbone frozen, thereby avoiding confounding the comparison with a different training dataset.

The full-token teacher supervises information lost after deletion without token labels or extra data. Flow matching preserves the generation objective, and distillation aligns compressed and full-token predictions. Random-budget training exposes Routers to many sequence lengths, so changing α adjusts the inference budget without retraining.

3.5 CONDITION-TOKEN REDUNDANCY MEASUREMENT VIA DEPENDENCY-DROP SELECTION (DDS)

DDS places Routers where condition tokens become redundant. On a stratified 40-case MICo-150K subset, it measures target-to-reference dependency across frozen-backbone layers and steps. The primary *contribution ratio* $C(\ell)$ is the L2-norm share of attention output attributable to reference tokens:

$$C(\ell) = \sqrt{\frac{\left\| \left(\text{softmax}(\mathbf{Q}^{\text{tgt}}\mathbf{K}^{\top}/\sqrt{d})|_{\text{ref}}\right) \mathbf{V}^{\text{ref}}\, \mathbf{W}_O \right\|^2}{\left\| \text{softmax}(\mathbf{Q}^{\text{tgt}}\mathbf{K}^{\top}/\sqrt{d})\, \mathbf{V}\, \mathbf{W}_O \right\|^2}}, \tag{7}$$

The auxiliary *attention ratio* $A(\ell)$ is the attention mass that target queries place on reference keys,

$$A(\ell) = \frac{1}{H\, N_q} \sum_{h=1}^{H} \sum_{i=1}^{N_q} \sum_{j \in \text{ref}} \text{softmax}\big(\mathbf{q}_{h,i}^{\text{tgt}} \mathbf{K}_h^{\top}/\sqrt{d}\big)_j, \tag{8}$$

averaged over heads and queries. We average both measures over cases and steps, normalize them, and form $S(\ell) = w_A \hat{A}(\ell) + w_C \hat{C}(\ell)$.

The terms are complementary: $A(\ell)$ measures attention mass on reference keys, whereas $C(\ell)$ measures their contribution to the resulting output. Their combination measures layerwise condition dependency before Router training.

DDS converts the dependency reduction into a layerwise redundancy signal:

$$R(\ell) = S(\ell) - S(\ell{+}1), \tag{9}$$

where a larger $R(\ell)$ means that more reference information has been absorbed into the target representation and the condition tokens are more redundant. We select the K transitions with the largest $R(\ell)$ (excluding the first and last layers) and place Routers there. This adapts Router placement to dependency patterns within each backbone. Figure 4(b) shows the selected layers.

A large positive signal indicates that reference information has entered the target sequence. Placing a Router there avoids repeated computation without deleting details too early. DDS is computed once and only selects Router locations; learned Routers make per-input decisions.

3.6 INSTANTIATION

CARVE applies unchanged to Qwen-Image, a 60-block DiT (D=3072) with reference VAE latents, and BAGEL, a 28-layer MoT LLM (D=3584) with reference ViT and clean-VAE tokens. BAGEL additionally deletes the corresponding rows and columns of its 2D attention mask. Both use the same Router, physical deletion, DDS, and random-budget distillation while freezing the backbone; only condition-segment and position/mask metadata differ.

Table 1: DreamOmni2Bench summary (PF / SC / Overall). Keep = fraction of reference-condition tokens retained. All methods use the Qwen-Image (DiT) backbone; Full tokens is excluded from ranking.

		Editing			Generation			All		
Method	Keep	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.
Full tokens	1.000	3.020	4.678	3.299	5.974	6.105	5.663	4.075	5.188	4.144
CAT-Pruning	0.500	2.737	4.620	3.021	5.693	5.412	5.256	3.793	4.903	3.820
OminiControl2	0.500	2.956	4.454	3.125	<u>6.228</u>	**6.211**	**5.842**	4.126	5.082	4.096
SiTo	0.500	3.273	4.800	3.479	6.132	5.561	5.449	4.295	5.072	4.183
ToMA	0.500	<u>3.537</u>	<u>4.834</u>	<u>3.718</u>	6.044	5.211	5.109	<u>4.433</u>	4.969	4.215
ToMeSD	0.500	3.259	**5.132**	3.524	6.149	<u>5.930</u>	5.624	4.292	**5.417**	<u>4.274</u>
CARVE (ours)	0.471	**3.776**	4.800	**3.823**	**6.377**	5.904	<u>5.819</u>	**4.705**	<u>5.194</u>	**4.537**

Table 2: DreamOmni2Bench results by reference-image number (Qwen-Image). Editing uses two references ($N = 205$); Generation uses one ($N = 39$), two ($N = 25$), three ($N = 29$), or four ($N = 21$) references. Each entry reports PF / SC / Overall. CARVE and OminiControl2 use the same independent per-column selection rule as Table 1.

	Editing-2ref			Gen-1ref			Gen-2ref			Gen-3ref			Gen-4ref		
Method	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.
Full tokens	3.020	4.678	3.299	6.718	4.949	5.061	5.600	6.880	6.011	5.345	6.483	5.758	5.905	6.810	6.235
CAT-Pruning	2.737	4.620	3.021	6.308	4.821	4.939	5.560	6.200	5.741	5.517	5.517	5.398	4.952	5.429	5.074
OminiControl2	3.107	4.502	3.220	6.795	<u>5.103</u>	<u>5.114</u>	<u>5.800</u>	**6.720**	<u>5.950</u>	5.586	<u>6.862</u>	6.085	5.571	6.286	5.857
SiTo	3.273	4.800	3.479	<u>6.923</u>	4.872	5.025	5.360	5.760	5.304	5.793	5.621	5.612	6.048	**6.524**	6.182
ToMA	<u>3.537</u>	4.834	<u>3.718</u>	6.333	3.692	3.854	5.040	5.520	4.910	6.241	6.069	6.055	**6.429**	<u>6.476</u>	**6.371**
ToMeSD	3.259	**5.132**	3.524	6.462	4.590	4.634	5.080	6.080	5.313	**6.621**	**7.276**	**6.817**	<u>6.190</u>	6.381	6.184
CARVE (ours)	**3.707**	<u>4.844</u>	**3.739**	**6.949**	**5.256**	**5.476**	**5.920**	<u>6.640</u>	**6.139**	<u>6.379</u>	6.414	<u>6.266</u>	**6.429**	6.381	<u>6.343</u>

4 EXPERIMENTS

Datasets. The Routers are trained on 10K multi-reference samples drawn from MICo-150K (2–11 references per sample). Evaluation uses two multi-reference benchmarks: OmniContext (400 cases), spanning single-/multi-subject and scene-conditioned composition, and DreamOmni2Bench (319 cases), covering reference-based editing (Add/Replace/Global/Local) and generation with 1–4 references.

Tasks and evaluation metrics. Following the LLM-as-judge protocol of Re-Align (He et al., 2026), both benchmarks are scored by a GPT-4.1 judge on Prompt Following (PF) and Subject Consistency (SC), with Overall reported as their geometric mean. Efficiency is evaluated using the configured input keep ratio, measured condition-token keep ratio, mean condition tokens processed by attention, attention FLOPs, DiT latency, and reference-token removal rate. Throughout, **bold** denotes the best result among compression methods and <u>underline</u> the second best.

Baselines. We instantiate every method on two architecturally distinct backbones: Qwen-Image (DiT) and BAGEL (UMM). We compare CARVE against CAT-Pruning, OminiControl2, SiTo, ToMA, and ToMeSD, and report the uncompressed model (Full tokens) as an upper bound that is excluded from ranking. The main tables report Qwen-Image results; detailed BAGEL results are provided in the appendix.

Implementation details. Each backbone is initialized from a checkpoint fine-tuned on MICo-150K, and only the lightweight Routers are trained on top of it, with the diffusion backbone frozen; each backbone uses $K{=}5$ Routers (hidden width $H{=}2048$) placed at the DDS-selected layers (Figure 4(b)). Training uses AdamW with a constant learning rate of $1{\times}10^{-5}$, bf16 mixed precision, and gradient accumulation of 8 on 8×H200 GPUs, at resolutions up to 1024^2. Loss weights are equal ($\lambda_{\text{fm}}{=}\lambda_{\text{dist}}{=}1$) with mask strength $\lambda_{\text{mask}}{=}10^4$. All inference latencies we report are measured on an

Table 3: OmniContext (400 cases): Single and Multiple categories with their averaged Overall. Keep = mean fraction of condition tokens retained. CARVE uses input keep ratios of 47% for Qwen-Image and 65%† for BAGEL. Full tokens is the upper bound; compression baselines use a 50% input keep ratio. Ov. denotes Overall.

	Qwen-Image (DiT)							*BAGEL (UMM)*						
		Single		Multiple			Avg.		Single		Multiple			Avg.
Method	Keep	Char	Obj	Char	Obj	C+O	Ov.	Keep	Char	Obj	Char	Obj	C+O	Ov.
Full tokens	1.000	8.018	7.788	6.787	7.843	6.490	7.385	1.000	8.163	8.507	7.684	7.840	8.311	8.101
ToMA	0.500	6.177	6.703	4.748	6.066	4.473	5.633	0.500	0.218	0.175	0.057	0.180	0.470	0.220
ToMeSD	0.500	7.318	7.127	5.680	7.289	6.579	6.799	0.500	4.192	**6.313**	3.076	6.364	5.438	5.077
CAT-Pruning	0.500	6.878	7.224	5.812	6.881	5.483	6.456	0.500	0.343	0.913	0.147	0.573	0.393	0.474
SiTo	0.500	7.358	7.449	5.950	7.465	6.406	6.926	0.500	3.363	4.532	2.102	5.986	4.235	4.044
OminiControl2	0.500	7.677	**7.798**	**6.299**	7.502	6.700	7.195	0.500	2.956	3.721	**4.680**	4.563	4.957	4.175
CARVE (ours)	0.471	**7.837**	7.666	6.118	**7.566**	**6.892**	**7.216**	0.440	**5.325**	6.166	3.616	**6.575**	**5.535**	**5.443**

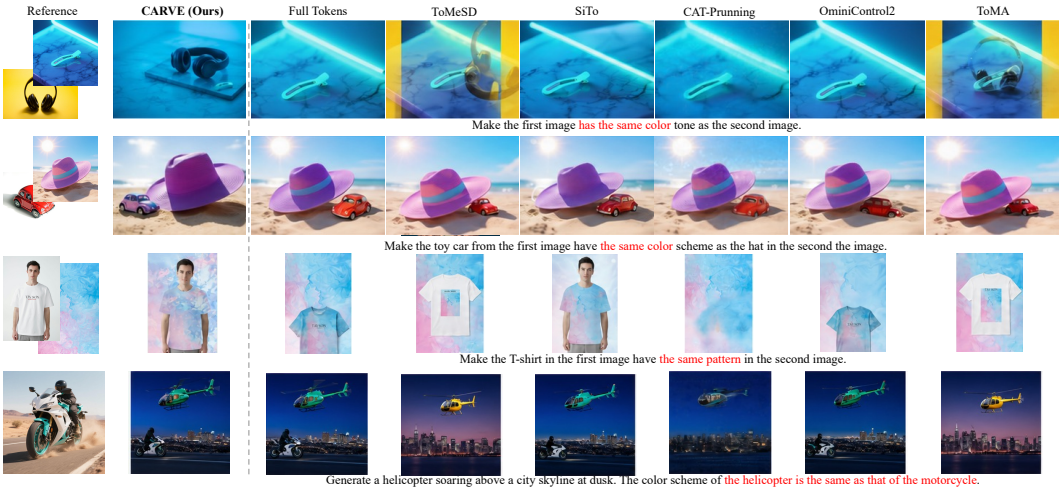


Figure 3: Qualitative comparison on Qwen-Image. Full tokens, ToMA, ToMeSD, SiTo, CAT-Pruning, OminiControl2, and CARVE are compared on four examples. CARVE uses an input keep ratio of 0.47; compression baselines use 0.50.

NVIDIA A100 GPU. We choose input keep ratios of 47% for Qwen-Image and 65% for BAGEL to compare quality gains at acceleration levels comparable to or greater than those of the baselines.

4.1 Qualitative comparison

Tables 3, 1, and 2 summarize the aggregate PF, SC, and Overall trade-offs at comparable compression budgets; Figure 3 shows the visual behaviors behind these differences. CARVE retains recognizable blue headphones and transfers the purple–cyan appearance to the toy car, whereas competing outputs weaken the product identity, leave the car red, or let the hat dominate the composition. It also keeps the person while confining the pastel texture to the T-shirt, and transfers the motorcycle's green–white palette to a coherent dusk helicopter scene; alternatives often lose the person, spread the texture, or produce an unrelated yellow helicopter. These failures are most apparent across these examples when the output must preserve both a source identity and a localized reference attribute. Thus, the table-level PF and SC gaps correspond to a recurring visual distinction: a method may retain a coarse color or object cue but fail to bind the requested reference attribute to the intended subject and region.

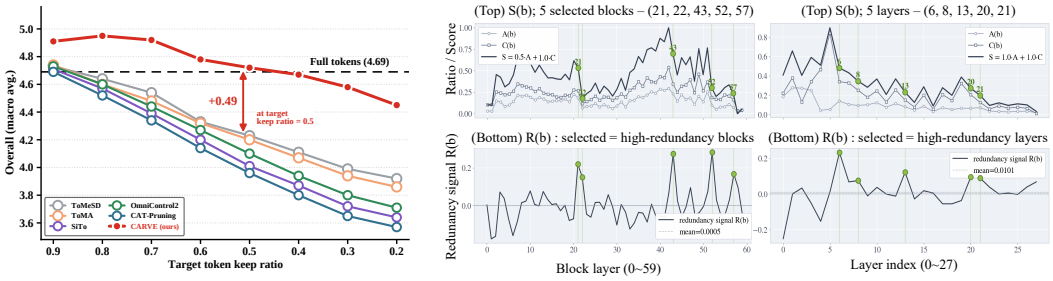
Router-mask visualization. Figure 5(a) reveals two properties of the Router. First, the score maps change across sampling steps, so selection depends on the current denoising state rather than a static

Table 4: Efficiency on DreamOmni2Bench with Qwen-Image.

Method	Input Keep	Mean Keep ↓	Mean $\bar{N}_c$ ↓	Attn. TFLOP ↓	DiT s ↓	Tokens removed/s ↑
Full tokens	1.00	1.000	2,181	833.5	40.645	–
CAT-Pruning	0.50	1.000	2,181	618.6	33.687	0.0
OminiControl2	0.50	0.496	1,082	576.5	33.813	33.4
SiTo	0.50	0.500	1,091	578.5	38.425	28.4
ToMA	0.50	0.500	1,091	578.5	36.851	29.6
ToMeSD	0.50	0.500	1,091	578.5	40.302	27.1
CARVE	**0.47**	**0.456**	**994**	**574.3**	**33.472**	**35.5**

Table 5: Component ablations on DreamOmni2Bench with Qwen-Image (✓/✗: on/off).

flow matching	teacher distillation	DDS	random budget	PF ↑	SC ↑	Overall ↑
Full tokens (upper bound)				4.769	5.423	4.694
✓	✓	✓	✓	5.099	5.533	4.892
✗	✓	✓	✓	4.239	5.063	4.133
✓	✗	✓	✓	4.371	5.148	4.221
✓	✓	✗	✓	4.196	4.519	3.851
✓	✓	✓	✗	4.677	5.039	4.401



(a) Quality across token keep ratios

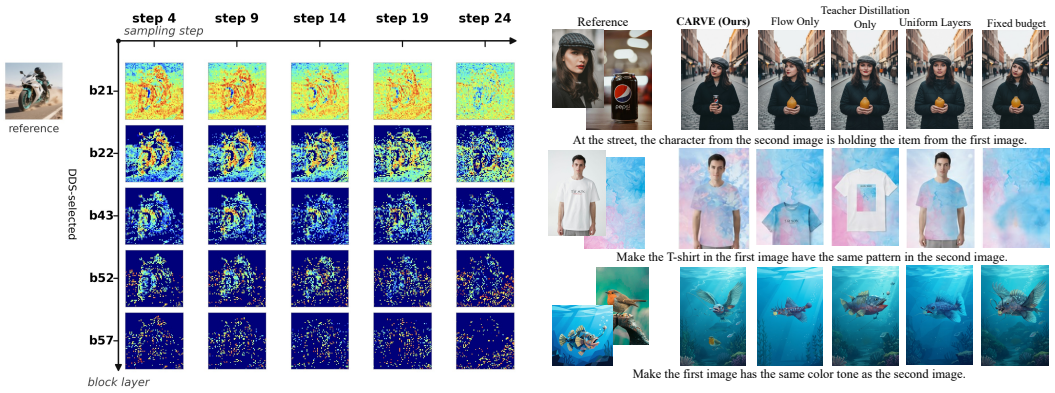
(b) DDS layer selection

Figure 4: Compression robustness and DDS layer selection. (a) Overall (macro-averaged over five tasks) vs. token keep ratio on DreamOmni2Bench with Qwen-Image. (b) DDS condition-token redundancy measurement on Qwen-Image (left) and BAGEL (right): condition-dependency score $S(\ell)$ (top) and redundancy signal $R(\ell)$ (bottom); green markers are the $K{=}5$ selected layers.

spatial mask. Second, the navy deleted regions grow from b21 to b57 as each Router removes part of the remaining condition tokens. Across layers, the retained high-score tokens stay concentrated on the motorcycle, while broad road and sky regions are removed. Thus, the Router does not select arbitrary locations: it preserves the reference regions the model needs for the current generation and discards less relevant content.

Efficiency. Table 4 compares reference-token processing and DiT latency on DreamOmni2Bench. With an input keep ratio of $B{=}0.47$, CARVE processes 994 condition tokens on average (Mean Keep $= 0.456$), yielding the lowest attention cost of 574.3 TFLOPs and the highest reference-token removal rate of 35.5 tokens/s. CARVE also achieves the lowest measured DiT latency among the compared methods, at 33.472 s per case. Relative to Full tokens, it reduces attention computation by 31.1% and latency by 17.6%, corresponding to a $1.21\times$ speedup. CAT-Pruning reduces attention computation while retaining all reference tokens, whereas CARVE reduces both condition-token processing and inference latency.

Robustness across compression budgets. Figure 4(a) sweeps the target keep ratio from 0.9 to 0.2 on DreamOmni2Bench (Qwen-Image), reporting Overall macro-averaged over the five tasks (Full $= 4.694$). A single CARVE model trained with random budgets achieves 4.91–4.95 Overall from 0.9 to 0.7 keep and then degrades gradually as compression strengthens, reaching 4.45 at 0.2 keep. CARVE remains above the Full baseline through 0.5 keep and is the best compression method at every tested budget. Its advantage over the strongest baseline is $+0.48$ at 0.5 keep (4.72 versus 4.24) and $+0.53$ at 0.2 keep (4.45 versus 3.92). Thus, random-budget training does not eliminate the quality cost of extreme compression, but it yields a single selection policy whose degradation is substantially more controlled than that of the fixed compression baselines.



(a) Router keep-score maps

(b) Qualitative component ablation

Figure 5: Router selection and qualitative component ablation on Qwen-Image. (a) Router keep-score maps across the five DDS-selected block layers and sampling steps; navy marks tokens deleted by preceding Routers. (b) Qualitative component ablation on DreamOmni2Bench.

4.2 Ablation Study

Flow matching anchors the requested edit. Removing flow matching produces the largest loss in PF (from 5.099 to 4.239). In the flow-free variant, the reference colors remain visible but are no longer constrained by the requested edit: the T-shirt transfer spreads color beyond the intended garment. Flow matching therefore keeps the compressed prediction aligned with the target image and instruction, especially for where an attribute should be applied.

Teacher distillation preserves reference-specific appearance. Removing teacher distillation reduces SC (from 5.533 to 5.148) and turns the T-shirt pattern into an almost uniform blue appearance. The full-token teacher supplies the reference-conditioned detail that is lost after token deletion. This supervision aligns compressed predictions with those using all tokens. Its role is to preserve fine texture, color variation, and source appearance rather than only the coarse editing direction.

DDS chooses when deletion is safe. Uniformly spacing Router layers gives the lowest Overall score (3.851). Its outputs lose a requested attribute or reference object even though the token budget is unchanged. DDS instead places deletion after reference information has become redundant, so the Router can retain tokens that still determine the shirt pattern, the can–person binding, or the target bottle and remove them only later. Thus, DDS improves compression not by retaining more tokens, but by postponing deletion until tokens are sufficient for the edit and reference binding.

Random-budget training stabilizes selection across budgets. The fixed-budget variant keeps coarse content but produces a flatter, less faithful reference-tone transfer. Training with random budgets exposes the Router to different sequence lengths. Figure 4(a) shows that one model maintains CARVE's advantage over the baselines from a 0.9 to a 0.2 target keep ratio. Its score declines smoothly as compression becomes more aggressive, rather than collapsing at budgets far from the evaluation setting. Random-budget training therefore learns a selection policy that adapts to the token budget instead of relying on one deletion pattern tied to a training ratio.

5 Conclusion

We present CARVE, the first method that reduces the cost of multi reference generation by compressing visual condition tokens. CARVE uses lightweight learnable input-adaptive Routers to select tokens from the current target and condition states, instead of ranking them with a fixed hand crafted criterion. It uses DDS to measure condition token redundancy across layers and places Routers where reference information has been absorbed into the target representation. CARVE trains only the Routers by distilling a full token model under random compression budgets, so one model can operate at different compression ratios. Experiments show that CARVE preserves stronger generation quality than the compared compression methods while reducing the number of processed tokens and inference time. Further extending this learned compression strategy to more generation backbones and diverse tasks is a promising direction for future work.

AI USE STATEMENT

We used AI tools for translation, language polishing, and grammar correction. The authors reviewed and take responsibility for all content.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4599–4603, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *Proceedings of the International Conference on Learning Representation*, 2025.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXI*, pp. 19–35, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73003-0. doi: 10.1007/978-3-031-73004-7_2. URL https://doi.org/10.1007/978-3-031-73004-7_2.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Tianxiao Gao, Shanwei Zhao, Shuo Fang, Shiai Zhu, and Chenguang Ma. QuietPrune: Query-guided early token pruning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3553–3562, 2026.

Yizhao Gao, Zhichen Zeng, Dayou Du, Shijie Cao, Peiyuan Zhou, Jiaxing Qi, Junjie Lai, Hayden Kwok-Hay So, Ting Cao, Fan Yang, and Mao Yang. SeerAttention: Learning intrinsic sparse attention in your LLMs, 2025. URL https://arxiv.org/abs/2410.13276.

Runze He, Yiji Cheng, Tiankai Hang, Zhimin Li, Yu Xu, Zijin Yin, Shiyi Zhang, Wenxun Dai, Penghui Du, Ao Ma, Chunyu Wang, Qinglin Lu, Jizhong Han, and Jiao Dai. Re-Align: Structured reasoning-guided alignment for in-context image generation and editing, 2026. URL https://arxiv.org/abs/2601.05124.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context LoRA for diffusion transformers, 2024. URL https://arxiv.org/abs/2410.23775.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. MInference 1.0: accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Ao Li, Yuxiang Duan, Jinghui Zhang, Congbo Ma, Yutong Xie, Gustavo Carneiro, Mohammad Yaqub, and Hu Wang. TransPrune: Token transition pruning for efficient large vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39529–39538, 2026a.

Dongxu Li, Junnan Li, and Steven CH Hoi. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2023.

Guangyuan Li, Rongzhen Zhao, Jinhong Deng, Yanbo Wang, and Joni Pajarinen. Object-centric vision token pruning for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7040–7049, 2026b.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *ArXiv*, abs/2202.07800, 2022. URL https://api.semanticscholar.org/CorpusID:246867285.

Junqing Lin, Xingyu Zheng, Pei Cheng, Bin Fu, Jingwei Sun, and Guangzhong Sun. Token pruning for in-context generation in diffusion transformers, 2026. URL https://arxiv.org/abs/2602.01609.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1X-Edit: A practical framework for general image editing, 2025. URL https://arxiv.org/abs/2504.17761.

Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. ToMA: Token merge with attention for diffusion models. In *International Conference on Machine Learning*, pp. 40930–40951. PMLR, 2025.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: efficient vision transformers with dynamic token sparsification. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Yuzhang Shang. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *International Conference on Computer Vision (ICCV)*, volume 2025, 2025.

Guohao Sun, Yufei Wang, Sizhuo Ma, Yuege Xie, Yuting Cheng, Zhiqiang Tao, and Jian Wang. IF-Prune: Information-flow guided token pruning for efficient vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3522–3531, 2026.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025a.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. OminiControl2: Efficient conditioning for diffusion transformers, 2025b. URL https://arxiv.org/abs/2503.08280.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025b.

Minhao Xiong, Zichen Wen, Zhuangcheng Gu, Xuyang Liu, Rui Zhang, Hengrui Kang, Jiabing Yang, Junyuan Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Prune2Drive: A plug-and-play framework for accelerating vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25215–25224, 2026.

Tong Xu, Hailong Shi, and Xingyu Gao. SCoRe: Salience-coverage reduction for vision token pruning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24686–24695, 2026.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19803–19813, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. ATP-LLaVA: Adaptive token pruning for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24972–24982, 2025.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1126. URL https://aclanthology.org/2025.acl-long.1126/.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: heavy-hitter oracle for efficient generative inference of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Chao Zhou, Tianyi Wei, Yiling Chen, Wenbo Zhou, and Nenghai Yu. Rethinking multi-condition DiTs: Eliminating redundant attention via position-alignment and keyword-scoping, 2026. URL https://arxiv.org/abs/2602.06850.

Qihui Zhu, Tao Zhang, Yuchen Wang, Zijian Wen, Mengjie Zhang, Shuangwu Chen, Xiaobin Tan, Jian Yang, Yang Liu, Zhenhua Dong, Xianzhi Yu, and Yinfei Pan. HAWK: Head importance-aware visual token pruning in multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

A DETAILED BAGEL RESULTS ON DREAMOMNI2BENCH

Tables 6 and 7 summarize the BAGEL results, while Figure 6 shows the corresponding visual differences.

Table 6: DreamOmni2Bench summary for BAGEL (PF / SC / Overall).

		Editing			Generation			All		
Method	Keep	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.
Full tokens	1.000	2.468	4.756	2.927	5.947	6.500	5.794	3.712	5.379	3.951
CAT-Pruning(50%)	0.500	0.771	0.288	0.249	0.781	0.474	0.392	0.774	0.354	0.300
OminiControl2(50%)	0.500	1.376	2.107	1.288	3.026	3.123	2.743	1.966	2.470	1.808
SiTo(50%)	0.500	2.756	3.434	2.750	4.254	3.860	3.714	3.292	3.586	3.094
ToMA(50%)	0.500	1.873	2.278	1.568	0.860	1.895	0.737	1.511	2.141	1.271
ToMeSD(50%)	0.500	2.907	3.971	3.185	5.026	**4.018**	**4.171**	3.665	3.987	3.537
CARVE (ours,65%)	0.440	**3.522**	**4.015**	**3.246**	**5.211**	3.991	4.155	**4.125**	**4.006**	**3.571**

Table 7: DreamOmni2Bench results for BAGEL by reference-image number. Editing uses two references; generation uses one to four references.

	Editing-2ref			Gen-1ref			Gen-2ref			Gen-3ref			Gen-4ref		
Method	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.	PF	SC	Ov.
Full tokens	2.468	4.756	2.927	6.795	5.564	5.454	4.760	6.600	5.217	5.517	7.034	6.070	6.381	7.381	6.734
CAT-Pruning(50%)	0.771	0.288	0.249	0.821	0.590	0.290	0.880	0.400	0.433	0.931	0.621	0.645	0.381	0.143	0.182
OminiControl2(50%)	1.376	2.107	1.288	2.436	2.000	1.709	1.600	2.080	1.524	4.310	**4.690**	4.215	4.048	4.286	4.082
SiTo(50%)	2.756	3.434	2.750	4.872	3.744	3.704	3.320	3.760	3.192	4.069	3.897	3.778	4.476	4.143	4.262
ToMA(50%)	1.873	2.278	1.568	0.692	1.949	0.466	1.240	2.040	1.026	0.690	1.724	0.659	0.952	1.857	1.002
ToMeSD(50%)	2.907	3.971	3.185	5.744	3.949	4.137	3.960	**4.200**	**3.776**	**4.931**	4.103	**4.416**	5.095	3.810	4.364
CARVE (ours,65%)	**3.522**	**4.015**	**3.246**	**6.051**	**4.205**	**4.328**	**4.160**	3.760	3.713	4.621	3.655	3.843	**5.714**	**4.333**	**4.792**

Qualitative comparison. In the café example, CARVE preserves a coherent person with the reference-associated hairstyle while retaining the coffee, book, and indoor café context. In the motorcycle example, it transfers the vase's warm ceramic appearance to a recognizable vintage motorcycle without replacing the requested object or disrupting the cobblestone scene. In the rainy-street example, CARVE jointly maintains the trench coat, umbrella, street setting, and reference-conditioned facial expression. By contrast, the fixed methods may retain an isolated color or broad object cue but alter the subject or garment, break the object–material association, or produce visibly degraded images. The comparison shows that faithful condition compression must preserve the binding among subject, local attribute, and requested scene, rather than only a coarse visual resemblance.

Why fixed condition-token compression degrades. All compared methods compress only reference *condition tokens*; the noised target tokens are unchanged. Their failures therefore arise from an incomplete or distorted reference representation, rather than from pruning the target latent. This is particularly consequential for BAGEL, whose pooled condition sequence contains ViT tokens that carry semantic information such as subject and expression, together with clean-VAE tokens that carry localized appearance, material, and spatial detail. A fixed reduction can damage both forms of information at once.

ToMeSD. Its bipartite cosine matching assigns each source condition token to a similar destination token, averages their features, and deletes the sources. Feature similarity is not equivalent to redundancy in BAGEL's joint ViT–VAE condition sequence: visually similar tokens may still encode different object parts, reference regions, or different semantic and appearance information. The average then replaces a specific cue with a mixed representative, and the deleted token cannot supply the exact cue when it becomes important during denoising. The motorcycle row of Figure 6, where



Figure 6: Qualitative comparison on DreamOmni2Bench with BAGEL. Each row shows the reference image(s), instruction, and outputs from Full tokens and compression methods.

ToMeSD produces an unrelated portrait instead of the requested motorcycle, illustrates this loss of condition binding.

OminiControl2. It uniformly downsamples every reference before both the ViT and clean-VAE encoders. Fine material cues, boundaries, and local color transitions are consequently lost before they can enter the condition sequence, while the coarser semantic evidence is also less precise. Uniform reduction also spends the same limited spatial budget on instruction-critical objects and irrelevant background. The café and motorcycle rows in Figure 6 show the resulting severe visual degradation, and the rainy-street row shows that a broad scene cue alone is insufficient for faithful reference-conditioned generation.

The same limitation applies to one-shot score-based pruning and other fixed merge rules: they neither distinguish the semantic and appearance roles of BAGEL's condition tokens nor reassess their importance against the evolving target states across depth and sampling steps. As a result, they can preserve a coarse cue while removing the information required to bind the requested subject, attribute, and scene.

Efficiency. Table 8 measures a lower actual condition-token budget for CARVE, not the same compression setting as the fixed baselines: its mean keep is 0.440 rather than 0.500, and it processes 1,617 rather than 1,836 condition tokens on average. Even with this smaller condition budget, CARVE delivers the strongest compressed BAGEL quality in Tables 6 and 7; its visual gains are especially clear for localized texture transfer and reference-attribute binding. It also reduces BAGEL attention from 477.4 to 187.1 TFLOPs, a $2.55\times$ reduction from Full tokens. The acceleration differs from Qwen-Image because the backbones use different attention structures: Qwen-Image applies dense self-attention to the concatenated target and condition sequence, whose cost grows quadratically with total token length, $O\big((L_{\mathrm{tgt}} + L_{\mathrm{cond}})^2\big)$. BAGEL instead packs ViT and clean-VAE condition tokens with text and target tokens under a structured 2D attention mask, for which added reference-condition computation grows approximately linearly with L_{cond}. Thus, the same Router design provides backbone-specific speed gains while consistently concentrating the smaller condition budget on the reference details that most improve the generated image.

Efficiency calculation. For the BAGEL results in Table 8, we estimate attention computation from the processed sequence lengths. For each case, we record the sequence length $S_{c,f,\ell}$ actually processed by attention layer ℓ in every denoising forward f. The reported attention cost is the case

average

$$\text{Attn. TFLOPs} = \frac{2d}{10^{12}|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{f=1}^{F} \sum_{\ell=1}^{L} S_{c,f,\ell}^2, \tag{10}$$

where $F{=}3{\times}49$ counts the three CFG forwards at each of 49 denoising steps, and $(d, L){=}(3584, 28)$ are BAGEL's hidden width and number of attention layers. The factor 2 counts the $QK^\top$ and attention–value products; VAE/ViT encoding, projection, and MLP cost are excluded consistently for every method. For CARVE, $S_{c,f,\ell}$ is measured after each Router's physical deletion, so later layers use the shortened sequence; the same actual-length accounting is applied to every baseline. We report $\bar{N}_c = \frac{1}{|\mathcal{C}|L} \sum_{c,\ell} N_{c,\ell}^{\text{cond}}$, the mean condition-token count across cases and layers, and define Mean Keep as $\bar{N}_c/\bar{N}_c^{\text{full}}$. Attn. speedup is $\text{TFLOPs}_{\text{full}}/\text{TFLOPs}_{\text{method}}$, computed separately against the Full tokens run of the same backbone.

Router selection visualization. Figure 7 shows that the Router selects structured condition content rather than isolated, arbitrary tokens. In both the VAE and ViT maps, high-score tokens form coherent regions over the person and the T-shirt, which are the reference details the model needs to perform the requested local texture transfer. At the first Router layer, both modalities retain broad reference coverage. As the cascade proceeds through later layers, the navy regions expand because already-removed tokens no longer enter the following blocks, while the remaining high-score regions continue to follow the relevant subject structure. The maps also change across sampling steps, showing that the Router selects the condition information needed by the current denoising state instead of reusing a static spatial mask.

B Implementation Details

Router parameterization. We use five independent condition Routers. Each Router mean-pools a linear projection of the noised target tokens, adds it to a per-condition-token projection, and applies a hidden linear layer followed by a scalar score head. The Router hidden width is 2,048. For BAGEL, whose language-model hidden size is 3,584, this gives 18.88M parameters per Router (94.41M for five Routers); the frozen BAGEL backbone is otherwise unchanged. DDS places the five BAGEL Routers at layers $[6, 8, 13, 20, 21]$. For Qwen-Image, the Router input width is 3,072, giving 16.79M parameters per Router (83.93M in total), and DDS selects blocks $[21, 22, 43, 52, 57]$. For both backbones, only the Router parameters are trained; the diffusion backbone, text encoder, and VAE remain frozen.

Training configuration. Both backbones are trained on the same 10K-example MICo training subset with $8\times$ NVIDIA H200 GPUs and bfloat16 precision. At every training forward pass, we sample a base budget $B \sim \mathcal{U}(0.3, 1.0)$ and use the exponential cascade schedule B^i at the i-th Router ($i = 1, \ldots, 5$). BAGEL uses full-shard FSDP, a constant learning rate of 5×10^{-5} with

Table 8: Reference-token processing efficiency on DreamOmni2Bench with BAGEL at 512^2. Input Keep is the configured condition-token budget. Mean Keep and $\bar{N}_c$ are averaged over all 319 cases. Attention FLOPs sum the three CFG forwards over 49 denoising steps and 28 BAGEL blocks; speedup is relative to Full tokens. Values use the actual per-case packed token counts and each method's configured sequence schedule.

Method	Input Keep	Mean Keep ↓	Mean $\bar{N}_c$ ↓	Attn. TFLOP ↓	Attn. speedup ↑
Full tokens	1.00	1.000	3,672	477.4	1.00×
CAT-Pruning	0.50	0.500	1,836	181.5	2.63×
OminiControl2	0.50	0.500	1,836	181.5	2.63×
SiTo	0.50	0.500	1,836	181.5	2.63×
ToMA	0.50	0.500	1,836	181.5	2.63×
ToMeSD	0.50	0.500	1,836	181.5	2.63×
CARVE (ours)	0.65	0.440	1,617	187.1	2.55×

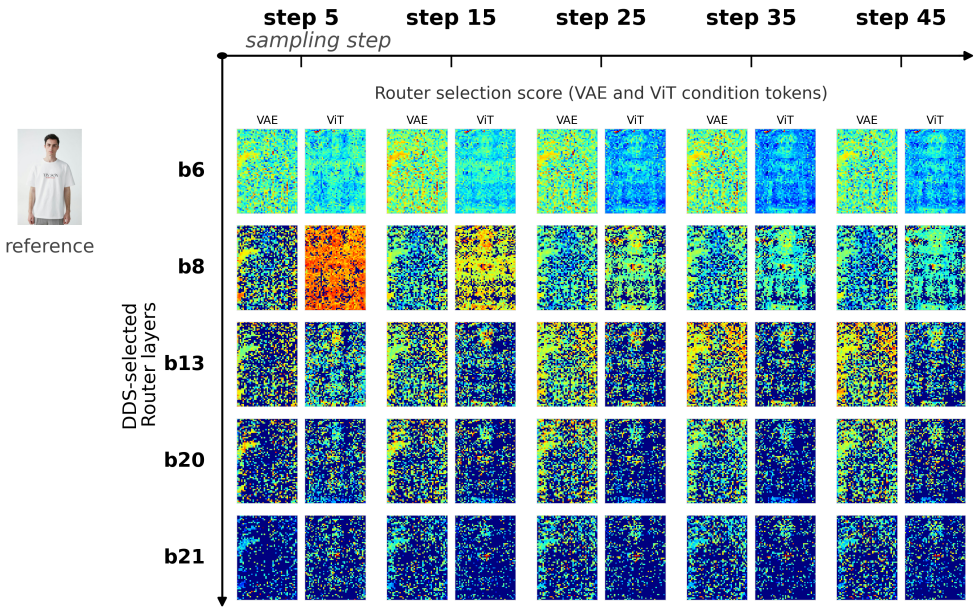


Figure 7: BAGEL Router selection scores for the T-shirt texture example. Rows are the DDS-selected Router layers and columns are sampling steps. Each step shows the VAE and ViT condition-token scores; navy marks tokens removed by preceding Routers.

100 warm-up steps, gradient accumulation of one, gradient clipping at 1.0, and equal flow-matching and distillation losses; only the Router parameters are optimized for 20,000 steps. Its VAE and ViT training-image maxima are 512 and 518 pixels, respectively, with an expected packed length of 8,192 tokens (at most 16,000). Qwen-Image uses AdamW with a constant learning rate of 10^{-5}, gradient accumulation of eight, gradient checkpointing, and equal flow-matching and distillation weights. It trains for one epoch at a maximum of $1{,}048{,}576$ image pixels; with 10K examples, $8\times$ NVIDIA H200 GPUs, and accumulation eight, this corresponds to approximately 157 optimizer updates.

C Efficiency Metrics for Qwen-Image

For Qwen-Image, we evaluate reference-token processing efficiency and DiT latency on 319 DreamOmni2Bench cases, with approximately 512^2-area reference images on an NVIDIA A100 GPU. Input Keep is the configured input keep ratio. Mean Keep is the measured mean fraction of reference-condition tokens processed by attention, and $\bar{N}_c$ is the corresponding mean token count. DiT latency, t_{DiT}, is the mean denoising-model execution time in seconds per case.

The reference-token removal rate is computed as

$$\mathrm{Removed/s} = \frac{\bar{N}_c^{\mathrm{Full}} - \bar{N}_c}{t_{\mathrm{DiT}}}, \tag{11}$$

where $\bar{N}_c^{\mathrm{Full}}$ is the mean condition-token count of the uncompressed model under the same evaluation setting. This metric measures the reduction in reference-condition tokens normalized by DiT latency. For the evaluated Qwen-Image CAT-Pruning implementation, Input Keep controls the fraction of noisy target tokens updated, while reference-condition tokens remain intact; its reference-token removal rate is therefore zero despite its reduced attention computation.

D DISCUSSION

Our primary objective is to improve inference efficiency for multi-reference generation while retaining the reference information required for faithful outputs. For Qwen-Image and BAGEL, the respective 47% and 65% input keep ratios support quality comparisons at comparable or greater efficiency than the baselines, measured through DiT latency for Qwen-Image and attention-FLOPs reduction for BAGEL. The observed quality gains over competing compression methods are an additional benefit beyond this efficiency objective, showing that compression can offer a favorable quality–efficiency trade-off.