WHEN VISUAL PROMPTS ROTATE: SPECTRAL GUIDANCE FOR TEST-TIME IN-CONTEXT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual in-context learning (VICL) predicts dense outputs from image–annotation demonstrations. Rotating a prompt image and its annotation together preserves the demonstrated task, yet can degrade predictions for an unchanged query. This prompt-only mismatch changes spatial relationships and introduces interpolation, aliasing, and padding; a matched padding control shows that boundary fill alone does not explain the degradation. We propose **GIST** (Graph-Informed Spectral Tuning), a test-time adaptation method that adds graph-guided attention and error weighting to VICT's support-reconstruction cycle. Spectral Relation Guidance (SRG) uses distances in a low-frequency graph subspace to bias token interactions during both query prediction and support reconstruction. Spectral Adaptation Guidance (SAG) defines a reconstruction metric on a fixed support graph and adjusts its spectral contribution using the preceding residual's frequency composition. The two components guide information exchange and parameter updates without query labels, additional trainable modules, or rotation-specific backbone pretraining. Across 15 corruptions, two prompt settings, and four depth/segmentation benchmarks, GIST reduces average Abs.Rel by 9.9% to 27.3% relative to VICT and increases average mIoU by 0.40 to 1.14 points under matched rotations. Component ablations, attention-bias controls, and geometric diagnostics support these two intervention points while showing that graph structure remains sensitive to real image transformations.

1 INTRODUCTION

In-context learning replaces task-specific retraining with inference from demonstrations (Brown et al., 2020). Visual prompting extends this interface to perception (Bar et al., 2022; Wang et al., 2023a;b): Painter arranges a prompt image, its annotation, a query image, and a blank target in a shared completion canvas. The prompt thus specifies both the task and the spatial reference used to relate query regions to demonstrated outputs. This dual role enables generalist dense prediction, but also permits the semantics to remain valid while the prompt geometry changes.

Reusable demonstrations can differ in orientation from later observations, for example in underwater imaging and endoscopy (Islam et al., 2020; Stoyanov et al., 2010). As Figure 1 illustrates, even a small mismatch can cause a visible prediction failure. Rotating the prompt image and annotation together preserves the demonstrated task while the query remains unchanged, yet shifts patch indices and relative coordinates and introduces interpolation, aliasing, padding, and graph perturbations. A matched padding-only control is less damaging than the full rotation transform. Because that control does not isolate reindexing from other resampling effects, we regard these factors as jointly corrupting the prompt's spatial reference rather than assign a single cause.

Existing robustness mechanisms do not directly resolve this asymmetric setting. Rotation augmentation and equivariant models build tolerance during training and generally transform the full input distribution (Cohen & Welling, 2016; Worrall et al., 2017; Kundu & Kondor, 2025); test-time canonicalization instead searches for a typical input view (Singhal et al., 2025). General test-time adaptation handles unlabeled distribution shift (Sun et al., 2020; Wang et al., 2021; Gandelsman et al., 2022), and VICT adapts VICL by reversing prompt and query roles to reconstruct the known prompt annotation (Xie et al., 2025). Its reconstruction objective does not explicitly encode patch-

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.

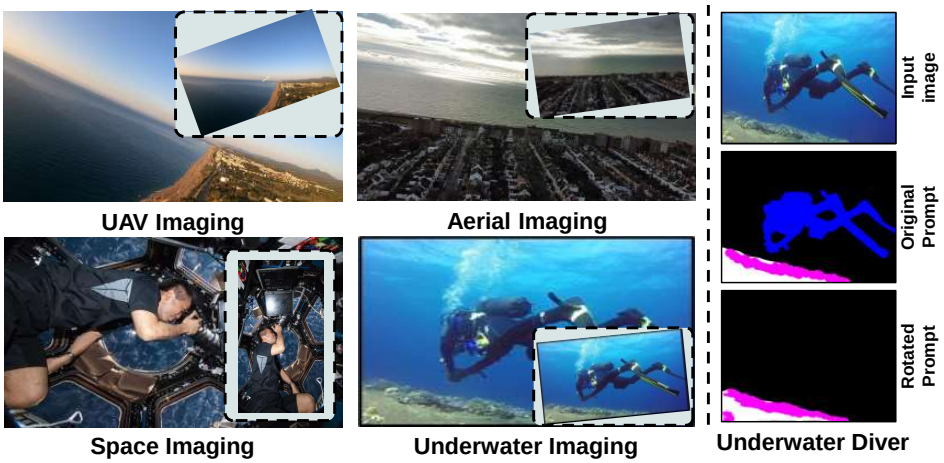


Figure 1: Prompt orientation mismatch: motivating scenarios and a prediction example. *Left:* Illustrative acquisition scenarios in which a stored prompt and a query may differ in orientation. *Right:* With the query unchanged, rotating only the prompt causes the visual in-context model to lose the diver segmentation, showing that a semantics-preserving prompt rotation can substantially degrade prediction.

graph relationships, and its original token interactions remain sensitive to prompt geometry. For example, one-shot NYUv2-C error rises from 0.205 to 0.342 under prompt rotation (Table 1).

We build on a deliberately limited spectral premise. A feature-affinity graph represents relations among patches rather than absolute grid coordinates, and its low-frequency Laplacian subspace summarizes coarse structure (Belkin & Niyogi, 2003; Kreuzer et al., 2021). This subspace is not rotation invariant: resampling perturbs the graph, and eigenvectors have sign and basis ambiguities (Lim et al., 2023; Huang et al., 2024). GIST therefore uses basis-invariant eigenspace distances only as a soft cue. It adds two forms of guidance inside VICT: SRG rebuilds temporary token graphs to guide attention in both completions, while SAG uses a fixed support graph to define a low-pass reconstruction metric whose weight follows the preceding residual spectrum. The former guides token interaction and the latter reshapes the adaptation signal, without query annotations or rotation-specific training (Sections 3.2 and 3.3).

Across NYUv2-C, ADE20K-C, Hamlyn-C, and SUIM-C, GIST improves the aggregate metric over VICT under matched random rotations in both prompt settings, while category-level exceptions preclude a universal claim. Clean-input and mechanism-specific controls further delimit the observed gains.

Contributions:

- We formulate prompt-only rotation as a distinct test-time mismatch and characterize its effect without assuming a single geometric cause.
- We propose **GIST**, combining graph-relative attention with adaptive spectral error weighting inside VICT's support-reconstruction cycle.
- Two standard and two real-domain benchmarks show improved average robustness, supported by clean-input and mechanism-specific controls.

2 Related Work

Visual In-Context Learning and Test-Time Adaptation. Visual prompting casts demonstrations and queries as image inpainting (Bar et al., 2022); Painter (Wang et al., 2023a) and SegGPT (Wang et al., 2023b) extend this interface to general dense prediction. Test-time training and adaptation update a model on unlabeled target samples using self-supervision, entropy, or prompt objectives (Sun et al., 2020; Gandelsman et al., 2022; Wang et al., 2021; Shu et al., 2022). VICT (Xie et al., 2025) makes this idea native to VICL: it exchanges the query and prompt roles, then tunes the model by reconstructing the known prompt output. GIST retains that label-free support-reconstruction cycle but targets sensitivity to prompt orientation by introducing graph relationships into attention and the reconstruction metric.

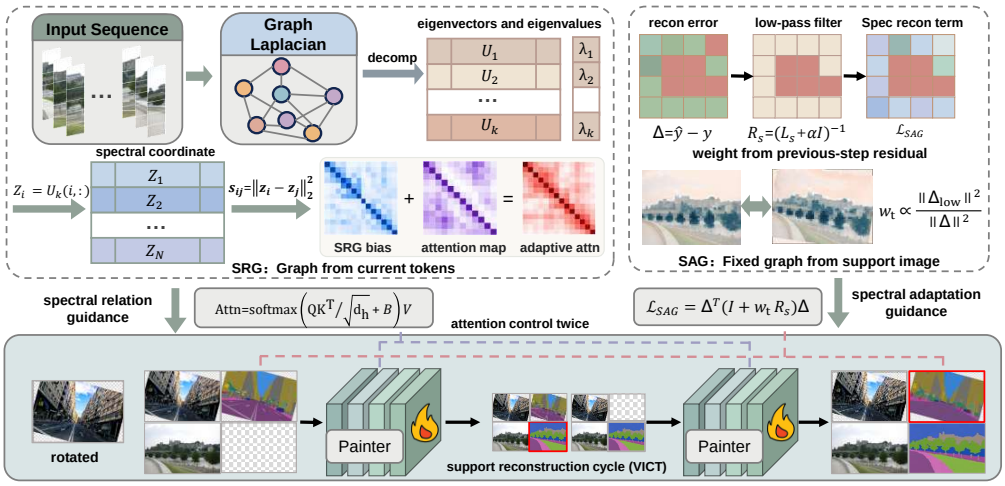


Figure 2: Overview of GIST. VICT first predicts the query annotation and then exchanges the prompt/query roles to reconstruct the known support annotation, providing label-free test-time supervision. GIST adds SRG, which rebuilds temporary token graphs and supplies a soft graph-relative attention bias in both completions, and SAG, which uses a fixed support graph to shape the reconstruction update from the preceding residual spectrum. No query annotation is used for adaptation.

Rotation Robustness. Group-equivariant and steerable CNNs encode rotation symmetry during training (Cohen & Welling, 2016; 2017; Worrall et al., 2017), with recent extensions to Transformers (Kundu & Kondor, 2025; Karella et al., 2024). These methods redesign or retrain the representation. FoCal instead searches over transformed inputs at test time and selects a canonical view using foundation-model priors (Singhal et al., 2025). Our setting is narrower but structurally different: only the task prompt is rotated, its annotation must undergo the same transform, and the query remains unchanged. GIST therefore regularizes prompt–query interaction rather than enforce global equivariance or select a canonical query view.

Spectral Graph Methods for Vision. Laplacian methods encode graph topology in low-frequency coordinates (Belkin & Niyogi, 2003), and graph neural networks interpret message passing through spectral filtering (Kipf & Welling, 2017). Spectral Transformers and positional encodings inject this structure while confronting eigenspace ambiguity and perturbation sensitivity (Kreuzer et al., 2021; Wang et al., 2022; Lim et al., 2023; Huang et al., 2024). Self-supervised ViT features exhibit structural clustering suitable for graph affinities (Caron et al., 2021). GSDTTA adapts 3D point clouds by optimizing low-frequency graph components (Wei et al., 2025). GIST neither learns a rotation-equivariant spectral encoder nor shifts input geometry; it constructs per-sample graph cues that control attention and support reconstruction inside Painter's cycle.

3 Method

GIST retains Painter's prediction interface and VICT's support-reconstruction cycle. The cycle supplies supervision from the known support annotation. Within it, SRG guides token interaction using temporary graphs, while SAG guides parameter updates using a fixed support-graph metric.

3.1 Task Setup and Reconstruction Cycle

Let (X_s, Y_s) denote a support image–annotation prompt and (X_q, Y_q) a query pair, where Y_q is unavailable at test time. We apply the same rotation to X_s and Y_s, leaving X_q unchanged; below, the support symbols denote this observed, transformed pair. Let $P_\theta((X_a, Y_a), X_b)$ denote Painter completion conditioned on (X_a, Y_a) with the annotation cell of X_b masked. The ordinary query prediction is

$$\hat{Y}_q = P_\theta((X_s, Y_s), X_q). \tag{1}$$

Painter implements the completion in Eq. (1) with a conceptual 2×2 image–annotation canvas and processes its image and annotation streams separately before early fusion. VICT makes this

interface usable for test-time adaptation by exchanging the roles of the query and support. Each step invokes Painter twice:

$$\hat{Y}_q = P_\theta((X_s, Y_s), X_q), \quad \hat{Y}_s = P_\theta((X_q, \hat{Y}_q), X_s). \tag{2}$$

The first completion is the prediction of interest; the second uses that prediction as a temporary annotation and reconstructs the known support annotation Y_s. Thus Y_s, rather than the unavailable Y_q, supplies the self-supervision signal. We initialize an independent parameter copy for each X_q, update only that copy's encoder, and reset it for the next query. GIST preserves this VICT cycle and adds graph cues without changing tokenization or requiring a query label.

3.2 Spectral Relation Guidance

SRG addresses the first question in the cycle: which token relationships should guide each completion? It constructs a temporary feature graph from the current stream, extracts a low-frequency Laplacian subspace, and uses distances in that subspace as a soft attention bias. The graph is recomputed for each completion and adaptation step, so it is a relation cue rather than a fixed positional encoding or a strict rotation-invariance claim.

Affinity graph construction. Here $r = \mathrm{fwd}$ predicts the query annotation, while $r = \mathrm{rev}$ reconstructs the support annotation. At injection layer $\ell \in \{0, 16\}$ and active stream s, let $H^{(r,\ell,s)} = [h_1, \ldots, h_{N_\ell}]^\top \in \mathbb{R}^{N_\ell \times C_\ell}$ denote the input tokens, where h_i is the i-th token feature. Layer 0 has separate image and annotation streams; layer 16 uses their fused representation. For each stream we compute

$$A_{ij} = \exp\left(-\frac{\|h_i - h_j\|_2^2}{2\sigma_{\mathrm{aff}}^2}\right). \tag{3}$$

In Eq. (3), σ_{aff} is the mean non-self pairwise token distance of this stream. We mask A_{ii} before retaining the k_g=16 most similar non-self neighbors and symmetrize with $A = \frac{1}{2}(\tilde{A} + \tilde{A}^\top)$, yielding a sparse feature-affinity graph. With $D = \mathrm{diag}(A\mathbf{1})$, the symmetric normalized Laplacian is

$$L = I - D^{-1/2} A D^{-1/2}. \tag{4}$$

Thus, the affinity in Eq. (3) and Laplacian in Eq. (4) satisfy $A^{(r,\ell,s)}, L^{(r,\ell,s)} \in \mathbb{R}^{N_\ell \times N_\ell}$ and match the corresponding attention logits. At layer 0, H includes Painter's native patch, segment, and positional embeddings; graph construction uses $\mathrm{stopgrad}(H)$, while the ordinary token path remains differentiable and no query annotation is used.

Graph-relative spectral embedding. An ideal node reindexing satisfies $L' = PLP^\top$. A real prompt rotation need not satisfy this identity because interpolation, padding, changed graph edges, and feature drift perturb the Laplacian (Huang et al., 2024). We therefore use low-frequency coordinates as a graph-relative cue rather than assume rotation equivariance. For a connected graph, the first eigenvector of the symmetric normalized Laplacian is the trivial degree-weighted constant mode, $u_1 \propto D^{1/2}\mathbf{1}$, with eigenvalue zero. We discard u_1 and retain the next eight ordered eigenvectors, $U_m = [u_2, \ldots, u_{m+1}]$ with m=8; disconnected graphs can retain additional zero modes under this fixed-index policy. Individual eigenvectors have sign and basis ambiguities (Lim et al., 2023), but pairwise distances in the retained subspace are unchanged by orthogonal basis changes. With $e_i = U_m[i, :]$, we define

$$B_{ij} = \mathrm{RowStd}_i\big(-\|e_i - e_j\|_2^2\big). \tag{5}$$

A larger B_{ij} in Eq. (5) favors graph-relative proximity. Row standardization is applied after all pairwise distances are formed.

Attention injection. The resulting $B^{(r,\ell,s)} \in \mathbb{R}^{N_\ell \times N_\ell}$ is broadcast across heads and added to the same-shaped pre-softmax logits:

$$\mathrm{Attn} = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + C_{\mathrm{pos}} + B\right) V, \tag{6}$$

Here d_h is the head dimension and C_{pos} denotes Painter's native relative-position logit term, which is retained. Equation (6) adds B before softmax. The detached graph and embedding are discarded after that attention call. The injection locations are evaluated in the supplementary ablation.

3.3 Spectral Adaptation Guidance

SRG controls token interaction in both completions; SAG addresses a separate question: how should the known support-reconstruction residual shape the adaptation signal? For this purpose, we construct a separate graph from patch features of the transformed support image. Its affinity A_s uses the same Gaussian kernel, self-edge masking, top-k_g selection, and symmetrization as the temporary SRG graph:

$$L_s = I - D_s^{-1/2} A_s D_s^{-1/2} = U_s \Lambda_s U_s^\top. \tag{7}$$

Unlike the temporary SRG graphs, the support Laplacian in Eq. (7) is cached and fixed throughout the query episode. Let $\Delta_t \in \mathbb{R}^{N_s \times D_p}$ denote the support reconstruction residual $Y_s - \hat{Y}_s$ arranged on the same N_s-patch grid, with each row retaining the encoded residual within one patch. Ordinary support reconstruction uses the identity metric,

$$\mathcal{L}_{\text{recon}} = \operatorname{tr}(\Delta_t^\top \Delta_t) = \|\Delta_t\|_F^2. \tag{8}$$

This is the usual squared support-reconstruction error, which weights all graph frequencies equally.

We replace this flat metric with a sample-adaptive spectral metric using

$$R_s = (L_s + \alpha I)^{-1} = U_s (\Lambda_s + \alpha I)^{-1} U_s^\top, \tag{9}$$

where $\alpha > 0$ makes the inverse in Eq. (9) finite at the zero-eigenvalue trivial mode. Its response $1/(\lambda_{s,k}+\alpha)$ emphasizes graph-low-frequency error. Define the spectral term $\mathcal{L}_{\text{spec}} = \operatorname{tr}(\Delta_t^\top R_s \Delta_t)$. The complete SAG loss is

$$\begin{aligned} \mathcal{L}_{\text{SAG}} &= \mathcal{L}_{\text{recon}} + w_t \mathcal{L}_{\text{spec}} = \operatorname{tr}\big[\Delta_t^\top (I + w_t R_s) \Delta_t\big] \\ &= \sum_k \left(1 + \frac{w_t}{\lambda_{s,k} + \alpha}\right) \|\widehat{\Delta}_{t,k,:}\|_2^2, \end{aligned} \tag{10}$$

where $\widehat{\Delta}_t = U_s^\top \Delta_t$. Equation (10) replaces the identity metric in Eq. (8) with $I + w_t R_s$ and recovers Eq. (8) when $w_t = 0$. The implemented resolvent retains this regularized trivial mode, so once $w_t > 0$ it also reweights the degree-weighted global residual component. By contrast, the statistic below excludes u_1; this exclusion affects the adaptive trigger, not the modes present in the reconstruction metric.

Adaptive weight w_t. Graph-low-frequency residuals vary slowly across strongly connected patches after degree normalization; they need not coincide with spatially smooth errors. SAG emphasizes these components as a design choice and reduces the added weight when they account for little residual energy. The weight measures residual composition, not graph reliability or whether an input is clean.

At $t = 0$, no previous error is available, so $w_0 = 0$ and Eq. (10) reduces to $\mathcal{L}_{\text{recon}}$ for the first update.

From step $t \geq 1$ onward, we compute w_t from the spectral concentration of the previous step's support residual. Let $\bar{\Delta}_{t-1} = \operatorname{stopgrad}(\Delta_{t-1})$ and $\bar{\widehat{\Delta}}_{t-1} = U_s^\top \bar{\Delta}_{t-1}$. Using support-graph mode indices 2 through $m{+}1$ with $m{=}8$, the concentration ratio and adaptive weight are

$$\rho_t = \frac{\sum_{k=2}^{m+1} \|\bar{\widehat{\Delta}}_{t-1,k,:}\|_2^2}{\|\bar{\Delta}_{t-1}\|_F^2 + \varepsilon}, \qquad w_t = w_{\max}\, \rho_t, \tag{11}$$

where the trivial mode u_1 is excluded, $w_{\max}{=}1$, and ε prevents division by zero. The ratio $\rho_t \in [0, 1]$ measures the fraction of previous residual energy in the retained nontrivial support-graph modes with approximate scale invariance when the residual energy dominates ε. It is a sample-specific structural-error statistic, not an angle estimator.

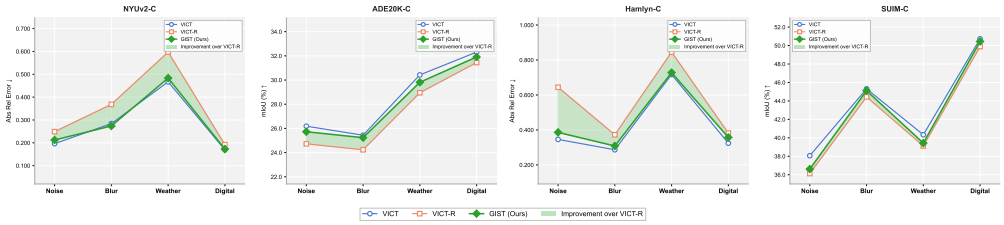


Figure 3: Corruption-category trends across four datasets. Each point is the unweighted mean over the corruption types in that category and the zero-/one-shot settings. Under identical randomly rotated prompts, GIST improves the category-averaged results over VICT-R with dataset- and category-level variation.

3.4 Test-Time Update and Prediction

We now combine SRG and SAG within the VICT cycle. Let $\hat{Y}_s(\theta')$ denote the support reconstruction from Eq. (2), including SRG biases recomputed from the current streams, while (L_s, R_s) remain fixed. With $\theta_q^{(0)} = \theta$, we optimize the sample-specific encoder while keeping the decoder fixed:

$$\begin{aligned} \theta_q^{(t+1)} &= \mathrm{AdamW}\Big(\theta_q^{(t)}, \nabla_{\theta_q^{(t)}} \mathcal{L}_{\mathrm{SAG}}^{(t)}\Big), \\ \hat{Y}_q^* &= P_{\theta_q^{(T)}}((X_s, Y_s), X_q), \qquad T = 60. \end{aligned} \tag{12}$$

Optimizer state is implicit in Eq. (12). At every step, Eq. (11) sets the weight in Eq. (10), which updates the sample-specific encoder through both completions. Graph construction and the previous-step statistic are detached. The final query prediction recomputes SRG and returns $\hat{Y}_q^*$; SAG affects the update through support reconstruction but is not a prediction head. At $w_0 = 0$, only the loss reduces to ordinary reconstruction; SRG remains active. The adapted copy is reset before the next query.

4 Experiments

We evaluate whether GIST improves robustness to prompt rotation without sacrificing Painter's general task interface.

4.1 Standard Benchmark Evaluation

NYUv2-C evaluates depth estimation with AbsRel (↓), and ADE20K-C evaluates semantic segmentation with mIoU (↑). Following the common-corruption protocol (Hendrycks & Dietterich, 2019), both contain 15 corruption types spanning noise, blur, weather, and digital artifacts at severity 5, providing different appearance shifts on top of prompt orientation mismatch.

Experimental setup. We use Painter with a ViT-Large encoder and clone its pretrained parameters for independent per-sample adaptation. Following VICT's one-prompt protocol (Xie et al., 2025), we use one paired prompt (N=1) per dataset and prompt setting and keep its identity fixed across queries and methods for paired comparison. Zero-shot prompts come from the clean training distribution, whereas one-shot prompts share the query corruption setting. VICT and GIST use identical query samples, prompt transformations, rotation draws, optimizer, and 60-step budget. In tables, the suffix -R marks the rotated-data condition for either method. The long-tailed angle distribution emphasizes small rotations: 70% in $[0°, 10°)$, 25% in $[10°, 30°)$, 4% in $[30°, 90°)$, and 1% in $[90°, 180°]$, with direction sampled symmetrically. We use $w_{\max}$=1, graph degree k_g=16, the eight ordered modes after u_1, and α=0.1. Optimizer and adaptation details are provided in the supplementary material.

Results. Table 1 shows an aggregate trend across depth and segmentation: prompt rotation weakens VICT, whereas GIST improves over VICT under the same rotated-data condition in both prompt settings and in many individual corruption categories. Category-level exceptions remain, so the result supports an average robustness gain rather than universal improvement. Figure 3 summarizes these per-category trends across all four evaluated datasets.

Table 1: Results across 15 corruption types on four datasets: (a) NYUv2-C depth, (b) ADE20K-C segmentation, (c) Hamlyn-C depth, and (d) SUIM-C segmentation. The suffix -R denotes evaluation after prompts and annotations are randomly rotated by the same angle distribution; it is an evaluation condition, not a method variant. Green/red cells indicate improvement/degradation relative to VICT-R.

Method	Bright	Cont	Defoc	Elast	Fog	Frost	Gauss	Glass	Impul	JPEG	Motion	Pixel	Shot	Snow	Zoom	Avg.
(a) Depth Estimation on NYUv2-C (Abs.Rel ↓)																
zero-shot setting:																
Painter	0.129	0.215	0.712	0.109	0.129	0.536	0.200	0.612	0.189	0.386	0.167	0.187	0.142	1.951	0.209	0.392
VICT	0.102	0.220	0.661	0.113	0.153	0.531	0.191	0.531	0.180	0.331	0.132	0.120	0.138	1.836	0.224	0.365
VICT-R	0.102	0.227	0.672	0.112	0.162	0.542	0.215	0.533	0.207	0.342	0.141	0.125	0.140	1.869	0.235	0.375
GIST-R	0.098	0.223	0.416	0.116	0.159	0.569	0.207	0.389	0.215	0.258	0.133	0.111	0.179	1.633	0.248	0.330
one-shot setting:																
Painter	0.126	0.285	0.743	0.109	0.132	0.853	0.901	0.622	0.901	0.392	0.174	0.194	0.440	1.964	0.212	0.537
VICT	0.097	0.193	0.180	0.107	0.128	0.227	0.241	0.210	0.278	0.159	0.121	0.114	0.150	0.662	0.214	0.205
VICT-R	0.116	0.230	0.539	0.110	0.143	0.280	0.361	0.494	0.368	0.272	0.122	0.118	0.202	1.564	0.210	0.342
GIST-R	0.097	0.201	0.358	0.118	0.164	0.295	0.243	0.309	0.243	0.245	0.126	0.110	0.187	0.852	0.211	0.251
(b) Semantic Segmentation on ADE20K-C (mIoU ↑)																
zero-shot setting:																
Painter	40.40	11.70	27.90	31.20	35.00	23.30	24.80	25.60	26.10	38.60	31.80	40.40	25.90	9.40	13.10	27.00
VICT	40.90	12.70	28.20	31.70	36.40	23.70	25.50	26.30	27.40	38.80	31.90	40.90	28.00	17.30	13.60	28.20
VICT-R	40.32	15.13	28.12	30.69	37.69	25.20	26.00	25.85	26.75	38.77	27.06	40.04	26.83	13.88	13.40	27.72
GIST-R	40.55	15.29	28.74	30.17	37.07	25.24	25.35	26.02	27.85	38.76	32.80	40.22	26.27	15.32	13.62	28.22
one-shot setting:																
Painter	40.90	11.70	27.50	31.30	35.00	22.10	19.70	25.80	20.30	38.50	31.30	40.30	23.20	6.50	13.30	25.80
VICT	41.70	21.00	28.80	32.20	37.00	24.70	24.40	27.00	25.20	39.70	32.80	41.50	26.60	21.70	14.90	29.30
VICT-R	40.88	17.34	27.85	30.53	36.53	25.18	22.71	25.51	22.06	38.89	32.61	40.17	24.01	11.90	13.47	27.31
GIST-R	40.36	20.33	28.19	31.26	37.61	25.09	23.78	25.43	24.77	38.69	32.85	40.53	26.32	17.28	14.22	28.45
(c) Depth Estimation on Hamlyn-C (Abs.Rel ↓)																
zero-shot setting:																
Painter	0.695	0.470	0.393	0.410	1.168	0.643	1.335	0.485	1.320	0.788	0.375	0.415	1.198	2.448	0.380	0.835
VICT	0.278	0.380	0.270	0.336	0.621	0.553	0.439	0.283	0.465	0.253	0.313	0.306	0.312	1.624	0.230	0.444
VICT-R	0.318	0.496	0.440	0.354	0.547	0.594	0.481	0.330	0.544	0.258	0.432	0.480	0.355	1.827	0.302	0.517
GIST-R	0.293	0.452	0.275	0.390	0.705	0.561	0.453	0.295	0.465	0.263	0.329	0.415	0.389	1.425	0.278	0.466
one-shot setting:																
Painter	0.858	0.830	0.445	0.420	0.943	0.713	1.315	0.540	1.308	0.830	0.388	0.448	1.165	2.420	0.418	0.869
VICT	0.288	0.341	0.257	0.376	0.614	0.383	0.289	0.279	0.306	0.245	0.308	0.356	0.265	1.377	0.353	0.402
VICT-R	0.294	0.417	0.376	0.374	0.710	0.574	0.774	0.376	0.927	0.253	0.302	0.424	0.785	1.899	0.421	0.594
GIST-R	0.299	0.335	0.260	0.363	0.643	0.395	0.330	0.285	0.365	0.260	0.411	0.380	0.315	1.502	0.334	0.432
(d) Semantic Segmentation on SUIM-C (mIoU ↑)																
zero-shot setting:																
Painter	57.60	28.12	47.77	61.83	42.16	26.72	31.29	47.48	33.48	51.16	47.17	57.94	32.30	17.46	33.55	41.07
VICT	60.21	27.62	48.42	63.58	44.10	29.32	36.05	48.91	38.96	52.76	47.89	58.55	36.63	25.32	34.20	43.50
VICT-R	59.35	27.69	47.83	62.92	42.25	28.77	34.10	48.32	37.05	51.73	47.60	58.21	35.14	25.68	33.71	42.69
GIST-R	59.21	28.36	48.29	62.53	43.17	29.12	33.89	48.47	38.56	52.78	48.51	58.31	35.71	25.33	34.23	43.09
one-shot setting:																
Painter	57.66	27.69	45.41	61.74	42.50	26.26	20.93	48.49	22.61	51.35	46.70	57.90	24.29	13.25	33.78	38.70
VICT	60.12	28.25	49.53	62.46	45.85	31.45	38.89	49.83	39.89	53.35	48.51	59.27	37.94	26.29	35.20	44.46
VICT-R	59.18	27.32	47.89	61.46	44.77	30.59	34.87	49.04	38.75	52.49	47.89	57.31	36.85	22.34	33.23	42.93
GIST-R	58.11	28.23	48.12	62.30	43.96	30.62	37.15	49.55	37.92	52.85	48.10	58.22	36.44	25.89	35.62	43.54

4.2 Domain-Specific Evaluation

Controlled rotation alone does not establish that the problem remains relevant outside standard benchmarks. We therefore test SUIM (Islam et al., 2020) and Hamlyn (Stoyanov et al., 2010), two real-world domains whose acquisition processes are rotation-prone. We apply the same controlled prompt-rotation protocol to isolate orientation mismatch.

Experimental setup. Because both datasets lie outside Painter's pretraining distribution, comparing methods directly from the generic checkpoint would confound domain knowledge with test-time adaptation. For fairness, every method starts from the same dataset-specific checkpoint and receives the same prompt and corruption. Only the per-sample adaptation strategy differs. The updated blocks, objectives, schedule, and training details are deferred to the supplementary material.

Results. The lower part of Table 1 shows the real-world evaluation results. GIST improves the aggregate metric over VICT under rotation in both domains and prompt settings, although gains are

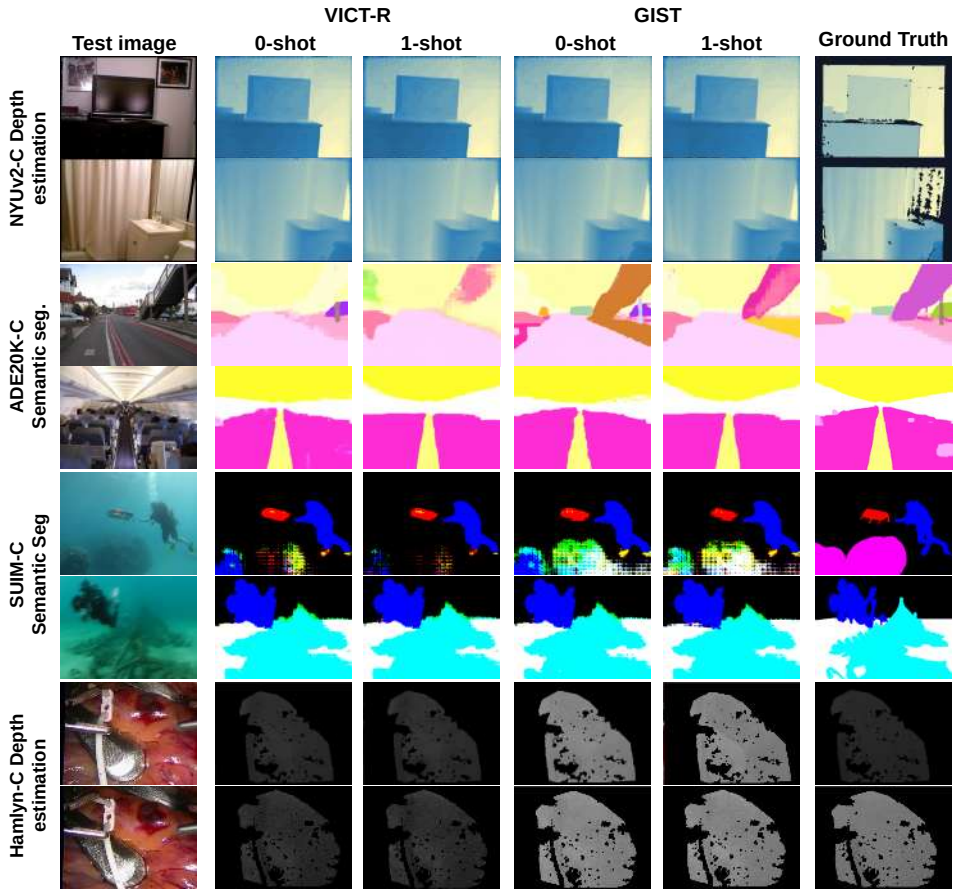


Figure 4: Representative predictions under rotated prompts across the four benchmarks. Compared with VICT, GIST better preserves region boundaries and suppresses prediction artifacts, particularly in the one-shot setting.

Table 2: Component ablation on rotated NYUv2-C (one-shot, Abs.Rel ↓).

Configuration	SRG	SAG	Avg.↓
VICT-R	–	–	0.342
+ SRG	✓	–	0.289
+ SAG	–	✓	0.295
GIST-R (full)	✓	✓	0.251

Table 3: Performance on clean, unrotated in-domain test samples.

Method	NYU↓	ADE↑	Hamlyn↓	SUIM↑
Painter	0.12	49.9	0.63	42.8
VICT	0.13	49.9	0.62	42.9
GIST	0.12	49.5	0.62	42.8

not observed for every corruption category. These experiments test controlled orientation mismatch on real-domain appearance and task distributions; they are not a measurement of native camera roll.

4.3 Ablation Study

Component contribution. Table 2 shows that SRG and SAG each improve the rotated baseline, and their combination performs best. This supports using both intervention points: SRG guides token interactions, whereas SAG changes reconstruction weighting. The comparison does not by itself establish a particular interaction mechanism.

Effect on clean in-domain samples. We run each adaptation method on the original clean test images with neither corruption nor prompt rotation. This control tests whether adaptation harms source-domain performance when no distribution or orientation shift is present. Table 3 shows small aggregate changes: relative to VICT, NYU error decreases from 0.13 to 0.12, Hamlyn is unchanged, and ADE/SUIM mIoU decreases by $0.4/0.1$ points. Thus the rotated-condition gains coexist with modest, dataset-dependent changes on clean inputs.

4.4 ATTENTION-BIAS DESIGN

Question and controls. The component ablation shows that SRG helps, but not whether spectral structure or merely an extra logit term causes the gain. We therefore fix the VICT objective, samples, rotations, and injection layers, disable SAG, and vary only the pre-softmax bias. *None* is VICT-R without a bias; *Raw feature affinity* directly uses pairwise similarity between the same detached tokens, bypassing eigendecomposition; *Random matched* uses an unstructured matrix rescaled to SRG's magnitude; and *SRG* uses the low-frequency eigenspace distances in Eq. (5). These controls respectively test raw locality, generic logit perturbation, and the proposed graph-relative construction.

Table 4: Attention-bias controls under random prompt rotation (zero-shot).

Bias	**NYU↓**	**ADE↑**
None	0.375	27.5
Raw feature affinity	0.392	26.7
Random matched	0.393	26.4
SRG	**0.330**	**28.4**

Result and interpretation. In Table 4, raw affinity and random matched both worsen NYUv2-C error and ADE20K-C mIoU relative to no bias. SRG instead reduces Abs.Rel from 0.375 to 0.330 and raises mIoU from 27.5 to 28.4. Because only the bias construction changes, the result argues against two simple explanations: that any feature similarity suffices or that arbitrary logit perturbation is beneficial. Together with Table 2, it supports low-frequency graph-relative structure as the effective forward control. This is a diagnostic, not proof of rotation invariance; SRG remains sensitive to graph construction and resampling, as Section 4.5 and the appendix examine.

4.5 FURTHER ANALYSIS

Rotation beyond padding. The full-angle control in Appendix E.4 shows a smaller performance decline from padding alone. A complementary feature diagnostic uses one brightness-corrupted NYUv2-C prompt and 256 sampled angles (Appendix E.10). After geometric patch matching, mean cosine similarity is 0.7666 under content rotation versus 0.9901 under matched padding; nearest-neighbor edge retention is 0.1499 versus 0.8293. This shows graph changes beyond boundary fill in the examined prompt, while interpolation, aliasing, and within-patch changes remain intertwined. The broader initialization diagnostic also finds subspace affinity decreasing from 0.785 at 5° to 0.524 at 120° (Table 11).

What spectral weighting measures. Appendix E.11 constructs low- and high-frequency graph residuals with identical MSE. At diagnostic $w_{\max} = 0.5$, their mean per-node spectral additions are 3.5649 and approximately zero. This illustrates SAG's frequency-selective weighting, not an empirical claim that rotation errors must be low-frequency. The probe is static; adaptation instead uses the preceding residual to set w_t. Together, these diagnostics explain why graph relations are used as soft guidance and how equal pixel errors can receive different update weights; downstream effectiveness is assessed by the component and attention-bias ablations.

5 CONCLUSION

Prompt rotation is a common mismatch for visual in-context adaptation because it preserves task semantics while changing token coordinates and resampling the prompt. GIST addresses this mismatch by first retaining VICT's support-reconstruction cycle, then adding SRG's graph-relative attention and SAG's adaptive reconstruction metric. Controlled studies show a smaller effect from padding alone, useful low-frequency regional correspondence in the examined cases, and selective spectral weighting. Across the evaluated compound shifts, the resulting adaptation improves average robustness without adding trainable modules or requiring rotation-specific backbone pretraining. Benefits vary across corruptions, and the graph computations increase runtime; the method supplies soft structural guidance rather than exact rotation equivariance.

AI USE STATEMENT

In this work, the authors used generative AI tools only to aid or polish the writing. All AI-assisted text was reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, pp. 2990–2999. PMLR, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations (ICLR)*, 2017.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A. Efros. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 29374–29385, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Yinan Huang, William Lu, Joshua Robinson, Yu Yang, Muhan Zhang, Stefanie Jegelka, and Pan Li. On the stability of expressive positional encodings for graphs. In *International Conference on Learning Representations (ICLR)*, 2024.

Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 1769–1776. IEEE, 2020.

Tomáš Karella, Adam Harmanec, Jan Kotera, Jan Blažek, and Filip Šroubek. Harmformer: Harmonic networks meet transformers for continuous roto-translation equivariance. *arXiv preprint arXiv:2411.03794*, 2024.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Soumyabrata Kundu and Risi Kondor. Steerable transformers for volumetric data. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, pp. 32011–32031. PMLR, 2025.

Derek Lim, Joshua Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. Sign and basis invariant networks for spectral graph representation learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 746–760. Springer, 2012.

U. Singhal, R. Feng, S. X. Yu, and A. Prakash. Test-time canonicalization by foundation models for robust perception. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 55788–55809, 2025.

Danail Stoyanov, Marco Visentini Scarzanella, Philip Pratt, and Guang-Zhong Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 275–282. Springer, 2010.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Haorui Wang, Haoteng Yin, Muhan Zhang, and Pan Li. Equivariant and stable positional encoding for more powerful graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6830–6839, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023b.

X. Wei, Q. Yang, Y. Fang, M. Zhu, and N. Wang. 3d test-time adaptation via graph spectral driven point shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26762–26771, 2025.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5028–5037, 2017.

Jiahao Xie, Alessio Tonioni, Nathalie Rauschmayr, Federico Tombari, and Bernt Schiele. Test-time visual in-context tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19996–20005, 2025.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Appendix for

When Visual Prompts Rotate: Spectral Guidance for Test-Time In-Context Adaptation

This supplementary material provides detailed experimental protocols, implementation clarifications, theoretical analyses, additional control experiments, and qualitative evidence for rotation-robust in-context adaptation with GIST.

Contents

A Additional Experimental Details **13**

A.1 Standard Benchmark Evaluation 13

A.2 Implementation Details 13

A.3 Environment and Reproducibility 14

B Additional Real-world Evaluation Details **14**

B.1 Dataset Description 14

B.2 Domain-specific Finetuning 14

C GIST Implementation Details **15**

C.1 Painter Streams, Feature Source, and Shapes 15

C.2 Fixed Support Graph and Adaptive Loss 16

C.3 Complete Adaptation Procedure 17

D Theoretical Properties of the Spectral Construction **17**

D.1 Idealized Node Relabeling and Eigenspace Ambiguities 18

D.2 Conditional Perturbation Bound after Patch Alignment 19

D.3 Properties of the Support Spectral Metric 19

E Additional Control Experiments **20**

E.1 Adaptive versus Fixed Spectral Weight 20

E.2 Injection-Layer Selection 21

E.3 Graph Connectivity and Near-zero Modes 21

E.4 Padding-only and Full-angle Rotation Control 22

E.5 Rotation-conditioned Spectral Diagnostics 23

E.6 Simple Geometric Controls 24

E.7 Computational Cost 26

E.8 Consistency across Rotation Quadrants 26

E.9 Sensitivity across Corruption Severities 26

E.10 Prompt-feature Drift beyond Padding 27

E.11 Equal-MSE Spectral Residual Diagnostic 27

F Additional Visualization **28**

F.1 Additional Layerwise Spectral Responses 28

F.2 Query-to-Prompt Attention Correspondence 30

A Additional Experimental Details

A.1 Standard Benchmark Evaluation

Datasets. We evaluate our method on two corruption benchmarks using identical settings for all compared methods.

For depth estimation, we use NYUv2-C (Silberman et al., 2012). The evaluation metric is Absolute Relative Error (AbsRel).

For semantic segmentation, we use ADE20K-C (Zhou et al., 2017). The evaluation metric is mean Intersection-over-Union (mIoU).

Corruption Types. Following common corruption benchmarks, we evaluate 15 corruption types covering four categories:

- Noise corruptions: Gaussian noise, impulse noise, and shot noise.
- Blur corruptions: Defocus blur, glass blur, motion blur, and zoom blur.
- Weather corruptions: Fog, frost, and snow.
- Digital corruptions: Brightness, contrast, elastic transformation, JPEG compression, and pixelation.

Unless otherwise specified, all corruption experiments use severity level 5.

Prompt Settings. We evaluate two visual prompt settings.

For each dataset and prompt setting, we select one underlying prompt pair once before evaluation and keep it fixed for every query and every compared method. The zero-shot prompt is drawn from the clean training distribution. In the one-shot setting, the fixed underlying prompt receives the corresponding corruption, with the same transformed prompt supplied to all methods. This paired protocol prevents prompt selection from confounding method comparisons.

Rotation Corruption Generation. To simulate orientation mismatch between visual prompts and query observations, we use the long-tailed angle distribution reported in the main paper. Rotation direction is sampled symmetrically between clockwise and counter-clockwise directions. The same sampled transformation is used by all compared methods for a given test instance.

A.2 Implementation Details

Standard-benchmark experiments initialize a sample-specific copy from the pretrained Painter ViT-Large checkpoint (Wang et al., 2023a). SUIM and Hamlyn instead use the shared dataset-specific checkpoints described below. Each test sample is optimized independently. Test-time adaptation is performed using AdamW:

- optimizer: AdamW
- learning rate: 1×10^{-6}
- $\beta_1 = 0.9$
- $\beta_2 = 0.999$
- weight decay: 0
- batch size: 1
- adaptation steps: 60

Each iteration uses Painter's native 2×2 canvas, where each sub-image is resized to 448×448. During TTT, only the encoder parameters of this copy are updated; the decoder remains fixed. The query prediction $\hat{Y}_q$ is not detached before the reverse completion, so reconstruction gradients pass through both Painter calls. Graph feature extraction, eigendecomposition, the attention bias B, and the statistic used to compute w_t are detached from this gradient path.

All experiments use graph degree $k_g = 16$, the eight ordered modes $u_2, \ldots, u_9$ after the first zero-eigenvalue mode, support resolvent $R_s = (L_s + \alpha I)^{-1}$ with $\alpha = 0.1$, and $w_{\max} = 1$. The affinity bandwidth σ_{aff} is not a fixed hyperparameter; it is the mean non-self pairwise feature distance of the current graph input. SRG and SAG use the same graph-construction rule and mode indices but different graph sources: spectral relation guidance uses the current attention stream, whereas the loss uses the fixed support-image graph. For every corruption setting, GIST and VICT under rotation use the same optimizer, prompt, rotation draws, and 60-step adaptation budget. The suffix -R in result tables marks this rotated-data condition for either method.

A.3 Environment and Reproducibility

The following table shows the environment for the reported spectral diagnostics and test-time adaptation controls.

Table 5: Hardware and software environment.

Component	**Configuration**
Host	I29c3e5fc530010124b
OS / kernel	Linux 5.15.0-174-generic, x86_64
CPU	AMD EPYC 7R12, 48 physical / 96 logical cores
RAM	251 GiB, no swap
GPU	2 × NVIDIA GeForce RTX 4090
GPU memory	49,140 MiB per GPU (approximately 48 GiB)
Driver / CUDA	570.211.01 / 12.8
Compute capability	8.9
Python	3.12.9 (Anaconda)
PyTorch / Torchvision	2.11.0+cu128 / 0.26.0+cu128
CUDA / cuDNN	12.8 / 9.24.0 (92400)
timm	0.3.2
NumPy / Pillow	2.4.6 / 12.3.0

B Additional Real-world Evaluation Details

B.1 Dataset Description

SUIM. SUIM (Islam et al., 2020) is an underwater semantic segmentation dataset containing diverse underwater scenes and object categories.

The dataset contains substantial appearance variations caused by underwater imaging conditions, including color distortion, illumination attenuation, and low contrast.

We use SUIM to evaluate whether our method preserves semantic structures under realistic domain variations.

Hamlyn. Hamlyn (Stoyanov et al., 2010) is an endoscopic depth estimation dataset collected from surgical environments. Due to camera manipulation and viewpoint changes during acquisition, the dataset contains geometric variations that challenge depth prediction consistency. We use Hamlyn to evaluate structural robustness for dense geometric prediction.

B.2 Domain-specific Finetuning

Since both datasets are outside the original Painter training distribution, we use a fixed 7:2:1 train/validation/test split of each official dataset release. We first perform supervised dataset-specific finetuning on the training subset before test-time adaptation. This offline stage updates all Painter parameters and uses neither test images nor test labels. SUIM and Hamlyn use separate dataset-specific checkpoints because they have different task outputs; within each dataset, all compared methods start from exactly the same checkpoint, so the only difference is the test-time adaptation strategy.

Table 6: Checkpoint and tokenization.

Field	Value
Checkpoint	`models/painter_vit_large/painter_vit_large.pth`
SHA-256	1bdd60ce013e82253e2a031726f36d07 587947438b0e561d4338d6f0c1113f44
Model	`painter_vit_large_patch16_input896x448_win_dec64_8glb_sl1`
Input canvas	896×448
Patch size	16×16
Attention-graph tokens	1,568 (56×28)
Support-loss tokens	784 (28×28)

Optimization Details. During dataset-specific finetuning, we update Painter's full encoder and decoder. Optimization uses AdamW with:

- learning rate: 2×10^{-5}
- layer-wise learning rate decay: 0.8
- weight decay: 0.05
- cosine learning rate scheduling
- training epochs: 10

SUIM Training Objective. For underwater segmentation, we optimize:

$$\mathcal{L}_{SUIM} = \mathcal{L}_{focal} + \mathcal{L}_{Dice} + \mathcal{L}_{RGB}.$$

The focal loss uses inverse-frequency class weighting to alleviate category imbalance. $\mathcal{L}_{RGB}$ is Painter's pixel reconstruction loss on the RGB-encoded segmentation target; it belongs only to the shared offline finetuning stage and is not an additional test-time regularizer.

Hamlyn Training Objective. For depth estimation, we use a combination of:

- relative depth error loss,
- multi-scale gradient matching loss,
- structural consistency between predicted and training-target depth maps.

After domain-specific finetuning, our method applies the test-time adaptation procedure described in the main paper.

C GIST Implementation Details

C.1 Painter Streams, Feature Source, and Shapes

Painter presents a conceptual 2×2 image–annotation layout but implements it as two aligned vertical streams: $S_X = [X_s; X_q]$ and $S_Y = [Y_s; M]$ in the forward completion. Each stream forms an 896×448 canvas and is patchified separately with 16×16 patches, yielding $N = 56 \times 28 = 1{,}568$ tokens. The streams are processed independently through blocks 0–2 and averaged tokenwise after block 2. Each attention operation and its bias are therefore $1{,}568 \times 1{,}568$ rather than one joint matrix over both streams.

For completion $r \in \{\mathrm{fwd}, \mathrm{rev}\}$ and injection layer $\ell \in \{0, 16\}$, let $H^{(r,\ell,s)} \in \mathbb{R}^{N_\ell \times C_\ell}$ be the input to the current attention stream s. At layer 0, s indexes the separate image and annotation streams; at layer 16, it denotes their fused representation. We construct

$$A^{(r,\ell,s)}, L^{(r,\ell,s)}, B^{(r,\ell,s)} \in \mathbb{R}^{N_\ell \times N_\ell}.$$

Algorithm 1 Causal layerwise spectral relation guidance in one Painter completion

Require: Image and annotation streams; L_{blocks} blocks $\{F_\ell\}$; injection layers $\mathcal{J} = \{0, 16\}$; $k_g = 16$; $m = 8$

Patchify both streams and add segment and positional embeddings
for $\ell = 0, \ldots, L_{\text{blocks}} - 1$ **do**
  if $\ell \in \mathcal{J}$ **then**
    for all current attention streams s **do**
      $Z \leftarrow \text{stopgrad}(H_\ell^{(s)})$; flatten Z to $N_\ell \times C_\ell$
      $d_{ij} \leftarrow \|Z_i - Z_j\|_2$; $\sigma_{\text{aff}} \leftarrow \text{mean}_{i \neq j}\, d_{ij}$
      $A_{ij} \leftarrow \exp[-d_{ij}^2/(2\sigma_{\text{aff}}^2)]$; mask A_{ii} before top-k_g
      Retain k_g non-self neighbors; $A \leftarrow \frac{1}{2}(A + A^\top)$; set $A_{ii} = 0$
      $L \leftarrow I - D^{-1/2} A D^{-1/2}$; $[\Lambda, U] \leftarrow \text{eigh}(L)$
      $E \leftarrow [u_2, \ldots, u_{m+1}]$
      $B \leftarrow \text{RowStd}(-\text{PairDist}^2(E))$
      $H_{\ell+1}^{(s)} \leftarrow F_\ell(H_\ell^{(s)}; B)$, adding B before softmax
    end for
  else
    Apply the ordinary Painter block F_ℓ
  end if
  if $\ell = 2$ **then**
    Fuse the aligned image and annotation streams tokenwise
  end if
end for
return completion features for Painter decoding

The bias is generated inside the same attention call from its input $H^{(r,\ell,s)}$ before softmax. Consequently, layer 0 uses patch, segment, and positional embeddings already available at its entrance; layer 16 uses the hidden representation entering block 16. No auxiliary forward pass or future-layer feature is required. $B^{(r,\ell,s)}$ is broadcast across heads and added to the same-shaped logits before the ordinary relative-position term and softmax.

For every current stream, σ_{aff} is its mean non-self pairwise feature distance. Self-affinity is masked before top-k_g selection, ensuring 16 non-self candidates, and the directed KNN graph is symmetrized as $A \leftarrow \frac{1}{2}(A + A^\top)$. We form the normalized Laplacian and use dense FP32 `torch.linalg.eigh`; no iterative sparse solver or explicit connected-component repair is applied. Following the implementation, we always discard u_1 and use the next ordered eigenvectors $U_m = [u_2, \ldots, u_9]$. If a graph is disconnected, this fixed-index policy can retain an additional zero mode; the distance construction remains defined, but its interpretation as a consistent correspondence cue is conditional on a suitable gap around the retained subspace. The negative pairwise squared distances in U_m are standardized independently along each row. Feature extraction for the graph, eigendecomposition, and $B^{(r,\ell,s)}$ are detached from optimization.

Under an ideal artificial node reindexing, $L' = PLP^\top$. A real prompt rotation is different: even after node alignment, interpolation, padding, changed KNN edges, and feature variation yield $L' \approx PLP^\top + E_L$. We therefore treat the spectrum as a graph-relative prior whose reliability depends on the perturbation and spectral gap, without assuming a fixed transformation law across physical rotations. This formulation follows Painter's actual stream and fusion schedule rather than imposing a new recognition pattern.

C.2 Fixed Support Graph and Adaptive Loss

The loss graph is distinct from the completion-specific spectral relation guidance graphs. At the first spectral-loss evaluation of an adaptation episode, we extract the patch embedding $Z_s = \text{PatchEmbed}(X_s)$ from the support image alone. We mask self-affinity before selecting $k_g = 16$ neighbors, symmetrize with $A_s \leftarrow \frac{1}{2}(A_s + A_s^\top)$, and form

$$L_s = I - D_s^{-1/2} A_s D_s^{-1/2} = U_s \Lambda_s U_s^\top.$$

The detached L_s and its eigendecomposition are cached and remain fixed for all 60 adaptation steps. For a 448×448 support image and 16×16 patches, this graph has $N_s = 28 \times 28 = 784$ nodes. Let $\mathcal{P}_p$ split a three-channel annotation image into this same non-overlapping patch grid and flatten each

Algorithm 2 Per-sample GIST adaptation

Require: Episode (X_s, Y_s, X_q); pretrained θ; steps $T = 60$
1: $\theta_q \leftarrow \theta$; build detached L_s from $\mathrm{PatchEmbed}(X_s)$
2: $R_s \leftarrow (L_s + \alpha I)^{-1}$; cache (L_s, U_s, R_s)
3: **for** $t = 0, \ldots, T-1$ **do**
4: $\quad \hat{Y}_q \leftarrow P_{\theta_q}((X_s, Y_s), X_q)$ with same-call spectral relation guidance
5: $\quad \hat{Y}_s \leftarrow P_{\theta_q}((X_q, \hat{Y}_q), X_s)$ with newly recomputed spectral relation guidance
6: $\quad \Delta_t \leftarrow \mathcal{P}_p(Y_s - \hat{Y}_s)$
7: $\quad$ Set $w_0 = 0$; otherwise compute w_t from detached Δ_{t-1} by the main-paper concentration rule
8: $\quad$ Compute $\mathcal{L}_{\mathrm{SAG}} = \mathrm{tr}[\Delta_t^\top (I + w_t R_s)\Delta_t]$
9: $\quad$ Update only the encoder of θ_q; keep the decoder fixed
10: **end for**
11: **return** $P_{\theta_q}((X_s, Y_s), X_q)$ with recomputed SRG

patch. The operator consumes the encoded support target delivered by Painter's task preprocessing. GIST adds no task-specific validity mask, spatial or channel averaging, or normalization by valid-pixel count. The reconstructed support residual is

$$\Delta_t = \mathcal{P}_p(Y_s - \hat{Y}_s) \in \mathbb{R}^{N_s \times D_p}, \qquad D_p = 3p^2.$$

Rows correspond one-to-one with support-image graph nodes, while columns preserve all within-patch pixel and channel components. Thus the loss is the sum over the complete encoded support cell; no query or blank-canvas region is introduced into the loss graph.

Because L_s contains a zero mode, directly inverting it is ill-posed. We therefore define the regularized support resolvent

$$R_s = (L_s + \alpha I)^{-1} = U_s(\Lambda_s + \alpha I)^{-1} U_s^\top.$$

The shift makes the inverse finite and caps its strongest low-frequency response. With $\mathcal{L}_{\mathrm{recon}} = \|\Delta_t\|_F^2$ and $\mathcal{L}_{\mathrm{spec}} = \mathrm{tr}(\Delta_t^\top R_s \Delta_t)$, the complete objective is

$$\mathcal{L}_{\mathrm{SAG}} = \mathcal{L}_{\mathrm{recon}} + w_t \mathcal{L}_{\mathrm{spec}} = \mathrm{tr}\left[\Delta_t^\top (I + w_t R_s)\Delta_t\right].$$

Equivalently, the trace sums D_p graph quadratic forms, one for each preserved residual component. It contains no separate smoothness regularizer. At $t = 0$, $w_0 = 0$, so the objective recovers the squared support-reconstruction loss, while SRG remains active. For $t \geq 1$, with $\bar{\Delta}_{t-1} = \mathrm{stopgrad}(\Delta_{t-1})$ and $\bar{\hat{\Delta}}_{t-1} = U_s^\top \bar{\Delta}_{t-1}$,

$$\rho_t = \frac{\sum_{k=2}^{m+1} \|\bar{\hat{\Delta}}_{t-1,k,:}\|_2^2}{\|\bar{\Delta}_{t-1}\|_F^2 + \varepsilon}, \qquad w_t = w_{\max}\rho_t,$$

where $m = 8$, $w_{\max} = 1$, and the first zero-eigenvalue mode is excluded. These are the same retained mode indices used by spectral relation guidance, but they belong to the independently constructed support graph L_s.

C.3 Complete Adaptation Procedure

The intermediate $\hat{Y}_q$ is not detached, so gradients from the reverse residual traverse both Painter completions. Only the spectral side computations are detached: spectral relation guidance graph construction and biases, cached (L_s, U_s, R_s), and the previous-step statistic used for w_t. No query label participates in graph construction, loss evaluation, model selection, or the final prediction.

D Theoretical Properties of the Spectral Construction

Scope: idealized correspondence. This section records algebraic properties of GIST's spectral bias and support-graph metric. The exact statements below assume an ideal node relabeling and, where stated, an orthogonal change of basis inside the same retained eigenspace. These premises are introduced before the corresponding propositions and should not be identified with physical image rotation, which also changes pixels, features, valid patches, and graph edges. For an aligned rotated graph, we give only a conditional perturbation bound. The treatment draws on analyses of

spectral ambiguity and perturbation (Lim et al., 2023; Huang et al., 2024); the statements below are derived for GIST's construction.

Let $A \in \mathbb{R}^{N\times N}$ be a symmetric nonnegative affinity matrix with strictly positive node degrees, $D = \mathrm{diag}(A\mathbf{1})$, and $L = I - D^{-1/2}AD^{-1/2}$. Let $U \in \mathbb{R}^{N\times m}$ contain the retained Laplacian modes and define

$$S(U)_{ij} = -\|U_{i,:} - U_{j,:}\|_2^2, \qquad B(U) = \mathrm{RowStd}(S(U)).$$

The row standardization is understood on nonconstant rows; the implementation may equivalently include a fixed numerical stabilizer in the row standard deviation.

D.1 Idealized Node Relabeling and Eigenspace Ambiguities

Proposition 1 (Normalized Laplacian under ideal node relabeling)**.** *For any permutation matrix P, if $A' = PAP^\top$, then*

$$D' = PDP^\top, \qquad L' = PLP^\top.$$

Consequently, if $Lu = \lambda u$, then $L'(Pu) = \lambda(Pu)$.

Proof. Because $P\mathbf{1} = \mathbf{1}$, the degree vector of A' is $A'\mathbf{1} = PA\mathbf{1}$, hence $D' = PDP^\top$. Functional calculus for a permuted diagonal matrix gives $(D')^{-1/2} = PD^{-1/2}P^\top$. Substitution yields

$$\begin{aligned} L' &= I - (PD^{-1/2}P^\top)(PAP^\top)(PD^{-1/2}P^\top) \\ &= P\big(I - D^{-1/2}AD^{-1/2}\big)P^\top. \end{aligned}$$

Multiplying $Lu = \lambda u$ by P proves the eigenvector statement.

Proposition 2 (SRG consistency under ideal node relabeling and basis change)**.** *Suppose two retained embeddings describe the same spectral subspace after node correspondence, so that*

$$U' = PUO$$

for a permutation matrix P and an orthogonal matrix $O \in O(m)$. Then

$$S(U') = PS(U)P^\top, \qquad B(U') = PB(U)P^\top.$$

This includes arbitrary eigenvector sign flips and orthogonal changes of basis inside repeated eigenspaces.

Proof. Let π be the node permutation represented by P. Orthogonality of O gives

$$\begin{aligned} \|U'_{i,:} - U'_{j,:}\|_2 &= \|(U_{\pi(i),:} - U_{\pi(j),:})O\|_2 \\ &= \|U_{\pi(i),:} - U_{\pi(j),:}\|_2. \end{aligned}$$

Thus every pairwise entry is permuted in both indices, proving the first identity. A simultaneous column permutation only reorders the entries within each row, so its row mean and standard deviation are unchanged; the row permutation merely reorders those statistics. Row standardization therefore commutes with $S \mapsto PSP^\top$, which proves the second identity. A sign flip is the special case in which O is diagonal with entries in $\{-1, 1\}$.

Corollary 1 (Attention consistency under ideal node relabeling)**.** *If $Q' = PQ$, $K' = PK$, and $V' = PV$, and any native positional logit term satisfies $C'_{\mathrm{pos}} = PC_{\mathrm{pos}}P^\top$, then attention with SRG satisfies*

$$\mathrm{Attn}(Q', K', V'; C'_{\mathrm{pos}}, B(U')) = P\,\mathrm{Attn}(Q, K, V; C_{\mathrm{pos}}, B(U)).$$

Proof. The content logits, native positional term, and spectral bias all transform by conjugation with P. Row-wise softmax commutes with simultaneous row and column permutation, and right multiplication by PV then yields the result.

The corollary isolates the algebra of spectral relation guidance under an ideal exact reindexing. It does not assert that a rotated image, Painter's learned positional terms, or its intermediate features satisfy the premise, and therefore makes no claim about a fixed response to physical rotation.

D.2 Conditional Perturbation Bound after Patch Alignment

Conditional setting. For analysis, assume two graphs of equal size admit a bijective node alignment P, and write

$$\bar{L} = P^\top L' P = L + E_L, \qquad \eta = \|E_L\|_2.$$

This assumption concerns full aligned graphs; restricting embeddings to valid patches in a diagnostic does not automatically satisfy it. The perturbation E_L collects interpolation, padding, changed KNN edges, and feature drift. Let U span the retained modes and let $\bar{U}$ be the corresponding retained subspace of $\bar{L}$. Define the boundary gap

$$\gamma_m = \min\{\lambda_2 - \lambda_1, \lambda_{m+2} - \lambda_{m+1}\},$$

with the natural one-sided modification if a boundary is absent.

Proposition 3 (Conditional bound on spectral-distance change)**.** *Assume $\gamma_m > 0$ and $\eta < \gamma_m/2$, so the retained subspace does not cross an eigenvalue cluster boundary. Then there exists $O_* \in O(m)$ such that*

$$\|\bar{U} - UO_*\|_F \le \frac{2\sqrt{2m}\,\eta}{\gamma_m}.$$

Moreover,

$$\max_{i,j} \left| \|\bar{U}_{i,:} - \bar{U}_{j,:}\|_2^2 - \|U_{i,:} - U_{j,:}\|_2^2 \right| \le \frac{16\sqrt{2m}\,\eta}{\gamma_m}.$$

If every row of $S(U)$ and $S(\bar{U})$ has standard deviation at least $\tau > 0$, then for a finite constant $C_{\tau,N}$ depending on τ, the row dimension, and the fixed numerical stabilizer,

$$\|B(\bar{U}) - B(U)\|_F \le C_{\tau,N} \frac{\eta}{\gamma_m}.$$

Proof. The population-gap form of the Davis–Kahan bound (Yu et al., 2015) gives $\|\sin\Theta(U,\bar{U})\|_F \le 2\sqrt{m}\,\eta/\gamma_m$. Orthogonal Procrustes alignment satisfies $\min_O \|\bar{U} - UO\|_F \le \sqrt{2}\|\sin\Theta(U,\bar{U})\|_F$, giving $\min_O \|\bar{U} - UO\|_F \le 2\sqrt{2m}\,\eta/\gamma_m$ under the stated separation condition. Pairwise distances are unchanged by O_*, so set $F = \bar{U} - UO_*$. Since rows of an orthonormal-column matrix have norm at most one,

$$\begin{aligned} &\left| \|\bar{U}_{i,:} - \bar{U}_{j,:}\|_2^2 - \|(U_{i,:} - U_{j,:})O_*\|_2^2 \right| \\ &\quad \le 4\big(\|F_{i,:}\|_2 + \|F_{j,:}\|_2\big) \le 8\|F\|_F, \end{aligned}$$

which gives the second bound. Finally, row standardization is Lipschitz on the set of rows whose standard deviation is bounded below by τ; applying this row by row to the distance bound yields the last inequality.

The ratio η/γ_m makes the scope explicit. For a disconnected graph or a repeated eigenvalue crossing the retained boundary, γ_m can vanish and the bound becomes vacuous. This agrees with the connectivity diagnostics in Section E: the result bounds an aligned perturbation under stated assumptions and does not guarantee consistent behavior for every transient graph.

D.3 Properties of the Support Spectral Metric

Proposition 4 (Adaptive spectral term under ideal node relabeling)**.** *Let $L'_s = PL_sP^\top$, $\Delta'_t = P\Delta_t$, and $\bar{\Delta}'_{t-1} = P\bar{\Delta}_{t-1}$. Then*

$$R'_s = (L'_s + \alpha I)^{-1} = PR_sP^\top.$$

Let $E_s = [u_{s,2}, \ldots, u_{s,m+1}]$ denote only the retained support modes, distinct from the full eigenbasis U_s. If $E'_s = PE_sO$ for an orthogonal $O \in O(m)$, then the concentration ratio, adaptive weight, and SAG loss are unchanged:

$$\rho'_t = \rho_t, \qquad w'_t = w_t, \qquad \mathcal{L}'_{\mathrm{SAG}} = \mathcal{L}_{\mathrm{SAG}}.$$

Proof. The resolvent identity follows from $P^\top P = I$. Furthermore,

$$(E'_s)^\top \bar{\Delta}'_{t-1} = O^\top E_s^\top \bar{\Delta}_{t-1},$$

so orthogonal invariance of the Frobenius norm preserves both the retained-mode energy and the total residual energy, and hence preserves ρ_t and w_t. Finally,

$$\begin{aligned}\mathcal{L}'_{\mathrm{SAG}} &= \mathrm{tr}\big[(P\Delta_t)^\top (I + w_t P R_s P^\top)(P\Delta_t)\big] \\ &= \mathrm{tr}\big[\Delta_t^\top (I + w_t R_s)\Delta_t\big]\,.\end{aligned}$$

Proposition 5 (Positive definiteness and controlled spectral emphasis)**.** *For $L_s \succeq 0$, $\alpha > 0$, and $0 \le w_t \le w_{\max}$, the metric $M_t = I + w_t R_s$ is positive definite and*

$$\|\Delta_t\|_F^2 \le \mathcal{L}_{\mathrm{SAG}} \le \left(1 + \frac{w_t}{\alpha}\right) \|\Delta_t\|_F^2.$$

Its gain on graph frequency $\lambda_{s,k}$ is $g_t(\lambda_{s,k}) = 1 + w_t/(\lambda_{s,k} + \alpha)$, which is nonincreasing in $\lambda_{s,k}$. In addition, the implemented concentration rule guarantees $0 \le w_t \le w_{\max}$, and $w_t = 0$ exactly recovers the pixel-reconstruction objective.

Proof. Diagonalizing L_s gives

$$M_t = U_s \,\mathrm{diag}\left(1 + \frac{w_t}{\lambda_{s,k} + \alpha}\right) U_s^\top.$$

Since every $\lambda_{s,k} \ge 0$, all eigenvalues of M_t lie in $[1, 1 + w_t/\alpha]$, proving positive definiteness and the quadratic-form bounds. Differentiating the gain gives $g'_t(\lambda) = -w_t/(\lambda + \alpha)^2 \le 0$. Finally, the numerator of ρ_t is the energy of an orthogonal projection onto a subset of modes, so it cannot exceed $\|\bar{\Delta}_{t-1}\|_F^2$; the positive denominator stabilizer therefore implies $0 \le \rho_t \le 1$. Thus $0 \le w_t = w_{\max}\rho_t \le w_{\max}$, and substitution of $w_t = 0$ leaves $M_t = I$.

E Additional Control Experiments

The following controls isolate the contributions of adaptive weighting, spectral bias, and geometric handling.

E.1 Adaptive versus Fixed Spectral Weight

Protocol. We compare pixel-only reconstruction, a validation-selected fixed weight, a predetermined monotone schedule, and GIST's adaptive w_t. All variants share checkpoints, prompts, rotation draws, learning rates, and the 60-step budget. Rotated scores use the full zero-shot corrupted benchmark over 15 corruption types, with standard deviations computed across three random seeds. Clean Δ is computed on clean zero-shot samples as the seed-averaged absolute difference from the matched pixel-only clean score.

Table 7: Ablation study on adaptive spectral weight. Rotated scores are mean±standard deviation across three seeds; Clean Δ reports seed-averaged absolute change from pixel-only adaptation.

Weighting	**NYU Rot.↓**	**ADE Rot.↑**	**NYU Clean Δ**	**ADE Clean Δ**
Pixel only ($w = 0$)	0.375 ± 0.003	27.7 ± 0.1	0.000	0.0
Best fixed w	0.353 ± 0.002	27.7 ± 0.1	0.015	0.3
Scheduled w	0.343 ± 0.002	27.9 ± 0.2	0.006	0.1
Adaptive w_t	0.330 ± 0.004	28.2 ± 0.2	0.021	0.3

Interpretation. Adaptive w_t gives the best rotated scores among the tested weighting rules. It does not minimize clean change: its NYU Clean Δ is 0.021, compared with 0.015 for fixed weighting and 0.006 for the schedule; on ADE, its 0.3 matches fixed weighting and exceeds the schedule's 0.1. Because Clean Δ is absolute, these values alone do not distinguish improvement from degradation. The table supports adaptive weighting for the rotated condition, without establishing automatic protection of clean performance or explaining when a schedule becomes harmful.

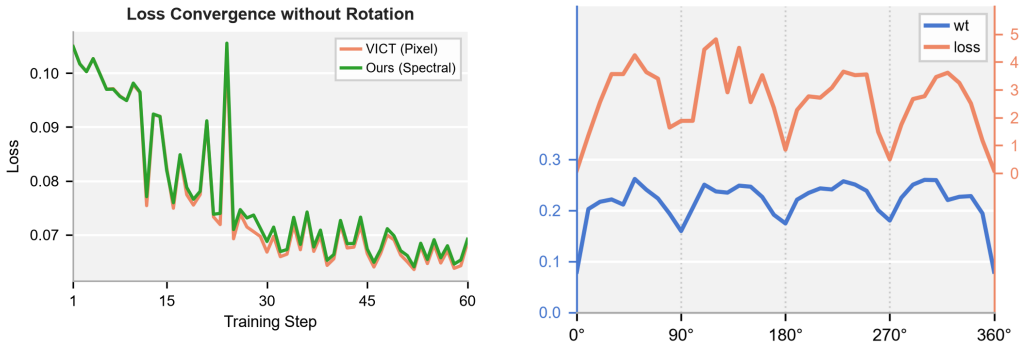


Figure E.1: Adaptive spectral weighting under clean and rotated prompts. Left: the displayed reconstruction-loss values for VICT and GIST follow similar trends on the clean 0° example. Right: the adaptive weight and reconstruction loss vary over a rotation sweep and both decline near realignment. Their different shapes illustrate distinct statistics; the definition of w_t, rather than this plot alone, establishes its dependence on residual spectral composition.

The similar scalar-loss curves in Figure E.1 do not establish identical parameter trajectories, especially because SAG and ordinary reconstruction use different metrics. The rotation plot is descriptive: the weight is computed from the preceding residual's spectral concentration, not from the angle or total loss. It is neither an angle estimator nor a graph-confidence score.

E.2 Injection-Layer Selection

We vary the spectral relation guidance injection point while keeping the adaptive support-graph loss and all other settings fixed.

Table 8: Injection-layer ablation on ADE20K-C under random rotation in the one-shot setting.

Injection layer	**mIoU** ↑
Layer 0	28.12
Layer 8	27.47
Layer 16	28.25
Layer 23	27.88
Layer 0+16	**28.45**

This behavior is consistent with Painter's staged input design. In the earliest blocks, the image stream and annotation stream are still processed separately, so a layer-0 bias controls low-level correspondences before the two modalities are mixed. After block 2 (zero-based indexing), Painter fuses the streams; a layer-16 bias therefore acts on an intermediate representation that has integrated image–annotation context but is not yet fully specialized by the decoder. Since these two locations regularize different stages of the same completion process, we also test their combination. Layer 0+16 gives the best result, showing that early pre-fusion control and middle fused-feature control provide complementary benefits.

E.3 Graph Connectivity and Near-zero Modes

Protocol. We diagnose graph health at $t = 0$ on the ADE20K-C stratified subset: one one-shot paired episode for each of 15 corruptions and eight signed rotations. The diagnostic separates two graph roles in GIST. The first role is the temporary SRG graph computed from Painter tokens inside each completion; these graphs affect attention logits only. The second role is the fixed support-loss graph built from the support image, which defines the spectral metric used by the SAG loss. The seven graph sources therefore comprise the image and annotation streams at layer 0 and the fused stream at layer 16 in both completions, plus the fixed support-loss graph. For the binary support of each averaged-symmetrized kNN graph, let $c(G)$ be its number of connected components. Over M

Table 9: ADE20K-C connectivity by graph source in the stratified one-shot, $t = 0$ diagnostic.

Graph source	Disc. rate	Components	Mean μ_0
Layer-0 image, forward	0.067	1.07	1.07
Layer-0 annotation, forward	1.000	2.01	2.01
Layer-16 fused, forward	0.904	1.90	1.90
Layer-0 image, reverse	0.067	1.07	1.07
Layer-0 annotation, reverse	1.000	2.00	2.00
Layer-16 fused, reverse	0.993	1.99	1.99
Fixed support-loss graph	0.000	1.00	1.00

graphs, the disconnection rate is

$$r_{\text{disc}} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}[c(G_j) > 1].$$

We also report component counts and the near-zero multiplicity $\mu_0 = \#\{i : \lambda_i \leq \tau_0\}$ with $\tau_0 = 10^{-5}$. Forward and reverse graphs remain separate records. These measurements diagnose the initial graph construction and do not represent graph health throughout 60-step adaptation.

Table 9 shows a clear separation between the two graph roles. The fixed support-loss graph remains connected for every original and rotated record across all 15 corruptions ($\mu_0 = 1$ and largest-component ratio 1.0), so the quadratic SAG metric is not dominated by extra zero modes in this diagnostic. By contrast, disconnections are concentrated in transient SRG graphs, especially annotation-stream and fused-stream graphs. This is expected from Painter's staged computation: early annotation tokens contain mask/label structure rather than only visual texture, and middle fused tokens mix support and query evidence before the decoder has settled the final prediction. These graphs are therefore better interpreted as temporary structural attention priors, not as the metric that defines the SAG loss.

Near-zero-mode sensitivity. The implementation retains the fixed modes $u_2, \ldots, u_9$, which skips the first trivial zero mode of a connected normalized Laplacian. If a graph is disconnected, however, additional near-zero eigenvectors may appear and could enter this fixed range. To test whether this choice drives the reported stability, we compare it with a zero-aware policy that skips every eigenvalue below τ_0 and then takes the next eight positive modes; no unreported eigengap-selected dimension or adaptive-k_g result is assumed.

Table 10: Retained-mode sensitivity on the ADE20K-C stratified $t = 0$ diagnostic. The last column is the mean per-pair absolute affinity change between policies.

Policy	Affinity↑	$\theta_{\max}$↓	Spectral Jaccard↑	$\lvert\Delta\rvert$ affinity
Fixed u_2:u_9	0.664	80.27	0.404	0.0378
Skip near-zero modes	0.679	78.02	0.404	0.0378

Skipping near-zero modes changes affinity only modestly and leaves spectral-neighbor Jaccard unchanged in this diagnostic. Thus the observed trends are not mainly explained by accidentally retaining extra zero modes. The result supports the fixed u_2:u_9 policy as a practical default for the current implementation, while identifying zero-aware selection as an alternative for disconnected transient graphs. We do not claim that the two policies are equivalent during full 60-step TTT; the comparison shows modest aggregate changes at initialization, not equivalence of individual graph embeddings.

E.4 Padding-only and Full-angle Rotation Control

Image rotation changes prompt coordinates while simultaneously introducing interpolation, aliasing, cropping, and blank boundaries. To separate boundary loss from the broader rotated-input effect, we evaluate ADE20K-C over a deterministic full-angle sweep and construct a matched padding-only input at every angle. The padding-only input receives the same valid-region mask and boundary fill but its interior content is not rotated. This control isolates the boundary component; it does not remove every resampling confound and therefore does not identify patch reindexing as the sole causal mechanism.

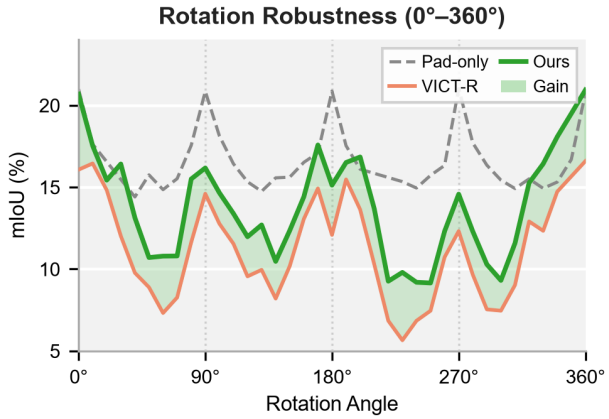


Figure E.2: Full-angle control on ADE20K-C. Matched padding alone causes limited degradation, whereas rotated prompts produce larger periodic drops as content is resampled and its patch-grid alignment changes. GIST attains higher mIoU than VICT-R over the displayed sweep.

Figure E.2 shows that padding is not harmless, but its effect is smaller than that of the complete rotation transform. VICT exhibits deeper periodic drops as the angle moves away from favorable grid alignments, while GIST follows the same broad periodicity at a higher performance level. Together with the matched construction, this supports the narrower conclusion that boundary fill alone does not explain the observed degradation. The remaining gap may combine changed patch correspondence, interpolation, aliasing, and content loss near the crop boundary.

E.5 Rotation-conditioned Spectral Diagnostics

Protocol. Eigenvectors cannot be compared columnwise because their signs and bases may change. We instead compare retained subspaces on paired original/rotated episodes at the same initial checkpoint before any parameter update, so the measured change is caused by prompt rotation rather than diverging adaptation trajectories. The known synthetic rotation defines a patch correspondence P_α: for the fixed support graph it maps valid support patches, while for a spectral relation guidance stream it is block diagonal, rotating the support half and leaving query nodes unchanged. After restricting to mutually valid one-to-one patch matches and orthonormalizing the restricted embeddings, we measure

$$S_{\mathrm{sub}}(\alpha) = \frac{1}{m} \left\| U_\alpha^\top P_\alpha U_0 \right\|_F^2 ,$$

the mean squared cosine of the principal angles, referred to below as subspace affinity. We additionally report the largest principal angle, the Spearman correlation between aligned pairwise spectral-distance matrices, and top-k_g spectral-neighbor Jaccard retention. These quantities are invariant to eigenvector sign and orthogonal basis changes. To prevent changing zero-mode multiplicity from being counted as rotation-induced subspace drift, this subsection uses the near-zero-aware policy only for the diagnostic; Section E.3 separately shows that switching from the implemented fixed indices changes the aggregate result only modestly. These are $t = 0$ diagnostics on the stratified subset, not full-test-set or post-adaptation estimates.

For the zero-mode-aware retained subspace, let c be the number of eigenvalues below τ_0. Define both boundary separations,

$$g_{\mathrm{low}} = \lambda_{c+1} - \lambda_c, \qquad g_{\mathrm{high}} = \lambda_{c+m+1} - \lambda_{c+m},$$

and use $g_{\min} = \min(g_{\mathrm{low}}, g_{\mathrm{high}})$ as the diagnostic boundary gap. Results are stratified using the main benchmark's angle bands, with 5° also reported separately because it motivates the small-rotation claim.

Table 11: Rotation-conditioned subspace diagnostics by absolute prompt angle on ADE20K-C. Values aggregate the seven graph-source/completion records under the near-zero-aware policy.

$\lvert\alpha\rvert$	**Affinity**↑	$\theta_{\max}$↓	**Spearman**↑
5°	0.785	61.24	0.798
20°	0.679	79.69	0.720
60°	0.629	85.78	0.683
120°	0.524	81.36	0.530

$\lvert\alpha\rvert$	**Spectral Jac.**↑	**Edge retain.**↑	**Raw Jac.**↑	**Eig. drift**↓
5°	0.458	0.603	0.652	0.208
20°	0.403	0.567	0.625	0.240
60°	0.379	0.546	0.606	0.273
120°	0.357	0.525	0.594	0.256

Interpretation. The angle-stratified results show graded degradation rather than a constant response across rotations. Affinity and all three neighborhood-retention measures decrease as rotation grows. Beyond 60°, the local-overlap measures change more slowly, but affinity falls further at 120°; this means that some local edges have reached a disruption plateau while the orientation of the retained subspace continues to change globally. The slightly smaller $\theta_{\max}$ and eigenvalue drift at 120° do not reverse this trend: $\theta_{\max}$ describes only the worst principal direction, eigenvalue drift does not determine eigenvector alignment near clustered eigenvalues, and affinity averages all retained directions. Thus Table 11 supports using a soft, perturbation-dependent prior rather than assuming a fixed spectral response across angles.

Connectivity and aligned positive-subspace correspondence also measure different properties. Annotation and fused SRG graphs often have two components because Painter tokens retain mask/label and support–query grouping, but removing the associated zero modes still leaves repeatable geometry within those components. Their relatively high affinity can therefore coexist with a high disconnection rate. This does not by itself validate every fixed-index SRG graph; Section E.3 tests that implementation choice, while the downstream injection ablation tests whether the resulting bias is useful.

The fixed support-loss graph shows the complementary case: it is connected, yet has lower cross-rotation correspondence. This is not a requirement violation. Each rotated episode builds L_s from the support image actually observed in that episode and then caches it for all 60 updates, so the SAG loss has one stationary, sample-conditioned metric. GIST does not require graphs built from two differently rotated versions of the support to be identical. Likewise, raw-feature neighbor Jaccard being higher than spectral-neighbor Jaccard measures repeatability alone, not the usefulness of either quantity as an attention bias; the attention-bias ablation provides that downstream comparison. Finally, the lower affinities under impulse and Gaussian noise show that reliability is corruption-dependent, which motivates using spectral structure as auxiliary guidance and adapting its loss weight rather than treating it as an exact correspondence rule.

E.6 Simple Geometric Controls

Protocol. We compare VICT under the rotated-prompt condition, rotate-and-average inference, label-free estimated canonicalization, oracle inverse rotation, and GIST. Let $\mathcal{R}_\alpha$ be the sampled transform applied jointly to the support image and annotation. The oracle is given the generator-recorded angle α, applies $\mathcal{R}_{-\alpha}$ to both rotated support cells, and then runs the same VICT adaptation pipeline while leaving the query unchanged. Estimated canonicalization follows the same inverse-correction principle but must infer the angle without labels or generator metadata. Because $\mathcal{R}_{-\alpha}\mathcal{R}_\alpha$ still contains interpolation, cropping, and padding artifacts, the oracle is a privileged, angle-informed reference rather than a strict performance upper bound.

To keep this control computationally tractable, we construct a fixed subset separately for NYUv2-C and ADE20K-C by selecting 200 query samples from each of the 15 severity-5 corruption types. All strategies are evaluated in the one-shot setting. Within every dataset–corruption pair, they use the same sample indices, one-shot prompt, and random prompt rotations. Only the oracle receives the generated angle; relative runtime accounts for additional inference. This subset control has a

Table 12: Rotation-conditioned subspace diagnostics by corruption on ADE20K-C under the near-zero-aware policy.

Corruption	Affinity	Spectral Jac.	Edge retain.	Raw Jac.
Brightness	0.743	0.447	0.609	0.653
Contrast	0.683	0.477	0.629	0.668
Defocus blur	0.793	0.477	0.600	0.656
Elastic transform	0.714	0.386	0.630	0.665
Fog	0.756	0.461	0.656	0.698
Frost	0.628	0.344	0.496	0.573
Gaussian noise	0.550	0.288	0.433	0.470
Glass blur	0.757	0.483	0.657	0.698
Impulse noise	0.440	0.253	0.365	0.413
JPEG compression	0.709	0.402	0.570	0.630
Motion blur	0.781	0.507	0.654	0.703
Pixelate	0.718	0.391	0.593	0.640
Shot noise	0.554	0.309	0.429	0.594
Snow	0.612	0.368	0.518	0.580
Zoom blur	0.753	0.468	0.641	0.688

Table 13: ADE20K-C subspace correspondence and connectivity by graph source. Neighbor retention denotes spectral-neighbor Jaccard.

Graph source	Disc. rate	Mean μ_0	Affinity↑	$\theta_{\max}$↓	Neighbor retain.↑
Layer-0 image, forward	0.067	1.07	0.649	74.39	0.379
Layer-0 annotation, forward	1.000	2.01	0.683	87.16	0.504
Layer-16 fused, forward	0.904	1.90	0.760	79.22	0.483
Layer-0 image, reverse	0.067	1.07	0.649	74.39	0.379
Layer-0 annotation, reverse	1.000	2.00	0.825	67.64	0.519
Layer-16 fused, reverse	0.993	1.99	0.756	78.80	0.427
Fixed support-loss graph	0.000	1.00	0.434	84.52	0.137

Table 14: Geometric handling strategies on the fixed 200-sample-per-corruption subsets under one-shot rotated prompts. The suffix -R denotes the rotated-data condition; oracle inverse rotation uses the generator-recorded angle and is not deployable.

Method	NYU Rot.↓	ADE Rot.↑	Time ×
VICT-R	0.395 ± 0.005	32.3 ± 0.3	1.00
Rotate-and-average	0.369 ± 0.006	32.6 ± 0.2	4.60
Estimated canonicalization	0.358 ± 0.004	32.7 ± 0.5	1.95
Oracle inverse rotation	0.302 ± 0.003	33.9 ± 0.3	1.27
GIST-R	0.330 ± 0.003	33.2 ± 0.2	1.40

different evaluation population from the full benchmark in the main paper, so its absolute values should not be compared directly with the main result table.

Interpretation. Among methods that do not receive the generated angle, GIST gives the strongest result on both datasets in this controlled subset and requires substantially less computation than rotate-and-average. Oracle inverse rotation performs better by directly removing the known synthetic transformation, but it answers a different, non-deployable question. Its result quantifies the benefit available from exact angle metadata and is not used to claim that GIST reproduces oracle correction.

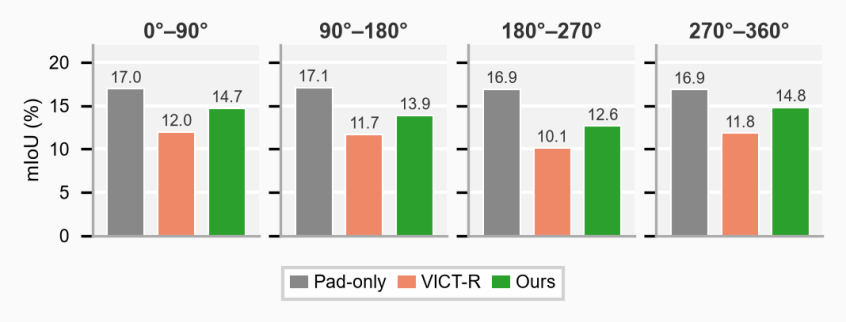


Figure E.3: ADE20K-C mIoU over four 90° intervals in the deterministic full-angle sweep. "Ours" denotes GIST-R.

E.7 Computational Cost

Table 15: Relative computational cost under rotated prompts with 60 adaptation steps per sample. Absolute runtime varies by dataset and corruption type.

Method	Relative time	Peak memory
VICT	1.00×	∼22 GB
GIST	1.40×	∼23 GB

Absolute runtime varies with dataset size, corruption type, storage throughput, and preprocessing. We therefore report only relative runtime in Table 15 and give one concrete example for scale. On a workstation with two 48-GB RTX 4090 GPUs, a complete NYUv2-C brightness evaluation takes approximately four hours for VICT under rotated prompts when every sample is adapted independently for 60 steps. Under the same checkpoint, samples, batch size, and step budget, GIST takes about 1.4× as long, or approximately 5.6 hours. Other datasets and corruption types have different absolute times, but exhibit the same relative trend. Peak memory for one running task increases from about 22 GB to 23 GB because graph construction and dense eigendecomposition add detached spectral tensors; other non-spectral adaptation baselines have approximately the same 22-GB footprint as VICT.

E.8 Consistency across Rotation Quadrants

We aggregate the full-angle ADE20K-C sweep used for the padding-only analysis in Figure E.2 into four 90° quadrants. These statistics summarize that deterministic angle sweep and are separate from the long-tailed random rotations used in the main benchmark tables.

Padding alone is less damaging than prompt rotation, while GIST improves over VICT under the same rotated condition in every quadrant. Thus boundary fill alone does not account for the observed degradation, and the gains occur in every quadrant; interpolation, aliasing, and correspondence changes remain intertwined.

E.9 Sensitivity across Corruption Severities

Protocol. We repeat the NYUv2-C evaluation at corruption severities 1–5. The checkpoint, query set, fixed underlying prompt, zero-/one-shot definitions, and random prompt-rotation distribution are identical to those in the main experiment; only corruption severity changes. Each result is the mean over the same 15 corruption types, so the table describes the reference gap across severities.

VICT-R has higher AbsRel than aligned VICT at every severity in both prompt settings, so the observed gap is not specific to severity 5. The zero-shot gap varies with corruption content and is not monotonic in severity. In contrast, the one-shot gap grows consistently as corruption becomes stronger and is largest at severity 5. This comparison measures the effect of the full prompt-rotation transform on VICT; it does not separate coordinate changes from interpolation, aliasing, and padding.

Table 16: Effect of prompt rotation across NYUv2-C corruption severities (AbsRel ↓), averaged over 15 corruption types. VICT uses aligned prompts, VICT-R uses randomly rotated prompts, and $\Delta_R = \text{VICT-R} - \text{VICT}$; a positive Δ_R denotes higher error under the full rotation transform.

	Zero-shot			One-shot		
Severity	**VICT**	**VICT-R**	Δ_R	**VICT**	**VICT-R**	Δ_R
1	0.127	0.133	0.006	0.094	0.103	0.009
2	0.208	0.227	0.019	0.138	0.152	0.014
3	0.253	0.273	0.020	0.164	0.194	0.030
4	0.310	0.341	0.031	0.215	0.253	0.038
5	0.365	0.375	0.010	0.205	0.342	0.137

E.10 Prompt-feature Drift beyond Padding

Protocol. We examine the severity-5 brightness-corrupted NYUv2-C prompt `study_room_0005b/rgb_00094.jpg` using the pretrained Painter patch embedding. We sample 256 signed angles from the main benchmark's long-tailed rotation distribution: 70% in $[0°, 10°)$, 25% in $[10°, 30°)$, 4% in $[30°, 90°)$, and 1% in $[90°, 180°]$, with symmetric signs. For each angle, the rotated-content view uses bicubic interpolation and RGB fill $(123, 116, 104)$. Its paired padding-only control applies the same fill mask while retaining the unrotated interior content. We extract 28×28 patch features and use geometric patch correspondences to compare the rotated view with the reference. We measure matched-patch cosine similarity and $k_g = 16$ nearest-neighbor edge retention. This diagnostic examines the prompt-image feature graph before adaptation, not all SRG streams or downstream prediction scores.

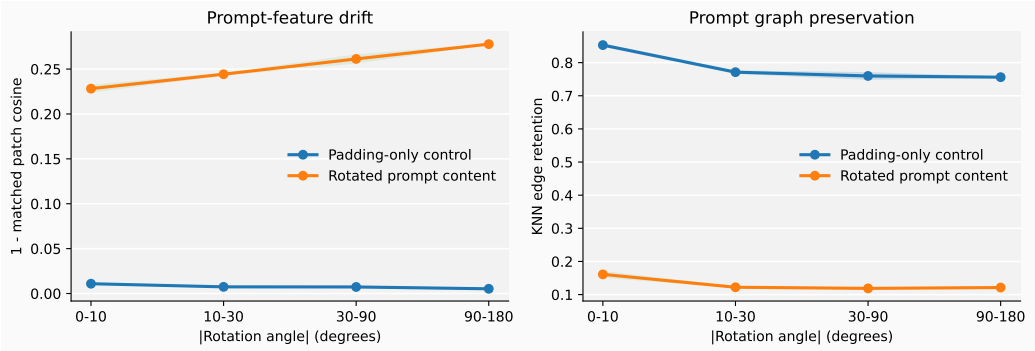


Figure E.4: Prompt-feature and graph changes for one brightness-corrupted NYUv2-C prompt over 256 sampled angles. Left: matched-patch cosine drift (1−cosine similarity). Right: nearest-neighbor edge retention. Curves show means within absolute-angle bins; shading denotes ±one standard error over sampled angles within each bin, not uncertainty across scenes.

Results and scope. Across the 256 draws, mean matched-patch cosine similarity is 0.7666 under content rotation and 0.9901 under padding alone, corresponding to drift 0.2334 and 0.0099. Mean edge retention is 0.1499 and 0.8293, respectively. Figure E.4 shows the same separation in every angle bin. Thus matched boundary fill alone does not explain the representation and graph changes in this example, even after geometric patch matching. The remaining difference includes interior resampling, aliasing, and changes in within-patch content; this experiment does not isolate a purely positional cause. It supports treating graph relations as rotation-sensitive guidance, rather than assuming that geometric alignment preserves the feature graph. Generalization across scenes and downstream benefit are separate from this single-prompt diagnostic.

E.11 Equal-MSE Spectral Residual Diagnostic

Construction. Using the same corrupted prompt and $k_g = 16$ support graph, we construct two scalar residual fields from the first nontrivial and largest-eigenvalue modes of its symmetric normalized Laplacian. Each field $z \in \mathbb{R}^{N_s}$ is scaled to $\|z\|_2^2/N_s = 1$, with $N_s = 784$. This fixes the ordinary mean squared error while changing its graph-frequency content. The diagnostic evaluates

the retained low-mode concentration $\rho(z)$ and uses $w(z) = 0.5\rho(z)$ with $\alpha = 0.1$. Unlike the main benchmark's $w_{\max} = 1$, this illustration uses $w_{\max} = 0.5$ and reports mean per-node extra energy,

$$E_{\text{extra}}(z) = \frac{w(z)}{N_s} z^\top (L_s + 0.1I)^{-1} z.$$

For a pure eigenmode of eigenvalue λ, unit MSE gives $E_{\text{extra}}(z) = 0.5\rho(z)/(\lambda + 0.1)$. The plotted value is the weighted spectral addition, not the complete SAG loss. This static probe evaluates the weight and energy on the same constructed field; it does not reproduce an adaptation trajectory, where the weight comes from the preceding detached residual.

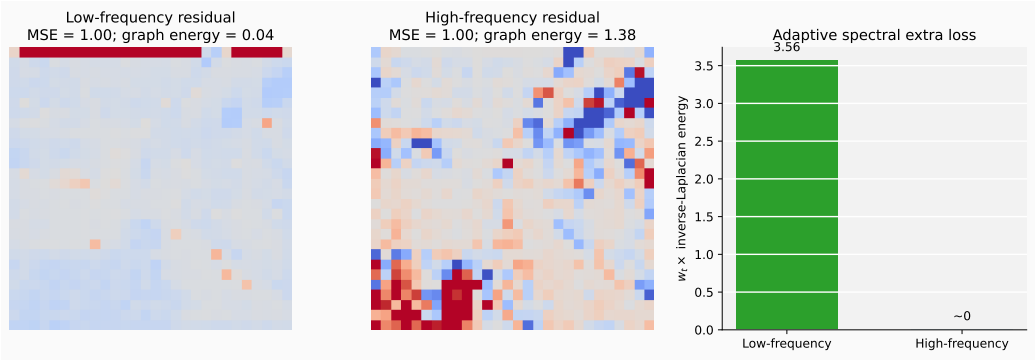


Figure E.5: Equal-MSE constructed residuals on a fixed support graph. Both fields have MSE 1.00; their normalized graph energies are 0.0403 and 1.3829. With diagnostic weight $w = 0.5\rho$, the mean per-node spectral additions are 3.5649 and approximately zero. Low/high frequency refers to the feature graph, not necessarily smoothness on the image grid.

Results and interpretation. The graph-low-frequency field has $\rho = 1.00$ and extra energy 3.5649, whereas the high-frequency field has $\rho = 2.3 \times 10^{-14}$ and extra energy 7.8×10^{-15}. Equal pixel error therefore need not produce equal spectral emphasis. The near-zero high-frequency addition here follows from its negligible adaptive weight; the resolvent alone has a positive response at every frequency. This construction illustrates the intended frequency selectivity of SAG, rather than establishing that actual rotation-induced residuals are predominantly low-frequency or that suppressing this constructed field improves task performance. Those distinctions keep the algebraic mechanism separate from the downstream evidence in the component ablations.

F Additional Visualization

F.1 Additional Layerwise Spectral Responses

We first examine a representative scene under its original orientation and two small prompt rotations. From a fixed Painter layer, we visualize a prompt-region crop of the low-frequency spectral partition together with the corresponding coordinate-conditioned attention response. Model parameters, layer selection, and visualization settings are held fixed across the displayed orientations.

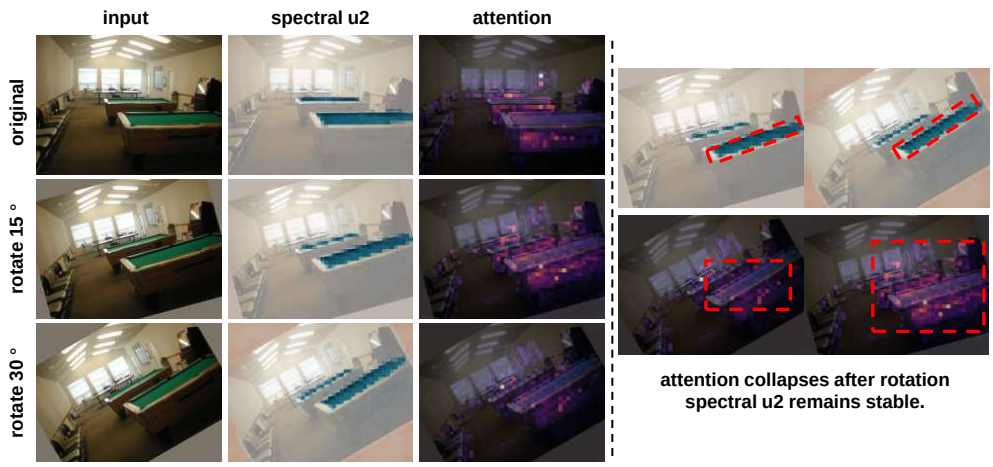


Figure F.1: Representative layerwise response under small prompt rotations. In this example, a low-frequency spectral partition remains associated with the same table region after grid displacement, while the corresponding attention response shifts and becomes less concentrated.

Figure F.1 illustrates the distinction motivating SRG: the displayed spectral grouping retains a recognizable regional association while coordinate attention changes with patch location. This is a qualitative case, not evidence of exact invariance. Eigenvector signs and bases are ambiguous, and real rotations alter features and graph edges through resampling and padding. The quantitative, basis-invariant comparisons in Table 11 provide the corresponding aggregate diagnostic.

Protocol. Figures F.2–F.4 extend Figure F.1's layerwise spectral-response example to three scenes with different structures: a dominant building, a road and skyline, and a water–vegetation landscape. For each scene, we keep the model, layer, and visualization procedure fixed and rotate the prompt by 15° and 30°. Here, attention denotes the final-layer weights from the query-side token to prompt patch tokens. Eigenvector sign and partition color indices are arbitrary: a color inversion in u_2 or a color permutation in the partition is not itself a structural change. We therefore compare region membership and boundaries relative to scene content rather than literal color values. These examples supplement, rather than replace, the quantitative $t = 0$ subspace diagnostics in Table 11.

Observation. The three examples show a recurring distinction between structural grouping and coordinate-conditioned attention. Although spectral coordinates can change sign, magnitude, or partition color, several major boundaries remain associated with the building facade, road and skyline, or pond and vegetation across the displayed rotations. Attention retains some semantic peaks but changes more visibly as the prompt moves to new patch indices. These cases are consistent with low-frequency graph structure providing an additional graph-relative cue when positional attention is perturbed. They are qualitative examples rather than evidence of a general rotation transformation law; interpolation, KNN-edge changes, eigenspace ambiguity, and scene-specific feature variation remain present.

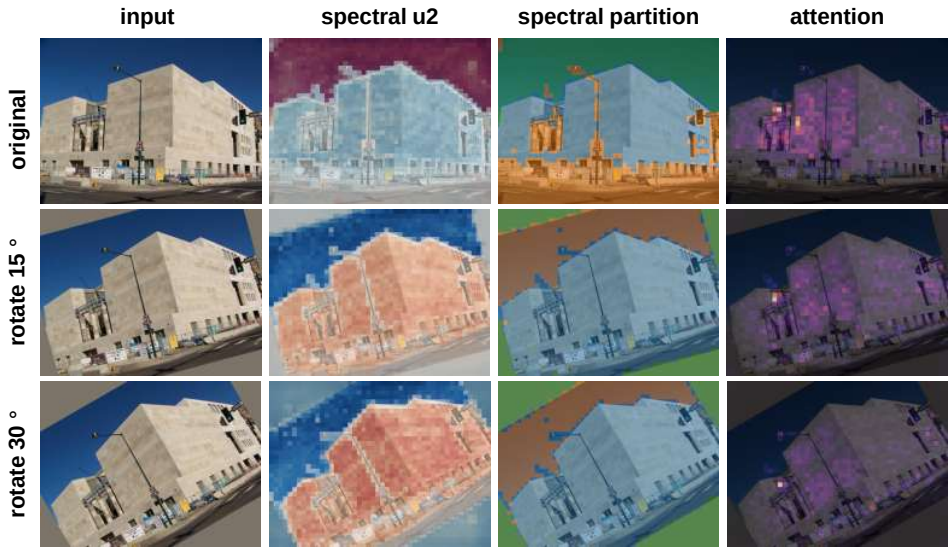


Figure F.2: Building scene at the original orientation and two prompt rotations. In this example, the low-frequency regions broadly follow the building, sky, and ground, while the attention response changes with the patch positions.

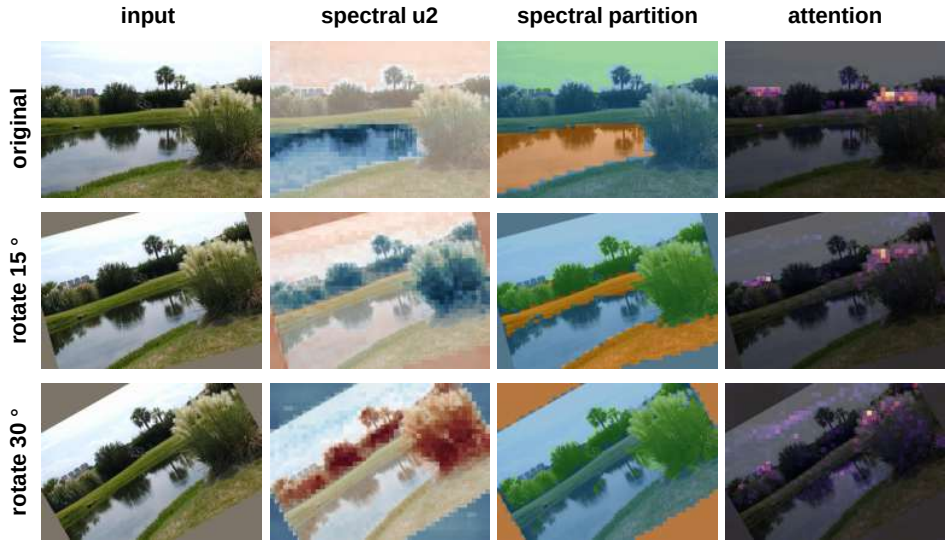


Figure F.3: Water–vegetation scene under the same settings. Although signs and displayed colors change, several partition boundaries remain associated with the pond, bank, and vegetation in the shown rotations.

F.2 Query-to-Prompt Attention Correspondence

Protocol. We further examine how prompt rotation changes the correspondence used by Painter attention. For the same adapted episode and query, we extract query-to-prompt attention from the final Transformer layer with the prompt at 0° and 35°; the prompt image and annotation are rotated together, while the query and all visualization settings remain fixed. The attention assigned to prompt tokens is arranged on the prompt patch grid. We then flatten the two maps over the entries used by the visualization and compute their Pearson correlation ρ. This compares the displayed token-grid maps without geometric realignment; it is a grid-space attention diagnostic rather than a task-performance or semantic-correspondence metric.

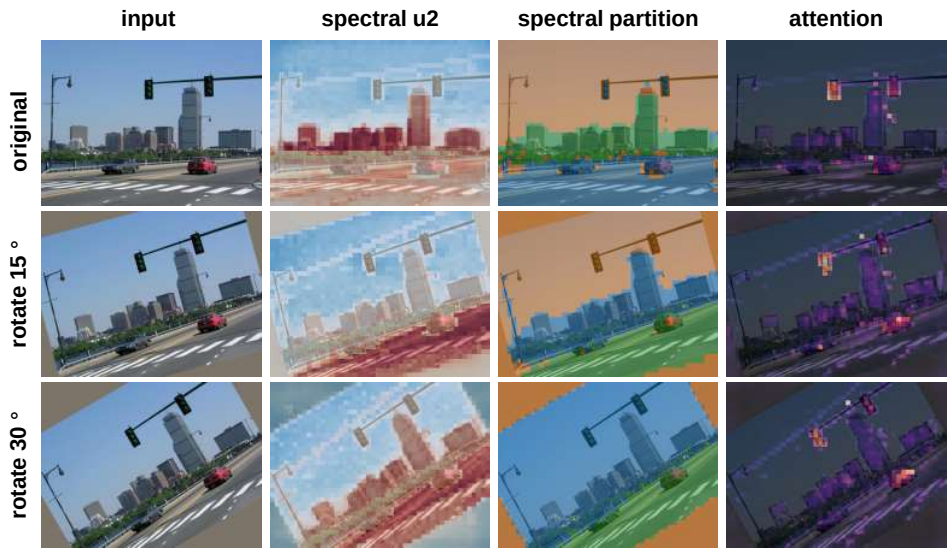


Figure F.4: Road scene under the same settings. In this example, several low-frequency region boundaries remain associated with the road, skyline, and sky despite rotation and padding, while attention peaks change location and strength.

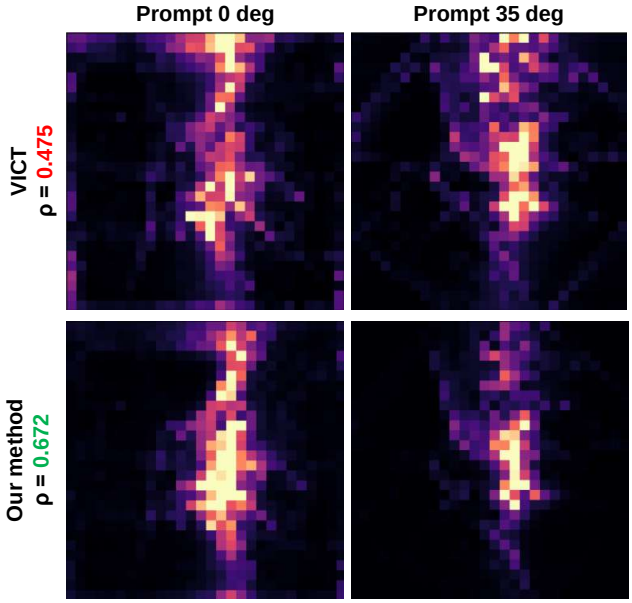


Figure F.5: Query-to-prompt attention from the final Transformer layer for the same episode with 0° and 35° prompts. The GIST maps have higher Pearson correlation in this example ($\rho = 0.672$) than the VICT maps ($\rho = 0.475$).

Observation. The displayed maps show changes in attention on the prompt token grid. GIST has higher Pearson correlation than VICT in this example, but the statistic compares fixed grid locations without undoing prompt rotation. It therefore measures grid-space repeatability, not geometric or semantic correspondence: attention that follows a rotated object can move on the grid and have lower correlation. The aligned, basis-invariant subspace diagnostic in Table 11 addresses a different question. Neither visualization establishes a general rotation law.